Noname manuscript No.
(will be inserted by the editor)

Convergence in Distribution of Randomized Algorithms: The Case of Partially Separable Optimization

D. Russell Luke

September 28, 2023

Abstract We present a Markov-chain analysis of blockwise-stochastic algorithms for solving partially block-separable optimization problems. Our main contributions to the extensive literature on these methods are statements about the Markov operators and distributions behind the iterates of stochastic algorithms, and in particular the regularity of Markov operators and rates of convergence of the *distributions* of the corresponding Markov chains. This provides a detailed characterization of the moments of the sequences beyond just the expected behavior. This also serves as a case study of how randomization restores favorable properties to algorithms that iterations of only partial information destroys. We demonstrate this on stochastic blockwise implementations of the forward-backward and Douglas-Rachford algorithms for nonconvex (and, as a special case, convex), nonsmooth optimization.

2010 Mathematics Subject Classification: Primary 65C40, 90C06, 90C26; Secondary 46N30, 60J05, 49M27, 65K05.

Keywords: Nonconvex optimization, Large-scale optimization, Markov chain, Random function iteration, Error bounds, Convergence rates

1 Introduction

We present a Markov-chain analysis of blockwise-stochastic algorithms for solving

$$\underset{x\in\mathcal{E}}{\text{minimize}}\ f(x) + \sum_{j=1}^{m} g_j(x). \tag{1}$$

Here $\mathcal{E}$ is a Euclidean space that is decomposed into a direct sum of the subspaces $\mathcal{E}_j$, denoted $\mathcal{E} = \bigoplus_{j=1}^{m}\mathcal{E}_j$, and for each $j = 1, 2, \ldots, m$, the function f is continuously differentiable with blockwise-Lipschitz gradients, g_i is everywhere subdifferentially regular (the regular and limiting subgradients coincide) and

$$g_j(x) = h_j(x_j) \tag{2}$$

for $h_j : \mathcal{E}_j \to (-\infty, +\infty]$ subdifferentially regular. This represents a partially separable structured optimization problem.

Problems with this structure are ubiquitous, and particular attention has focused on iterative algorithms for large-scale instances where the iterates are generated from only partial evaluation of the objective. Which partial information to access in each iteration is randomly selected and computations can be done in parallel across distributed systems [34,36,37,43]. There is a rich literature on the analysis of these methods, focusing mainly on deterministic properties of the objective function and expectations, iteration complexity, convergence of objective values, and acceleration strategies [8,10,13,23,31–33,35,36].

Institute for Numerical and Applied Mathematics, University of Goettingen, 37083 Goettingen, Germany. DRL was supported in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 432680300 – SFB 1456. `r.luke@math.uni-goettingen.de`; ORCID 0000-0002-4508-7360.

Address(es) of author(s) should be given

Our own contributions to the literature on such stochastic methods has focused on a stochastic block-coordinate primal-dual method for the instance of (1) where $f(x)$ is the indicator function of an affine subspace [24]. We will touch on primal-dual approaches via a stochastic blockwise Douglas-Rachford Algorithm 3, but more practical primal-dual approaches to nonsmooth problems are not on the agenda of the present study.

Our main contributions to the extensive literature on these methods are statements about the Markov operators and distributions behind the iterates of stochastic algorithms in the most complete sense possible. By that we mean not only statements about the limits of the ergodic sequences, which only tell one about the expectation, but rather the limiting distributions of the sequence of measures behind the iterates, when viewed as a Markov chain (see Theorem 7 and Proposition 17). This allows one to access the moments of the limiting sequence, not just its mean.

Getting a handle on the distributions behind iterates of randomized algorithms is significant not only for its generality, but also for the range of practical applications this encompasses. To explain this we note that, in its most general form, *consistency* of the update functions generating the Markov operators is not assumed. In plain terms, the update functions in the Markov chain need not have common fixed points. To see why this matters, it is first important to recognize that the literature on randomized algorithms is exclusively concerned with *almost sure* convergence. In [19, Proposition 2.5] it is shown that almost sure convergence of the iterates of such Markov chains can only happen when the update functions have common fixed points. Situations where the update functions do not have common fixed points are only a small perturbation away: consider any fixed point iteration with numerical error. To be sure, the consistent case allows for tremendous simplifications, and we show this in sections 3.2 and 4.2.1; the point is, however, that our approach goes far beyond this idealized case.

Previous work has established a foundation for this based on a fixed point theoretic approach [5–7,18–20]. A different perspective, modeled after a more direct analysis of the *descent* properties of algorithms in an optimization context has been established recently by Salzo and Villa [39]. This was further developed in the masters thesis of Kartamyschew [21]. In the present work we extend the results of [21] to a fully nonconvex setting for more general mappings.

A noteworthy feature of blockwise methods, and what distinguishes the present study from [18–20] is that, even when the objective in (1) is convex, blockwise algorithms do *not* satisfy the usual regularity properties enjoyed by convex optimization algorithms that lead generically to global convergence. This is demonstrated in Example 2. The stochastic implementations for convex problems, however, *do* enjoy nice properties *in expectation* (see Theorem 1), and this is enough to guarantee generic global convergence (Theorem 7, Proposition 17). While this fact lies implicitly behind the convergence analysis of, for instance, [24] and many others, it was recognized in [39] as the important property of *descent in expectation*. We place these observations in the context of Markov operators with update functions that satisfy desirable properties in expectation (see Theorem 3). These notions, at the level of the Markov operator, have already been defined in [18–20]; the convergence results presented in those works, however, are based on the assumption that each of the update functions that generate the Markov operator have the same *class* of regularity that they have in expectation. Blockwise algorithms for partially separable optimization do not enjoy this structure, and therefore many of the results of [18–20] do not immediately apply; indeed, we conjecture that some of the stronger convergence results of [18,19] are not true without additional compactness assumptions, hence our analogous global convergence statement for the convex case Proposition 6, is weaker than its counterparts [18, Theorem 3.6] or [19, Theorem 2.9].

The basic machinery of stochastic blockwise function iterations (Algorithm 1) and Markov chains is reviewed in section 2. In section 3 we review and establish the chain of regularity lifted from the regularity of the individual mappings on the sample space, Theorem 1, to the regularity of the corresponding Markov operators on the space of probability measures, Theorem 3. In section 3.2 the special case of consistent stochastic feasibility is detailed, showing in particular how the abstract objects for the general case simplify (see Theorem 5). In section 4 we present abstract convergence results, with and without rates (Proposition 6 and Theorem 7). The key to quantitative results in the space of probability measures is *metric subregularity* of the *invariant Markov transport discrepancy* (41). This is shown in the case of consistent stochastic feasibility to be *necessary* for quantitative convergence of *paracontractive* Markov operators in Theorem 9.

We return to the specialization of stochastic partial blockwise splitting algorithms in section 5, where we develop a case study of stochastic blockwise forward-backward splitting (Algorithm 2) and stochastic blockwise Douglas-Rachford (Algorithm 3), establishing the regularity of the corresponding fixed point operators (Propositions 11-14) and convergence in distribution of the corresponding Markov chains (Proposition 17).

2 Notation and Random Function Iterations

As usual, $\mathbb{N}$ denotes the natural numbers including 0. We denote by $\mathscr{P}(G)$ the set of all probability measures on $G \subset \mathcal{E}$; the measurable sets are given by the Borel sigma algebra on a subset $G \subset \mathcal{E}$, denoted by $\mathcal{B}(G)$. The notation $X \sim \mu \in \mathscr{P}(G)$ means that the law of X, denoted $\mathcal{L}(X)$, satisfies $\mathcal{L}(X) := \mathbb{P}^X := \mathbb{P}(X \in \cdot) = \mu$, where $\mathbb{P}$ is the probability measure on some underlying probability space. The open ball centered at $x \in \mathcal{E}$ with radius $r > 0$ is denoted $\mathbb{B}(x, r)$; the closure of the ball is denoted $\overline{\mathbb{B}}(x, r)$. The distance of a point $x \in \mathcal{E}$ to a set $A \subset \mathcal{E}$ in the metric d is denoted by $d(x, A) := \inf_{w \in A} d(x, w)$. The *projector* onto a set A is denoted by P_A and $P_A(x)$ is the set of all points where $d(x, A)$ is attained. This is empty if A is open, and a singleton if A is closed and convex; generically, P_A is a (possibly empty) set-valued mapping, for which we use the notation $P_A : \mathcal{E} \rightrightarrows \mathcal{E}$. For the ball of radius r around a subset of points $A \subset \mathcal{E}$, we write $\mathbb{B}(A, r) := \bigcup_{x \in A} \mathbb{B}(x, r)$.

Let $\mathbb{I}$ denote an index set, each element $i \in \mathbb{I}$ of which is a unique assignment to nonempty subsets of $\{1, 2, \ldots, m\}$: $M_i \in 2^{\{1,2,\ldots,m\}} \setminus \emptyset$ for $i \in \mathbb{I}$, where $\cup_{i \in \mathbb{I}} M_i = 2^{\{1,2,\ldots,m\}} \setminus \emptyset$ and $M_i \neq M_j$ for $i \neq j$. For convenience we will let the first such subset be the set itself: $M_1 := \{1, 2, \ldots, m\}$. For $i \in \mathbb{I}$ we denote the subspace $\mathcal{E}_{M_i} := \bigoplus_{j \in M_i} \mathcal{E}_j$ where $\{\mathcal{E}_1, \ldots, \mathcal{E}_m\}$ is a collection of mutually orthogonal subspaces of $\mathcal{E}$. The complement to this space in $\mathcal{E}$ is denoted $\mathcal{E}^\circ_{M_i} := \mathcal{E} \setminus \mathcal{E}_{M_i}$; likewise, denote the complement to the subset M_i in $\{1, 2, \ldots, m\}$ by $M_i^\circ = \{1, 2, \ldots, m\} \setminus M_i$. The *affine* embedding of the subspace $\mathcal{E}_{M_i}$ in $\mathcal{E}$ at a point $z \in \mathcal{E}$ is denoted $\mathcal{E}_{M_i} \bigoplus \{z\}$; the canonical embedding of $\mathcal{E}_{M_i}$ in $\mathcal{E}$ is thus $\mathcal{E}_{M_i} \bigoplus \{0\}$ where it is understood that $0 \in \mathcal{E}$. We use the corresponding notation for subsets $G \subset \mathcal{E}$: $G_j \subset \mathcal{E}_j$ and the affine embedding of a subset G_{M_i} at a point $z \in G_{M_i^\circ}$ is given by $G_{M_i} \bigoplus \{z\}_{M_i^\circ}$. The blockwise mappings $T_i : \mathcal{E} \to \mathcal{E}$ corresponding to this structure are defined by

$$[T_i(x)]_j := \begin{cases} T'_j(x), & j \in M_i, \\ x_j & \text{else}, \end{cases} \qquad \text{for} \quad T'_j : \mathcal{E} \to \mathcal{E}_j\,, \quad j = 1, 2, \ldots, m. \tag{3}$$

Note that T'_j is some action with respect to the j'th block in $\mathcal{E}_j$, though with input from $x \in \mathcal{E}$.

The measure space of indexes is denoted $(\mathbb{I}, \mathcal{I})$ and ξ is an $\mathbb{I}$-valued random variable on a probability space. The random variables ξ_k in the sequence $(\xi_k)_{k \in \mathbb{N}}$ (abbreviated (ξ_k)) are independent and identically distributed (i.i.d.) with ξ_k distributed as ξ ($\xi_k \sim \xi$). At each iteration k of the algorithm one selects at random a nonempty subset of blocks $M_{\xi_k} \subset \{1, 2, \ldots, m\}$ and performs an update to each block as follows:

Algorithm 1: Stochastic Block Iteration (SBI)

Initialization: Select a random variable X_0 with distribution μ, $t = (t_1, t_2, \ldots, t_m) > 0$, and $(\xi_k)_{k \in \mathbb{N}}$ an i.i.d. sequence with values on $\mathbb{I}$ and X_0 and (ξ_k) independently distributed. Given $T'_j : \mathcal{E} \to \mathcal{E}_j$ for $j = 1, 2, \ldots, m$.

for $k = 0, 1, 2, \ldots$ **do**

$$X^{k+1} = T_{\xi_k}(X^k) \quad \text{where} \quad [T_{\xi_k}(X^k)]_j := \begin{cases} T'_j(X^k), & j \in M_{\xi_k}, \\ X^k_j & \text{else} \end{cases}. \tag{4}$$

This is a special instance of a *random function iteration* studied in [18–20]. Convergence of such an iteration is understood in the sense of distributions and is a consequence of two key properties: that the mapping T_i is almost *α-firmly nonexpansive (abbreviated $a\alpha$-fne) in expectation* ((29) and (32a)), and that the *invariant Markov transport discrepancy* defined in (41) is *gauge metrically subregular* (55) at invariant measures. The latter of these two properties has been shown in many settings to be necessary for quantitative convergence of the iterates [18, 26]. The first property, with the qualifier "almost" removed, is enough to guarantee that the sequence of measures is *asymptotically regular* with respect to the *Wasserstein metric.* All this is formally defined below.

The following assumptions hold throughout.

Assumption 1. *(a) $\xi_0, \xi_1, \ldots, \xi_k$ are i.i.d random variables for all $k \in \mathbb{N}$ on a probability space with values on $\mathbb{I}$. The variable X_0 is an random variable with values on $\mathcal{E}$, independent from ξ_k.*
(b) The function $\Phi : \mathcal{E} \times \mathbb{I} \to \mathcal{E}$, $(x, i) \mapsto T_i x$ is measurable.

Let $(X_k)_{k\in\mathbb{N}}$ be a sequence of random variables with values on $G \subset \mathcal{E}$. Recall that a Markov chain with *transition kernel* p satisfies

(i) $\mathbb{P}\left(X_{k+1} \in A \,|\, X_0, X_1, \ldots, X_k\right) = \mathbb{P}\left(X_{k+1} \in A \,|\, X_k\right)$;
(ii) $\mathbb{P}\left(X_{k+1} \in A \,|\, X_k\right) = p(X_k, A)$

for all $k \in \mathbb{N}$ and $A \in \mathcal{B}(G)$ almost surely in probability, $\mathbb{P}$-a.s. In [19] it is shown that the sequence of random variables (X_k) generated by Algorithm 1 is a Markov chain with transition kernel p given by

$$(x \in G)(A \in \mathcal{B}(G)) \qquad p(x, A) := \mathbb{P}(T_\xi x \in A) \tag{5}$$

for the measurable *update function* $\Phi : G \times \mathbb{I} \to G$ given by $\Phi(x, i) := T_i x$.

The Markov operator $\mathcal{P}$ associated with this Markov chain is defined pointwise for a measurable function $f : G \to \mathbb{R}$ via

$$(x \in G) \qquad \mathcal{P}f(x) := \int_G f(y)p(x, \mathrm{d}\, y),$$

when the integral exists. Note that

$$\mathcal{P}f(x) = \int_\Omega f(T_{\xi(\omega)}x)\mathbb{P}(\mathrm{d}\,\omega) = \int_{\mathbb{I}} f(T_i x)\mathbb{P}^\xi(\mathrm{d}\, i).$$

Let $\mu \in \mathscr{P}(G)$. The dual Markov operator acting on a measure μ is indicated by action on the right by $\mathcal{P}$:

$$(A \in \mathcal{B}(G)) \qquad (\mathcal{P}^*\mu)(A) := (\mu\mathcal{P})(A) := \int_G p(x, A)\mu(\mathrm{d}\, x).$$

The distribution of the k'th iterate of the Markov chain generated by Algorithm 1 is therefore easily represented as follows: $\mathcal{L}(X_k) = \mu_0\mathcal{P}^k$, where $\mathcal{L}(X)$ denotes the law of the random variable X. Of course in general random variables do not converge, but distributions associated with the sequence of random variables (X_k) of Algorithm 1, if they converge to anything, do so to *invariant measures* of the associated Markov operator. An invariant measure of the Markov operator $\mathcal{P}$ is any distribution $\pi \in \mathscr{P}$ that satisfies $\pi\mathcal{P} = \pi$. The set of all invariant probability measures is denoted by $\mathrm{inv}\,\mathcal{P}$. The underlying problem we seek to solve is to

$$\text{Find} \qquad \pi \in \mathrm{inv}\,\mathcal{P}. \tag{6}$$

This is the *stochastic fixed point problem* studied in [19,20]. When the mappings T_i have common fixed points, the problem reduces to the *stochastic feasibility* problem studied in [18].

Let (ν_k) be a sequence of probability measures on $G \subset \mathcal{E}$, and let $C_b(G)$ denote the set of bounded and continuous functions from G to $\mathbb{R}$. The sequence (ν_k) is said to converge in distribution to ν whenever $\nu \in \mathscr{P}(G)$ and for all $f \in C_b(G)$ it holds that $\nu_k f \to \nu f$ as $k \to \infty$, where $\nu f := \int f(x)\nu(\mathrm{d}\, x)$. In other words, a sequence of random variables (X_k) converges in distribution if their laws $(\mathcal{L}(X_k))$ do. We use the *weighted Wasserstein metric* for the space of measures. Let

$$\mathscr{P}_2(G) = \left\{ \mu \in \mathscr{P}(G) \;\middle|\; \exists\, x \in G : \int \|x - y\|^2_{\mathbf{p}}\mu(\mathrm{d}\, y) < \infty \right\} \tag{7}$$

where $\|\cdot\|_{\mathbf{p}}$ is the Euclidean norm weighted by $\mathbf{p}$. This will be made explicit below. The Wasserstein 2-metric on $\mathscr{P}_2(G)$, with respect to the weighted Euclidean norm $\|\cdot\|_{\mathbf{p}}$ denoted $d_{W_{2,\mathbf{p}}}$, is defined by

$$d_{W_{2,\mathbf{p}}}(\mu, \nu) := \left(\inf_{\gamma\in\mathcal{C}(\mu,\nu)} \int_{G\times G} \|x - y\|^2_{\mathbf{p}}\gamma(dx, dy) \right)^{1/2} \tag{8}$$

where $\mathcal{C}(\mu,\nu)$ is the set of couplings of μ and ν:

$$\mathcal{C}(\mu,\nu) := \{\gamma \in \mathscr{P}(G\times G) \mid \gamma(A\times G) = \mu(A),\, \gamma(G\times A) = \nu(A) \quad \forall A \in \mathcal{B}(G)\}. \tag{9}$$

The principle mode of convergence in distribution that we use is convergence in distribution of the sequence $(\mathcal{L}(X_k))$ to a probability measure $\pi \in \mathscr{P}(G)$, i.e. for any $f \in C_b(G)$

$$\mathcal{L}(X_k)f = \mathbb{E}[f(X_k)] \to \pi f, \qquad \text{as } k \to \infty.$$

This is a stronger form of convergence than convergence of Cesàro averages sometimes seen in the literature. Since we are considering the The Wasserstein 2-metric, convergence in this metric implies that also the second moments converge in this metric. For more background on the analysis of sequences of measures we refer interested readers to [4, 16, 40–42].

2.2 Stochastic blockwise splitting algorithms

The concrete targets of the analysis presented here are two fundamental templates for solving problems of the form (1), forward-backward splitting as formulated in [39] and Douglas-Rachford splitting; the latter has not been studied in this context.

Denote by $\partial_{x_j} f : \mathcal{E} \rightrightarrows \mathcal{E}_j$ the partial limiting subdifferential of f with respect to the block $x_j \in \mathcal{E}_j$:

$$\partial_{x_j} f(\overline{x}) := \left\{ v \in \mathcal{E}_j \;\middle|\; f(x) \geq f(\overline{x}) + \left\langle v \bigoplus \{0\},\ x - \overline{x} \right\rangle + o\{\|x-\overline{x}\|\}, x \in \mathcal{E}_j \bigoplus \{\overline{x}\} \right\}. \tag{10}$$

When f is continuously differentiable, then this coincides with the partial gradient $\nabla_{x_j} f : \mathcal{E} \to \mathcal{E}_j$. The prox mapping of a function $h : \mathcal{E} \rightrightarrows (-\infty, +\infty]$ is defined by

$$\operatorname{prox}_{h,\lambda}(x) := \operatorname{argmin}_{y\in\mathcal{E}} \left\{ h(y) + \frac{1}{2\lambda}\|y-x\|^2 \right\}. \tag{11}$$

The prox mapping is nonempty and single-valued whenever h is proper, lsc and convex [30]. To allow for generalization to *nonconvex* functions we use instead the *resolvent* $J_{\partial h,\lambda} : \mathcal{E} \rightrightarrows \mathcal{E}$:

$$J_{\partial h,\lambda}(x) := \{ y \in \mathcal{E} \mid (\lambda \partial h + \mathrm{Id})(y) \ni x \}. \tag{12}$$

It is clear from this that, in general, $\operatorname{prox}_{h,\lambda}(x) \subset J_{\partial h,\lambda}(x)$ for all x.

Note that g_j defined in (2) is just the extension by zero of h_j to a mapping on $\mathcal{E}$. This yields

$$J_{\partial g_j,\lambda_j}(x) = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{j-1} \\ J_{\partial h_j,\lambda_j}(x_j) \\ x_{j+1} \\ \vdots \\ x_m \end{pmatrix} \quad \text{and} \quad \left(J_{\partial g_j,\lambda_j} - \mathrm{Id}\right)(x) = \left(J_{\partial h_j,\lambda_j}(x_j) - x_j\right) \bigoplus \{0\}. \tag{13}$$

Let $\partial_j f : \mathcal{E} \to \mathcal{E}$ denote the canonical embedding of $\partial_{x_j} f$ by zero into $\mathcal{E}$:

$$\partial_j f(x) := \partial_{x_j} f(x) \bigoplus \{0\}. \tag{14}$$

The corresponding resolvent, $J_{\partial_j f,\lambda}(x)$ is given by

$$J_{\partial_j f,\lambda}(x) := \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{j-1} \\ J_{\partial f_j(\cdot;x),\lambda}(x_j) \\ x_{j+1} \\ \vdots \\ x_m \end{pmatrix} \tag{15}$$

where $f_j(\cdot;x) : \mathcal{E}_j \to \mathbb{R}$, with $x \in \mathcal{E}$ a *parameter*, denotes

$$f_j(y;x) := f\left(x + (y - x_j) \bigoplus \{0\}\right) \tag{16a}$$

so that

$$\partial f_j(y;x) = \partial_{x_j} f\left(x + (y - x_j) \bigoplus \{0\}\right) \quad \text{and} \tag{16b}$$

$$J_{\partial f_j(\cdot;x),\lambda}(x_j) = \left\{ y \in \mathcal{E}_j \;\middle|\; y + \partial_{x_j} f(x + (y - x_j) \bigoplus \{0\}) \ni x_j \right\}. \tag{16c}$$

We recognize that the resolvent of a function that is not fully separable is not considered *prox friendly* from a computational standpoint, and this can only be evaluated numerically with some error. The framework presented here is well suited for algorithms with numerical error, and this is discussed at some length in [19, Section 4]. In the interest of keeping the presentation simple, we present results for exact evaluation of all the relevant operators; the incorporation of appropriate noise models for inexact computation builds on the structure introduced here and does not require any assumption of summable errors or increasing accuracy, though the noise model does require some careful consideration (see [19, Section 4.3]).

The abstract template is Algorithm 1 where the mappings T'_j specialize to

$$T'_j(x) := \frac{1}{2}\left(R_{\partial f_j(\cdot;y),t_j} R_{\partial h_j,t_j}(x_j) + x_j\right), \quad y = R_{\partial g_j,t_j}(x) \qquad \text{(blockwise Douglas-Rachford)}$$

where

$$R_{\partial h_j,t_j}(x_j) = 2J_{\partial h_j,t_j}(x_j) - x_j \quad \text{and} \quad R_{\partial f_j(\cdot;x),t_j}(x_j) := 2J_{\partial f_j(\cdot;x)}(x_j) - x_j.$$

or, when f is continuously differentiable,

$$T'_j(x) := J_{\partial h_j,t_j}\left(x_j - t_j \nabla_{x_j} f(x)\right) \qquad \text{(blockwise forward-backward).}$$

Using the resolvent instead of the prox mapping, the blockwise forward-backward algorithm studied in [39] consists of iterations of randomly selected mappings $T_i^{FB} : \mathcal{E} \to \mathcal{E}$:

$$T_i^{FB} := \left(\mathrm{Id} + \sum_{j \in M_i} \left(J_{\partial g_j,t_j}\left(\mathrm{Id} - t_j \nabla_j f\right) - \mathrm{Id}\right)\right) \quad (i \in \mathbb{I}). \tag{17}$$

Algorithm 2: Stochastic Blockwise Forward-Backward Splitting (S-BFBS)

Initialization: Select a random variable X_0 with distribution μ, $t = (t_1, t_2, \ldots, t_m) > 0$, and $(\xi_k)_{k \in \mathbb{N}}$ an i.i.d. sequence with values on $\mathbb{I}$ and X_0 and (ξ_k) independently distributed.

for $k = 0, 1, 2, \ldots$ **do**

$$X^{k+1} = T_{\xi_k}^{FB}(X^k) := \left(\mathrm{Id} + \sum_{j \in M_{\xi_k}} \left(J_{\partial g_j,t_j}\left(\mathrm{Id} - t_j \nabla_j f\right) - \mathrm{Id}\right)\right)(X^k), \tag{18a}$$

or equivalently

for $j = 0, 1, 2, \ldots, m$ **do**

$$X_j^{k+1} = [T_{\xi_k}^{FB}(X^k)]_j := \begin{cases} J_{\partial h_j,t_j}\left(X_j^k - t_j \nabla_{x_j} f(X^k)\right) & \text{if } j \in M_{\xi_k} \\ X_j^k & \text{else .} \end{cases} \tag{18b}$$

The blockwise Douglas-Rachford algorithm consists of iterations of randomly selected mappings $T_i^{DR} : \mathcal{E} \to \mathcal{E}$:

$$T_i^{DR} := \frac{1}{2}\left(\sum_{j \in M_i} \left(R_{\partial_j f,t_j} R_{\partial g_j,t_j} - \mathrm{Id}\right) + 2\,\mathrm{Id}\right) \quad (i \in \mathbb{I}). \tag{19}$$

In addition to its own merits, in the convex setting the Douglas-Rachford algorithm has the interpretation

Algorithm 3: Stochastic Blockwise Douglas-Rachford Splitting (S-BDRS)

Initialization: Select a random variable X_0 with distribution μ, $t = (t_1, t_2, \ldots, t_m) > 0$, and $(\xi_k)_{k\in\mathbb{N}}$ an i.i.d. sequence with values on $\mathbb{I}$ and X_0 and (ξ_k) independently distributed.

for $k = 0, 1, 2, \ldots$ **do**

$$X^{k+1} = T^{DR}_{\xi_k} X^k := \frac{1}{2}\left(\sum_{j\in M_{\xi_k}} \left(R_{\partial_j f, t_j} R_{\partial g_j, t_j} - \mathrm{Id}\right) + 2\,\mathrm{Id}\right)(X^k) \tag{20a}$$

or equivalently

for $j = 0, 1, 2, \ldots, m$ **do**

$$X^{k+1}_j = [T^{DR}_{\xi_k}(X^k)]_j := \begin{cases} \frac{1}{2}\left(R_{\partial_j f, t_j} R_{\partial h_j, t_j}(X^k_j) + X^k_j\right) & \text{if } j \in M_{\xi_k} \\ X^k_j & \text{else .}\end{cases} \tag{20b}$$

as the ADMM algorithm [15] applied to the "pre-primal" problem to (1) [11, 14]:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\; p(x) + q(Ax) \qquad \text{where} \quad p^*(-A^T x) = f(x) \quad \text{and} \quad g(x) = q^*(x). \tag{21}$$

The stochastic blockwise Douglas-Rachford Algorithm 3 therefore can be understood as a stochastic blockwise ADMM algorithm for solving (21). The discussion above about the separability of f is yet another way of understanding the observed computational difficulty of implementing this algorithm; it is quite unlikely that f given by (21) will be separable in the standard basis and therefore the resolvent (16c) will have to be computed numerically. Alternative primal-dual methods that circumvent this are the topic of future research.

Before we begin, however, it will be helpful to give an example delineating consistent from inconsistent feasibility.

Example 1 (consistent/inconsistent stochastic feasibility problems)**.** *Examples for partially separable optimization and blockwise algorithms abound, particularly in machine learning, but seldom is the distinction made between consistent and inconsistent problems. This is illustrated here for the problem of set feasibility, or, when feasible points don't exist, best approximation. Consider the problem*

$$\mathsf{Find}\ \overline{x} \in \cap_{j=1}^m \Omega$$

where $\Omega_j \subset \mathbb{R}^n$ are closed sets. This can be recast as the following optimization problem on the product space $(\mathbb{R}^n)^m$:

$$\textit{minimize}\ f(x) + \sum_{j=1}^m \iota_{\Omega_j}(x_j) \tag{22}$$

where $x_j \in \mathbb{R}^n$,

$$\iota_\Omega(x) = \begin{cases} 0 & \textit{when } x \in \Omega \\ +\infty & \textit{else,}\end{cases}$$

and f is some reasonable coupling function that promotes similarity between the blocks x_j. In the context of problem (1) $h_j = \iota_{\Omega_j}$. Common instances of the coupling function are $f(x) = \frac{1}{2}d(x, D)^2$ for $D := \{x = (x_1, x_2, \ldots, x_m) \mid x_i = x_j\ \forall i \neq j\}$ or the more strict indicator function $f(x) = \iota_D(x)$. The prox operators associated with the indicator functions are just projectors, while the gradient of the function f in the smooth case can be constructed from the projection onto D (just the averaging operator).

When $\cap_{j=1}^m \Omega \neq \emptyset$, the solutions to the feasibility problem and problem (22) coincide for both instances of f. In this case the blockwise operators T_i in (2) and (3) have common fixed points, which are (perhaps not exclusively) points where the sets intersect, and so, when all goes well, fixed points of these algorithms coincide with points in $\cap_{j=1}^m \Omega$; at the very least fixed points of the algorithms coincide with critical points*. Viewed as random function iterations, the iterates of (2) and (3) in this consistent case are random variables whose distributions converge to delta functions with support in $\cap_{j=1}^m \Omega$ and the algorithms converge to solutions of a* stochastic feasibility problem *studied in [18]:*

$$\mathsf{Find}\ \overline{x} \in \{x \mid \mathbb{P}(x \in \mathsf{Fix}\ T_\xi) = 1\}.$$

If the intersection is empty, as will often be the case in practice regardless of noise considerations, then it is easy to see that the blockwise operators in (2) *and* (3) *do not have common fixed points when* $f(x) = \iota_D(x)$. *The random algorithms do not have fixed points in this case, but viewed as random function iterations, the distributions of the iterates converge to invariant measures of the Markov operator corresponding to either Algorithm* (2) *or* (3). *These algorithms therefore find solutions to the more general stochastic fixed point problem 6 studied in [19]. How to interpret such invariant measures is an open issue in general. For this example, in the case of just two convex sets with empty intersection, the invariant probability measures will consist of equally weighted pairs of delta functions centered at best approximation pairs between the sets.*

The numerical behavior of deterministic versions of (2) *and* (3)*, and many others, has been thoroughly studied for the broad class of* cone and sphere problems*, which includes sensor localization, phase retrieval, and computed tomography [25]. For the example of set feasibility presented here, convergence depends on the regularity properties of the projectors onto the respective sets, which as shown in [27] is derived from the regularity of the sets. The main contribution of this article is to show that randomization can lead to Markov operators with better regularity than that of the individual operators generating its transition kernel.*

3 Regularity

Our main results concern convergence of Markov chains under regularity assumptions that are lifted from the generating mappings T_i. In [27] a framework was developed for a quantitative convergence analysis of set-valued mappings T_i that are one-sided Lipschitz continuous in the sense of set-valued-mappings with Lipschitz constant slightly greater than 1. We begin with the regularity of T_i and follow this through to the regularity of the resulting Markov operator.

3.1 Almost α-firmly nonexpansive mappings

Let $G \subset \mathcal{E}$ and let $F : G \rightrightarrows \mathcal{E}$. The mapping F is said to be *pointwise almost nonexpansive at* $x_0 \in G$ *on* G whenever

$$\exists \epsilon \in [0,1): \quad \|x^+ - x_0^+\| \le \sqrt{1+\epsilon}\,\|x - x_0\|, \qquad \forall x \in G, \forall x^+ \in Fx, x_0^+ \in Fx_0. \tag{23}$$

The *violation* is a value of ϵ for which (23) holds. When the above inequality holds for all $x_0 \in G$ then F is said to be *almost nonexpansive on* G. When $\epsilon = 0$ the mapping F is said to be *(pointwise) nonexpansive*. The mapping F is said to be *pointwise almost* α*-firmly nonexpansive at* $x_0 \in G$ *on* G, abbreviated *pointwise a*α*-fne* whenever

$$\begin{aligned} &\exists \epsilon \in [0,1) \text{ and } \alpha \in (0,1): \\ &\quad \|x^+ - x_0^+\|^2 \le (1+\epsilon)\|x - x_0\|^2 - \tfrac{1-\alpha}{\alpha}\psi(x, x_0, x^+, x_0^+) \\ &\qquad\qquad\qquad\qquad\qquad \forall x \in G, \forall x^+ \in Fx, \forall x_0^+ \in Fx_0, \end{aligned} \tag{24}$$

where the *transport discrepancy* ψ of F at x, x_0, $x^+ \in Fx$ and $x_0^+ \in Fx_0$ is defined by

$$\begin{aligned} &\psi(x, x_0, x^+, x_0^+) := \\ &\|x^+ - x\|^2 + \|x_0^+ - x_0\|^2 + \|x^+ - x_0^+\|^2 + \|x - x_0\|^2 - \|x^+ - x_0\|^2 - \|x - x_0^+\|^2. \end{aligned} \tag{25}$$

When the above inequality holds for all $x_0 \in G$ then F is said to be *a*α*-fne on* G. The *violation* is the constant ϵ for which (24) holds. When $\epsilon = 0$ the mapping F is said to be *(pointwise)* α*-firmly nonexpansive*, abbreviated *(pointwise)* α*-fne.*

The transport discrepancy ψ is a central object for characterizing the regularity of mappings in metric spaces and ties the regularity of the mapping to the geometry of the space. A short calculation shows that, in a Euclidean space, this has the representation

$$\psi(x, x_0, x^+, x_0^+) = \|(x - x^+) - (x_0 - x_0^+)\|^2. \tag{26}$$

The definition of pointwise aα-fne mappings in Euclidean spaces appeared first in [27]. This generalizes the notion of *averaged* mappings dating back to Mann, Krasnoselskii, and others [2, 9, 12, 22, 29].

A partial blockwise mapping T_i that is α-fne on an affine subspace $\mathcal{E}_{M_i} \bigoplus \{z\}$ may not be α-fne on $\mathcal{E}$, as the next example from [21, Remark 3.9] shows.

Example 2. *Let $\mathcal{E} = \mathbb{R}^2$ and define $f(x_1, x_2) = (x_1 + x_2)^2$, $g_1(x_1, x_2) = h_1(x_1) = 0$ and $g_2(x_1, x_2) = h_2(x_2) = x_2^2$. Here f is convex and differentiable with global gradient Lipschitz constant $L = 4$ and the functions g_j are clearly convex. The proximal gradient algorithm applied to the function $F = f + \sum_{j=1}^{2} g_j$ is $x^{k+1} = T^{FB}(x^k) = \operatorname{prox}_g(\operatorname{Id} - t\nabla f)(x^k)$. For all $t \in (0, 1/2)$ it can be shown that the fixed point mapping T^{FB} is α-firmly nonexpansive with the unique fixed point $(0, 0)$, the global minimum of the objective function F. Hence from any initial point x^0 this iteration converges to the global minimum $(0, 0)$. A blockwise implementation of this algorithm would involve computing the proximal gradient step with respect to x_1, leaving x_2 fixed; that is at some iterations k one computes*

$$x^{k+1} = T_1^{FB}(x^k) := \operatorname{prox}_{g_1}((\operatorname{Id} - t\nabla_{x_1^k} f)(x_1^k, x_2^k) = ((1 - 2t)x_1^k - 2tx_2^k, x_2^k). \tag{27}$$

A straightforward calculation shows that the blockwise mapping T_1^{FB} is not α-fne on $\mathbb{R}^2$ for any $t > 0$, although it is α-fne on $\mathbb{R} \times \{z\}$ for any $z \in \mathbb{R}$ whenever $t \in (0, 1/2)$. Being α-fne on $\mathbb{R} \times \{z\}$ for any $z \in \mathbb{R}$ is not much help, however, since this means that repeated application of T_1^{FB} defined by (27) converges to the minimum of F restricted to the affine subspace $\mathbb{R} \times \{z\}$, namely $(-z, z)$.

In light of the above counterexample, Theorem 1 below shows how randomization in the blockwise forward-backward algorithm restores the α-fne property *in expectation* [20, Definition 3.6]. This is the fixed point analog to descents in expectation introduced in [39].

In the stochastic setting we consider only *single-valued* mappings T_i that are aα-fne in expectation. We can therefore write $x^+ = T_i x$ instead of always taking some selection $x^+ \in T_i x$ (which then raises issues of measurability and so forth). On a closed subset $G \subset \mathcal{E}$ for a general self-mapping $T_i : G \to G$ for $i \in \mathbb{I}$, the mapping $\Phi : G \times \mathbb{I} \to G$ be given by $\Phi(x, i) = T_i x$ is said to be *pointwise almost nonexpansive in expectation at $x_0 \in G$ on G*, abbreviated *pointwise almost nonexpansive in expectation*, whenever

$$\exists \epsilon \in [0, 1) : \quad \mathbb{E}\left[\|\Phi(x, \xi) - \Phi(x_0, \xi)\|\right] \le \sqrt{1 + \epsilon}\, \|x - x_0\|, \qquad \forall x \in G. \tag{28}$$

When the above inequality holds for all $x_0 \in G$ then Φ is said to be *almost nonexpansive in expectation on G*. As before, the violation is a value of ϵ for which (28) holds. When the violation is 0, the qualifier "almost" is dropped. The mapping Φ is said to be *pointwise almost α-firmly nonexpansive in expectation at $x_0 \in G$ on G*, abbreviated *pointwise aα-fne in expectation*, whenever

$$\begin{aligned} &\exists \epsilon \in [0, 1), \alpha \in (0, 1) : \quad \forall x \in G, \\ &\mathbb{E}\left[\|\Phi(x, \xi) - \Phi(x_0, \xi)\|^2\right] \le (1 + \epsilon)\|x - x_0\|^2 - \tfrac{1-\alpha}{\alpha}\mathbb{E}\left[\psi(x, x_0, \Phi(x, \xi), \Phi(x_0, \xi))\right]. \end{aligned} \tag{29}$$

When the above inequality holds for all $x_0 \in G$ then Φ is said to be *almost α-firmly nonexpansive (aα-fne) in expectation on G*. The violation is a value of ϵ for which (29) holds. When the violation is 0, the qualifier "almost" is dropped and the abbreviation *α-fne in expectation* is used. The defining inequalities (28) and (29) will be amended below in (32a) to account for *weighted norms*.

The next result, derived from [21, Proposition 5.5] shows in particular that any collection of self-mappings $\{T_i\}_{i \in \mathbb{I}}$ on $G \subset \mathcal{E}$ that is aα-fne on $G_{M_i} \bigoplus \{z\}$ is aα-fne in expectation with respect to a weighted norm on G. In particular, denote by η_i the probability of selecting the i'th collection of blocks, M_i, and let p_j denote the probability that the j'th block is among the randomly selected collection of blocks:

$$0 < p_j = \sum_{i \in \mathbb{I}} \eta_i \cdot \chi_{M_i}(j) \le 1 \quad \text{where} \quad \chi_{M_i}(j) = \begin{cases} 1 & \text{if } j \in M_i \\ 0 & \text{else} \end{cases} \quad (j = 1, 2, \ldots, m). \tag{30}$$

Define the corresponding weighted norm

$$\|z\|_{\mathbf{p}} := \left(\sum_{j=1}^{m} \tfrac{1}{p_j} \|z_j\|_{\mathcal{E}_j}^2\right)^{1/2}. \tag{31}$$

Theorem 1 (almost α-firmly nonexpansive in expectation (aα-fne in expectation))**.** *Let the single-valued self-mappings $\{T_i\}_{i\in\mathbb{I}}$ on the subset $G \subset \mathcal{E}$ satisfy*

(a) for each i, T_i is the identity mapping on $\mathcal{E}_{M_i^\circ}$;
(b) T_1 is aα-fne on G with constant $\overline{\alpha}$ and violation no greater than $\overline{\epsilon}$ where $M_1 := \{1, 2, \ldots, m\}$.

Then

(i) for all i and each $z \in G$, T_i is aα-fne on $G_{M_i} \bigoplus \{z\}_{M_i^\circ}$ with constant at most $\overline{\alpha}$ and violation no greater than $\overline{\epsilon}$;
(ii) the mapping $\Phi : G \times \mathbb{I} \to G$ given by $\Phi(x, i) = T_i x$ satisfies

$$\mathbb{E}\left[\|\Phi(x,\xi) - \Phi(y,\xi)\|_{\mathbf{p}}^2\right] \le (1 + \overline{p\epsilon})\|x - y\|_{\mathbf{p}}^2 - \tfrac{1-\overline{\alpha}}{\overline{\alpha}}\mathbb{E}\left[\psi_{\mathbf{p}}(x, y, \Phi(x,\xi), \Phi(y,\xi))\right] \quad \forall x, y \in G \tag{32a}$$

where

$$\psi_{\mathbf{p}}(x, y, \Phi(x,i), \Phi(y,i)) := \|\,(x - \Phi(x,i)) - (y - \Phi(y,i))\,\|_{\mathbf{p}}^2 \quad \text{and} \quad \overline{p} := \max_j\{p_j\}. \tag{32b}$$

A mapping $\Phi : G \times \mathbb{I} \to G$ that satisfies (32a) is called *aα-fne in expectation* with respect to the weighted norm $\|\cdot\|_{\mathbf{p}}$ with constant $\overline{\alpha}$ and violation no greater than $\overline{p\epsilon}$.
Proof. The proof of part (i) follows immediately from the observation that T_i on $G_{M_i} \bigoplus \{z\}_{M_i^\circ}$ is equivalent to T_1 restricted to the same subset.

To see part (ii), fix any $x, y \in G$, and let $T_j' : G_j \to G_j$ $(j = 1, 2, \ldots, m)$ be the j'th block mapping for $j \in M_i$. Hence, $T_i(x) = P_{\mathcal{E}_{M_i}} T_i(x) + P_{\mathcal{E}_{M_i^\circ}}(x)$ and $T_1(x) = \bigoplus_{j=1}^m T_j'(x)$ where $P_{\mathcal{E}_{M_i}} : \mathcal{E} \to \mathcal{E}$ is the orthogonal projection onto the subspace $\mathcal{E}_{M_i}$ and likewise for $P_{\mathcal{E}_{M_i^\circ}}$. We begin with the left hand side of the defining inequality:

$$\begin{aligned}
\mathbb{E}\left[\|\Phi(x,\xi) - \Phi(y,\xi)\|_{\mathbf{p}}^2\right] &= \sum_{i\in\mathbb{I}} \eta_i \|T_i(x) - T_i(y)\|_{\mathbf{p}}^2 \\
&= \sum_{i\in\mathbb{I}} \eta_i \left\|\left(P_{\mathcal{E}_{M_i}} T_i(x) + P_{\mathcal{E}_{M_i^\circ}}(x)\right) - \left(P_{\mathcal{E}_{M_i}} T_i(y) + P_{\mathcal{E}_{M_i^\circ}}(y)\right)\right\|_{\mathbf{p}}^2 \\
&= \sum_{i\in\mathbb{I}} \eta_i \left(\left\|P_{\mathcal{E}_{M_i}} T_i(x) - P_{\mathcal{E}_{M_i}} T_i(y)\right\|_{\mathbf{p}}^2 + \left\|P_{\mathcal{E}_{M_i^\circ}}(x - y)\right\|_{\mathbf{p}}^2\right) \\
&= \sum_{i\in\mathbb{I}} \eta_i \left(\sum_{j\in M_i} \tfrac{1}{p_j} \left\|T_j'(x) - T_j'(y)\right\|_{\mathcal{E}_j}^2 + \sum_{k\in M_i^\circ} \tfrac{1}{p_k} \|x_k - y_k\|_{\mathcal{E}_k}^2\right).
\end{aligned} \tag{33}$$

Then (33) rearranges to

$$\begin{aligned}
\mathbb{E}\left[\|\Phi(x,\xi) - \Phi(y,\xi)\|_{\mathbf{p}}^2\right] &= \sum_{i\in\mathbb{I}} \eta_i \left(\sum_{j\in M_i} \tfrac{1}{p_j} \left\|T_j'(x) - T_j'(y)\right\|_{\mathcal{E}_j}^2 + \sum_{k\in M_i^\circ} \tfrac{1}{p_k} \|x_k - y_k\|_{\mathcal{E}_k}^2\right) \\
&= \sum_{j=1}^m p_j \tfrac{1}{p_j} \left\|T_j'(x) - T_j'(y)\right\|_{\mathcal{E}_j}^2 + (1 - p_j)\tfrac{1}{p_j} \|x_j - y_j\|_{\mathcal{E}_j}^2 \\
&= \|T_1(x) - T_1(y)\|^2 - \|x - y\|^2 + \|x - y\|_{\mathbf{p}}^2.
\end{aligned} \tag{34}$$

We simplify the expectation of the weighted transport discrepancy (32b) next.

$$\begin{aligned}
\mathbb{E}\left[\psi_{\mathbf{p}}(x, y, \Phi(x,\xi), \Phi(y,\xi))\right] &= \sum_{i\in\mathbb{I}} \eta_i \left(\|(x - T_i(x)) - (y - T_i(x))\|_{\mathbf{p}}^2\right) \\
&= \sum_{j=1}^m p_j \tfrac{1}{p_j} \left\|\left(x_j - T_j'(x)\right) - \left(y_j - T_j'(y)\right)\right\|_{\mathcal{E}_j}^2 \\
&= \|(x - T_1(x)) - (y - T_1(y))\|^2.
\end{aligned} \tag{35}$$

Combining (34) with $\frac{1-\overline{\alpha}}{\overline{\alpha}}$ times (35) yields

$$\mathbb{E}\left[\|\Phi(x,\xi)-\Phi(y,\xi)\|_{\mathbf{p}}^2\right]+\tfrac{1-\overline{\alpha}}{\overline{\alpha}}\mathbb{E}\left[\psi_{\mathbf{p}}(x,y,\Phi(x,\xi),\Phi(y,\xi))\right]$$
$$=\|T_1(x)-T_1(y)\|^2-\|x-y\|^2+\|x-y\|_{\mathbf{p}}^2+\tfrac{1-\overline{\alpha}}{\overline{\alpha}}\|(x-T_1(x))-(y-T_1(y))\|^2. \tag{36}$$

Now by assumption (b), T_1 is aα-fne with constant $\overline{\alpha}$ and violation no greater than $\overline{\epsilon}$ on G. Therefore (36) is bounded by

$$\mathbb{E}\left[\|\Phi(x,\xi)-\Phi(y,\xi)\|_{\mathbf{p}}^2\right]+\tfrac{1-\overline{\alpha}}{\overline{\alpha}}\mathbb{E}\left[\psi_{\mathbf{p}}(x,y,\Phi(x,\xi),\Phi(y,\xi))\right]\leq\overline{\epsilon}\,\|x-y\|^2+\|x-y\|_{\mathbf{p}}^2$$
$$\leq(1+\overline{p\epsilon})\,\|x-y\|_{\mathbf{p}}^2 \tag{37}$$

for all $x,y\in G$ as claimed. □

Following [20], we lift these notions to the analogous regularity of Markov operators on the space of probability measures. Let $\mathcal{P}$ be the Markov operator with transition kernel

$$(x\in G\subset\mathcal{E})(A\in\mathcal{B}(G))\qquad p(x,A):=\mathbb{P}(\Phi(x,\xi)\in A)$$

where ξ is an $\mathbb{I}$-valued random variable and $\Phi:\ G\times\mathbb{I}\to G$ is a measurable update function. The Markov operator is said to be *pointwise almost nonexpansive in measure at* $\mu_0\in\mathscr{P}(G)$ on $\mathscr{P}(G)$, abbreviated *pointwise almost nonexpansive in measure*, whenever

$$\exists\epsilon\in[0,1):\quad d_{W_{2,\mathbf{p}}}(\mu\mathcal{P},\mu_0\mathcal{P})\leq\sqrt{1+\epsilon}\,d_{W_{2,\mathbf{p}}}(\mu,\mu_0),\qquad\forall\mu\in\mathscr{P}(G). \tag{38}$$

When the above inequality holds for all $\mu_0\in\mathscr{P}(G)$ then $\mathcal{P}$ is said to be *almost nonexpansive in measure on* $\mathscr{P}(G)$. As before, the violation is a value of ϵ for which (38) holds. When the violation is 0, the qualifier "almost" is dropped. Let $\mathcal{C}_*(\mu_1,\mu_2)$ denote the set of couplings where the distance $d_{W_{2,\mathbf{p}}}(\mu_1,\mu_2)$ is attained (i.e. the optimal couplings between μ_1 and μ_2) The Markov operator $\mathcal{P}$ is said to be *pointwise almost α-firmly nonexpansive in measure at* $\mu_0\in\mathscr{P}(G)$ on $\mathscr{P}(G)$, abbreviated *pointwise aα-fne in measure*, whenever

$$\exists\epsilon\in[0,1),\alpha\in(0,1):\qquad\forall\mu\in\mathscr{P}(G),\forall\gamma\in C_*(\mu,\mu_0)$$
$$d_{W_{2,\mathbf{p}}}(\mu\mathcal{P},\mu_0\mathcal{P})^2\leq(1+\epsilon)d_{W_{2,\mathbf{p}}}(\mu,\mu_0)^2-$$
$$\tfrac{1-\alpha}{\alpha}\int_{G\times G}\mathbb{E}\left[\psi_{\mathbf{p}}(x,y,\Phi(x,\xi),\Phi(y,\xi))\right]\gamma(dx,dy). \tag{39}$$

When the above inequality holds for all $\mu_0\in\mathscr{P}(G)$ then $\mathcal{P}$ is said to be *aα-fne in measure on* $\mathscr{P}(G)$. The violation is a value of ϵ for which (39) holds. When the violation is 0, the qualifier "almost" is dropped and the abbreviation *α-fne in measure* is employed. The notions above were defined in [20, Definition 2.8] on more general metric spaces.

Proposition 2 (Proposition 2.10, [20])**.** *Let $G\subset\mathcal{E}$, let $\Phi:\ G\times\mathbb{I}\to G$ be given by $\Phi(x,i)=T_ix$ and let $\psi_{\mathbf{p}}$ be defined by* (32b)*. Denote by $\mathcal{P}$ the Markov operator with update function Φ and transition kernel p defined by* (5)*. If Φ is aα-fne in expectation on G with constant $\alpha\in(0,1)$ and violation $\epsilon\in[0,1)$, then the Markov operator $\mathcal{P}$ is aα-fne in measure on $\mathscr{P}_2(G)$ with constant α and violation at most ϵ, that is, $\mathcal{P}$ satisfies*

$$d^2_{W_{2,\mathbf{p}}}(\mu_1\mathcal{P},\mu_2\mathcal{P})\leq(1+\epsilon)d^2_{W_{2,\mathbf{p}}}(\mu_1,\mu_2)-\tfrac{1-\alpha}{\alpha}\int_{G\times G}\mathbb{E}\left[\psi_{\mathbf{p}}(x,y,\Phi(x,\xi),\Phi(y,\xi))\right]\ \gamma(dx,dy)$$
$$\forall\mu_2,\mu_1\in\mathscr{P}_2(G),\ \forall\gamma\in C_*(\mu_1,\mu_2). \tag{40}$$

Theorem 3 (stochastic block iterations)**.** *Let the single-valued self-mappings $\{T_i\}_{i\in\mathbb{I}}$ on the convex subset $G\subset\mathcal{E}$ satisfy*

(a) T_i is the identity mapping on $\mathcal{E}_{M_i^\circ}$;
(b) T_1 is aα-fne on G with constant $\overline{\alpha}$ and violation no greater than $\overline{\epsilon}$.

Then the Markov operator $\mathcal{P}$ with update function Φ is aα-fne in measure with constant $\overline{\alpha}$ and violation at most $\overline{p\epsilon}$.

Proof. This is an immediate consequence of Theorem 1 and Proposition 2. □

Note also that, since $\psi_{\mathbf{p}}$ is nonnegative, T_i is also almost nonexpansive in expectation whenever T_1 is aα-fne; the corresponding Markov operator is almost nonexpansive in measure with the corresponding violation whenever conditions (a)-(b) of Theorem 3 are satisfied.

In preparation for the next refinements, following [20] we lift the weighted transport discrepancy $\psi_{\mathbf{p}}$ to the corresponding *invariant Markov transport discrepancy* $\Psi : \mathscr{P}(G) \to \mathbb{R}_+ \cup \{+\infty\}$ on the subset $G \subset \mathcal{E}$ defined by

$$\Psi(\mu) := \inf_{\pi \in \operatorname{inv} \mathcal{P}} \inf_{\gamma \in \mathcal{C}_*(\mu,\pi)} \left(\int_{G \times G} \mathbb{E}\left[\psi_{\mathbf{p}}(x, y, T_\xi x, T_\xi y)\right] \; \gamma(dx, dy) \right)^{1/2}. \tag{41}$$

It is not guaranteed that both $\operatorname{inv} \mathcal{P}$ and $\mathcal{C}_*(\mu, \pi)$ are nonempty; when at least one of these is empty $\Psi(\mu) := +\infty$. It is clear that $\Psi(\pi) = 0$ for any $\pi \in \operatorname{inv} \mathcal{P}$.

3.2 Special Case: consistent stochastic feasibility

The stochastic fixed point problem (6) is called *consistent* in [18–20] when, for some closed subset $G \subset \mathcal{E}$,

$$C := \{x \in G \mid \mathbb{P}(x = T_\xi x) = 1\} \neq \emptyset. \tag{42}$$

In this case, the notions developed above can be sharpened.

Recall that a *paracontraction* is a continuous mapping $T : G \to G$ possessing fixed points that satisfies

$$\|T(x) - y\| < \|x - y\| \quad \forall y \in \mathsf{Fix}\, T, \forall x \in G \setminus \mathsf{Fix}\, T.$$

Any α-fne mapping on a Euclidean space, for example, is a paracontraction.

The notion of paracontractions extends to random function iterations for consistent stochastic feasibility. Continuous self-mappings $T_i : G \to G$ $(i \in \mathbb{I})$ are *paracontractions in expectation* with respect to the weighted norm $\|z\|_{\mathbf{p}}$ whenever

$$C \neq \emptyset \quad \text{and} \quad \mathbb{E}\left[\|T_\xi x - y\|_{\mathbf{p}}\right] < \|x - y\|_{\mathbf{p}} \quad \forall y \in C, \forall x \in G \setminus \mathsf{Fix}\, T. \tag{43}$$

The next result shows that, for consistent stochastic feasibility, collections of mappings T_i defined in Theorem 1 with $\overline{\epsilon} = 0$ are paracontractions in expectation.

Corollary 4 (paracontractions in expectation)**.** *Let the single-valued self-mappings $\{T_i\}_{i \in \mathbb{I}}$ on G satisfy*

(a) T_i is the identity mapping on $\mathcal{E}_{M_i^\circ}$;
(b) for every $z \in M_i^\circ$, T_i is α-fne on $G_{M_i} \bigoplus \{z\}_{M_i^\circ}$ with constant $\overline{\alpha}$ for all i;
(c) $C := \{x \in G \mid \mathbb{P}(x = T_\xi x) = 1\} \neq \emptyset$.

Then the mapping $\Phi : G \times \mathbb{I} \to G$ given by $\Phi(x, i) = T_i x$ is a paracontraction in expectation:

$$\mathbb{E}\left[\|\Phi(x, \xi) - \Phi(y, \xi)\|_{\mathbf{p}}^2\right] < \|x - y\|_{\mathbf{p}}^2 \quad \forall x \in G \setminus C, \forall y \in C. \tag{44}$$

Proof. Note that $\psi_{\mathbf{p}}$ takes the value 0 only when x and y are both in $\mathsf{Fix}\, T_i$; hence, for all $y \in C$

$$\mathbb{E}\left[\|T_\xi(x) - T_\xi(y)\|_{\mathbf{p}}^2\right] < \|x - y\|_{\mathbf{p}}^2 \quad \forall x \in G \setminus C. \tag{45}$$

□

To show the analogous result for the Markov operator $\mathcal{P}$ requires more work. A Markov operator is a paracontraction with respect to the weighted Wasserstein metric $d_{W_{2,M}}$ whenever

$$\operatorname{inv} \mathcal{P} \neq \emptyset \quad \text{and} \quad d_{W_{2,M}}(\mu\mathcal{P}, \pi) < d_{W_{2,M}}(\mu, \pi) \quad \forall \pi \in \operatorname{inv} \mathcal{P}, \forall \mu \in \mathscr{P}(G) \setminus \operatorname{inv} \mathcal{P}. \tag{46}$$

In the case of consistent stochastic feasibility, the invariant Markov transport discrepancy reduces to a very simple form. Indeed, note first of all that a δ-distribution centered on any point $x \in C$ is invariant with respect to $\mathcal{P}$ so the set of invariant measures supported on C,

$$\mathscr{C} := \{\mu \in \operatorname{inv} \mathcal{P} \mid \operatorname{supp} \mu \subset C\}, \tag{47}$$

is nonempty whenever C is. Now suppose $\pi \in \mathscr{C}$. Then $y = T_\xi y$ almost surely whenever $y \in \operatorname{supp} \pi$ and (35) yields

$$\begin{aligned}
\inf_{\gamma\in\mathcal{C}_*(\mu,\pi)} \left(\int_{G\times G} \mathbb{E}\left[\psi_{\mathbf{p}}(x,y,T_\xi x,T_\xi y)\right] \ \gamma(dx,dy) \right)^{1/2} &= \inf_{\gamma\in\mathcal{C}_*(\mu,\pi)} \left(\int_{G\times G} \mathbb{E}\left[\|x-T_\xi x\|_{\mathbf{p}}^2\right] \ \gamma(dx,dy) \right)^{1/2} \\
&= \left(\int_G \mathbb{E}\left[\|x-T_\xi x\|_{\mathbf{p}}^2\right] \ \mu(dx) \right)^{1/2} \\
&= \left(\int_G \|x-T_1 x\|^2 \, \mu(dx) \right)^{1/2} \qquad \forall \pi \in \mathscr{C}. \qquad (48)
\end{aligned}$$

Thus the invariant Markov transport discrepancy defined in (41) has the following simple upper bound:

$$\begin{aligned}
\Psi(\mu) &:= \inf_{\pi\in\operatorname{inv}\mathcal{P}} \inf_{\gamma\in\mathcal{C}_*(\mu,\pi)} \left(\int_{G\times G} \mathbb{E}\left[\psi_{\mathbf{p}}(x,y,T_\xi x,T_\xi y)\right] \ \gamma(dx,dy) \right)^{1/2} \\
&\leq \inf_{\pi\in\mathscr{C}} \inf_{\gamma\in\mathcal{C}_*(\mu,\pi)} \left(\int_{G\times G} \mathbb{E}\left[\psi_{\mathbf{p}}(x,y,T_\xi x,T_\xi y)\right] \ \gamma(dx,dy) \right)^{1/2} \\
&= \left(\int_G \|x-T_1 x\|^2 \, \mu(dx) \right)^{1/2}, \qquad (49)
\end{aligned}$$

where the last equality follows from (48). Inequality (49) is tight for all μ supported on C, so clearly $\mu \in \mathscr{C}$ implies that $\Psi(\mu) = 0$. On the other hand, if $\Psi(\mu) = 0$ implies that $\operatorname{supp} \mu \subset C$, then $\mathscr{C} = \operatorname{inv}\mathcal{P}$ and (49) holds with equality for all μ. This holds, in particular, when T_i is a paracontraction in expectation (see [18, Lemma 3.3] and Theorem 5 below).

Let's assume, then, that $\Psi(\mu) = 0$ if and only if $\operatorname{supp} \mu \subset C$. Then

$$d_{W_{2,\mathbf{p}}}(\mu, \operatorname{inv}\mathcal{P}) = \left(\int_G \inf_{z\in C} \|x-z\|_{\mathbf{p}}^2 \mu(dx) \right)^{1/2},$$

and (49) holds with equality, so

$$d_{W_{2,\mathbf{p}}}(\mu, \operatorname{inv}\mathcal{P}) = d_{W_{2,\mathbf{p}}}(\mu, \Psi^{-1}(0)) = \left(\int_G \inf_{z\in C} \|x-z\|_{\mathbf{p}}^2 \mu(dx) \right)^{1/2}. \qquad (50)$$

Theorem 5 (Markov operators of paracontractions in expectation)**.** *Let $G \subset \mathcal{E}$ be closed. If the continuous self-mappings $T_i : G \to G$ $(i \in \mathbb{I})$ defined by* (3) *are paracontractions in expectation on G with respect to the weighted norm $\|\cdot\|_{\mathbf{p}}$ defined by* (31)*, then*

(i) the associated Markov operator $\mathcal{P}$ is a paracontraction with respect to $d_{W_{2,\mathbf{p}}}$;
(ii) if G is bounded, the set of invariant measures for $\mathcal{P}$ is $\{\pi \in \mathscr{P}(G) \mid \operatorname{supp} \pi \subset C\}$;
(iii) if G is bounded,

$$(\forall x \in G) \quad \Psi(\delta_x) = \|x - T_1(x)\| \qquad \textit{and} \qquad (51a)$$

$$\frac{1}{\sqrt{p}} \inf_{z\in C} \|x-z\| \leq \inf_{z\in C} \|x-z\|_{\mathbf{p}} = d_{W_{2,\mathbf{p}}}(\delta_x, \operatorname{inv}\mathcal{P}) = d_{W_{2,\mathbf{p}}}\left(\delta_x, \Psi^{-1}(0)\right). \qquad (51b)$$

Proof. (i). For a random variable $X \sim \mu$, we have $T_\xi X = \Phi(X,\xi) \sim \mu\mathcal{P}$, and for a random variable $Y \sim \pi \in \operatorname{inv}\mathcal{P}$ we have $T_\xi Y = \Phi(Y,\xi) \sim \pi\mathcal{P} = \pi$, so

$$\begin{aligned}
d_{W_{2,\mathbf{p}}}(\mu\mathcal{P}, \pi) &= \left(\inf_{\gamma\in\mathcal{C}(\mu\mathcal{P},\pi)} \int_{G\times G} \|x^+ - y\|_{\mathbf{p}}^2 \gamma(dx^+, dy) \right)^{1/2} \\
&\leq \left(\inf_{\gamma\in\mathcal{C}(\mu,\pi)} \int_{G\times G} \mathbb{E}\left[\|T_\xi x - y\|_{\mathbf{p}}^2\right] \gamma(dx,dy) \right)^{1/2} \\
&< \left(\inf_{\gamma\in\mathcal{C}(\mu,\pi)} \int_{G\times G} \|x - y\|_{\mathbf{p}}^2 \gamma(dx,dy) \right)^{1/2} \\
&= d_{W_{2,\mathbf{p}}}(\mu,\pi) \qquad \forall \pi \in \operatorname{inv}\mathcal{P}, \forall \mu \in \mathscr{P}_2(G) \setminus \operatorname{inv}\mathcal{P}, \qquad (52)
\end{aligned}$$

where the last inequality follows from the assumption that T_i defined by (3) are a paracontractions in expectation with respect to the weighted norm $\|\cdot\|_{\mathbf{p}}$. This establishes that $\mathcal{P}$ is a paracontraction in the $d_{W_{2,\mathbf{p}}}$ metric as claimed.

(ii). Our proof follows the proof of [18, Lemma 3.3]. It is clear that $\pi \in \mathscr{P}(G)$ with $\operatorname{supp}\pi \subset C \subset G$ is invariant, since $p(x, \{x\}) = \mathbb{P}(T_\xi x \in \{x\}) = \mathbb{P}(x \in \mathsf{Fix}\, T_\xi) = 1$ for all $x \in C$ and hence $\pi\mathcal{P}(A) = \int_C p(x, A)\pi(\mathrm{d}\, x) = \pi(A)$ for all $A \in \mathcal{B}(G)$.

Suppose, on the other hand, that $\operatorname{supp}\pi \setminus C \neq \emptyset$ for some $\pi \in \operatorname{inv}\mathcal{P}$ with $\operatorname{supp}\pi \subset G$. Then due to compactness of $\operatorname{supp}\pi$ (it is closed in the compact set G) we can find $s \in \operatorname{supp}\pi$ maximizing the continuous function $d(\cdot, C) := \inf_{z\in C} \|\cdot - z\|$ on G. So $d_{\max} = d(s, C) > 0$. We show that this leads only to contradictions, so the assumption of the existence of such a π must be false.

Define the set of points being more than $d_{\max} - \epsilon$ away from C:

$$K(\epsilon) := \{x \in G \mid d(x, C) > d_{\max} - \epsilon\}, \qquad \epsilon \in (0, d_{\max}).$$

This set is measurable, i.e. $K(\epsilon) \in \mathcal{B}(G)$, because it is open. Let $M(\epsilon)$ be the event in the sigma algebra $\mathcal{F}$, that $T_\xi s$ is at least ϵ closer to C than s, i.e.

$$M(\epsilon) := \left\{\omega \in \Omega \;\middle|\; d(T_{\xi(\omega)}s, C) \leq d_{\max} - \epsilon\right\}.$$

There are two possibilities, either there is an $\epsilon \in (0, d_{\max})$ with $\mathbb{P}(M(\epsilon)) > 0$ or no such ϵ exists. In the latter case we have $\mathbb{E}\left[d(T_\xi s, C)\right] = d_{\max} = d(s, C)$ since T_i is a paracontraction in expectation. By compactness of C there exists $c \in C$ such that $0 < d_{\max} = \|s - c\|$. Hence the probability of the set of $\omega \in \Omega$ such that $s \notin \mathsf{Fix}\, T_{\xi(\omega)}$ is positive and so $\mathbb{E}\left[d(T_{\xi(\omega)}s, C)\right] \leq \mathbb{E}\left[\|T_{\xi(\omega)}s - c\|\right] < \|s - c\|$ - a contradiction.

Suppose next that there is an $\epsilon \in (0, d_{\max})$ with $\mathbb{P}(M(\epsilon)) > 0$. In view of continuity of the mappings T_i around s, $i \in \mathbb{I}$, define

$$A_n := \left\{\omega \in M(\epsilon) \;\middle|\; \|(T_{\xi(\omega)}x - T_{\xi(\omega)}s\| \leq \tfrac{\epsilon}{2} \quad \forall x \in \mathbb{B}(s, \tfrac{1}{n})\right\} \quad (n \in \mathbb{N}).$$

It holds that $A_n \subset A_{n+1}$ and $\mathbb{P}(\bigcup_n A_n) = \mathbb{P}(M(\epsilon))$. So in particular there is an $m \in \mathbb{N}$, $m \geq 2/\epsilon$ with $\mathbb{P}(A_m) > 0$. For all $x \in \mathbb{B}(s, \frac{1}{m})$ and all $\omega \in A_m$ we have

$$d(T_{\xi(\omega)}x, C) \leq \|T_{\xi(\omega)}x - T_{\xi(\omega)}s\| + d(T_{\xi(\omega)}s, C) \leq d_{\max} - \frac{\epsilon}{2},$$

which means $T_{\xi(\omega)}x \in G \setminus K(\frac{\epsilon}{2})$. Hence, in particular we conclude that

$$p(x, K(\tfrac{\epsilon}{2})) < 1 \quad \forall x \in \mathbb{B}(s, \tfrac{1}{m}).$$

Since $p(x, K(\epsilon)) = 0$ for $x \in G$ with $d(x, C) \leq d_{\max} - \epsilon$ by the assumption that T_i is a paracontraction in expectation, it holds by invariance of π that

$$\pi(K(\epsilon)) = \int_G p(x, K(\epsilon))\pi(\mathrm{d}\, x) = \int_{K(\epsilon)} p(x, K(\epsilon))\pi(\mathrm{d}\, x).$$

It follows, then, that

$$\begin{aligned}
\pi(K(\tfrac{\epsilon}{2})) &= \int_{K(\frac{\epsilon}{2})} p(x, K(\tfrac{\epsilon}{2}))\pi(\mathrm{d}\, x) \\
&= \int_{\mathbb{B}(s, \frac{1}{m})} p(x, K(\tfrac{\epsilon}{2}))\pi(\mathrm{d}\, x) + \int_{K(\frac{\epsilon}{2})\setminus\mathbb{B}(s, \frac{1}{m})} p(x, K(\tfrac{\epsilon}{2}))\pi(\mathrm{d}\, x) \\
&< \pi(\mathbb{B}(s, \tfrac{1}{m})) + \pi(K(\tfrac{\epsilon}{2}) \setminus \mathbb{B}(s, \tfrac{1}{m})) = \pi(K(\tfrac{\epsilon}{2}))
\end{aligned}$$

which leads again to a contradiction. So the assumption that $\operatorname{supp}\pi \setminus C \neq \emptyset$ is false, i.e. $\operatorname{supp}\pi \subset C$ as claimed.

(iii). By part (ii), $\operatorname{inv}\mathcal{P} = \{\pi \in \mathscr{P}(G) \mid \operatorname{supp}\pi \subset C\}$, so (49) holds with equality, and $\Psi(\mu) = 0$ if and only if $\operatorname{supp}\mu \subset C$, hence writing (50) pointwise (i.e., for $\mu = \delta_x$) reduces the expression to

$$\inf_{z\in C} \|x - z\|_{\mathbf{p}} = d_{W_{2,\mathbf{p}}}(\delta_x, \operatorname{inv}\mathcal{P}) = d_{W_{2,\mathbf{p}}}(\delta_x, \Psi^{-1}(0)).$$

The representation (51) then follows from $\overline{p} := \max_j\{p_j\}$. $\square$

4 Convergence

Contractive Markov operators have been extensively, almost exclusively, studied. When the update function Φ is a contraction in expectation, then [20, Theorem 2.12] shows that the corresponding Markov operator $\mathcal{P}$ is α-fne, and the sequence of measures (μ_k) converges Q-linearly (geometrically) to an invariant measure from any starting measure $\mu_0 \in \mathscr{P}(\mathcal{E})$. When the mappings T_i are only α-firmly nonexpansive on $\mathcal{E}$, then μ_k converges in the Prokhorov-Levi metric to an invariant measure from any initial measure [19, Theorem 2.9]. To obtain generic (weak) convergence of the iterates μ_k one must show that the sequence is *tight*. This has been established for Markov operators with nonexpansive update functions [19, Lemma 3.19]. We skirt a study of whether tightness can be established under the assumption that the update functions $\Phi(x, i)$ are only nonexpansive in expectation; we suspect, however, that this is not the case.

4.1 Generic proto-convergence

We establish a few properties that are cornerstones of a generic global convergence analysis. In particular, we show that when the Markov operator is α-fne (which, as shown above, does not require that all the mappings T_i be α-fne) this property together with an additional assumption about the decay of the invariant Markov transport discrepancy yields boundedness and *asymptotic regularity* of the sequence of measures.

Proposition 6 (asymptotic regularity)**.** *Let the Markov operator $\mathcal{P} : \mathscr{P}_2(\mathcal{E}) \to \mathscr{P}_2(\mathcal{E})$ with update functions $\Phi(x, i)$ possess at least one invariant measure and be pointwise α-fne in measure at all $\pi \in \operatorname{inv} \mathcal{P}$. If the invariant Markov transport discrepancy satisfies*

$$\exists c > 0 : \quad \Psi(\mu) \geq c d_{W_{2,\mathbf{p}}}(\mu, \mu\mathcal{P}) \quad \forall \mu \in \mathscr{P}_2(\mathcal{E}), \tag{53}$$

then the sequence $(\mu_k)_{k\in\mathbb{N}}$ defined by $\mu_{k+1} = \mu_k\mathcal{P}$ for any $\mu_0 \in \mathscr{P}_2(\mathcal{E})$ is bounded and asymptotically regular, i.e. satisfies $d_{W_{2,\mathbf{p}}}(\mu_k, \mu_{k+1}) \to 0$.

Proof. Note that (53) implies that there is a $c > 0$ such that

$$c^2 d_{W_{2,\mathbf{p}}}(\mu, \mu\mathcal{P})^2 \leq \int_{\mathcal{E}\times\mathcal{E}} \mathbb{E}\left[\psi_{\mathbf{p}}(x, y, \Phi(x, \xi), \Phi(y, \xi))\right] \gamma(dx, dy) \quad \forall \pi \in \operatorname{inv} \mathcal{P},\ \forall \gamma \in C_*(\pi, \mu).$$

This together with the assumption that $\mathcal{P}$ is α-fne yields

$$\begin{aligned} 0 \leq d_{W_{2,\mathbf{p}}}(\mu\mathcal{P}, \pi)^2 &\leq d_{W_{2,\mathbf{p}}}(\mu, \pi)^2 - \frac{1-\alpha}{\alpha} \int_{\mathcal{E}\times\mathcal{E}} \mathbb{E}\left[\psi_{\mathbf{p}}(x, y, \Phi(x, \xi), \Phi(x, \xi))\right] \gamma(dx, dy) \\ &\leq d_{W_{2,\mathbf{p}}}(\mu, \pi)^2 - \frac{1-\alpha}{\alpha} c^2 d_{W_{2,\mathbf{p}}}(\mu, \mu\mathcal{P})^2 \quad \forall \pi \in \operatorname{inv} \mathcal{P},\ \forall \gamma \in C_*(\pi, \mu). \end{aligned} \tag{54}$$

Applying (54) to the sequence of measures generated by $\mu_{k+1} = \mu_k\mathcal{P}$ with $\mu_0 \in \mathscr{P}_2(\mathcal{E})$ yields

$$\frac{1-\alpha}{\alpha} c^2 \sum_{k=1}^{N} d_{W_{2,\mathbf{p}}}(\mu_k, \mu_{k+1})^2 \leq d_{W_{2,\mathbf{p}}}(\mu_0, \pi)^2 \quad \forall \pi \in \operatorname{inv} \mathcal{P}, \forall N \in \mathbb{N}.$$

Letting $N \to \infty$ establishes that the left hand side is summable, hence $\liminf d_{W_{2,\mathbf{p}}}(\mu_k, \mu_{k+1}) = 0$. But $\mathcal{P}$ is also pointwise nonexpansive at all $\pi \in \operatorname{inv} \mathcal{P}$ since it is pointwise α-fne there, so $d_{W_{2,\mathbf{p}}}(\mu_k, \pi) \leq d_{W_{2,\mathbf{p}}}(\mu_0, \pi)$ for all k and $d_{W_{2,\mathbf{p}}}(\mu_k, \mu_{k+1}) \to 0$; i.e. the sequence is bounded and asymptotically regular as claimed. □

In the next section we pursue a quantitative local convergence analysis under the assumption of *metric subregularity* of the invariant Markov transport discrepancy.

Recall the inverse mapping $\Psi^{-1}(y) := \{\mu \mid \Psi(\mu) = y\}$, which clearly can be set-valued. It is important to keep in mind that an invariant measure need not correspond to a fixed point of any individual mapping T_i, unless these have common fixed points. See [19, 20] instances of this. We require that the invariant Markov transport discrepancy Ψ takes the value 0 at μ if and only if $\mu \in \operatorname{inv}\mathcal{P}$, and is *gauge metrically subregular for* 0 *relative to* $\mathscr{P}_2(G)$ *on* $\mathscr{P}_2(G)$:

$$d_{W_{2,\mathbf{p}}}(\mu, \operatorname{inv}\mathcal{P}) = d_{W_{2,\mathbf{p}}}(\mu, \Psi^{-1}(0)) \leq \rho(\Psi(\mu)) \quad \forall \mu \in \mathscr{P}_2(G). \tag{55}$$

Here $d_{W_{2,\mathbf{p}}}(\mu, \operatorname{inv}\mathcal{P}) = \inf_{\pi \in \operatorname{inv}\mathcal{P}} d_{W_{2,\mathbf{p}}}(\mu, \pi)$, and $\rho : [0,\infty) \to [0,\infty)$ is a *gauge function*: it is continuous, strictly increasing with $\rho(0) = 0$, and $\lim_{t\to\infty} \rho(t) = \infty$. The gauge of metric subregularity ρ is constructed implicitly from another nonnegative function $\theta_{\tau,\epsilon} : [0,\infty) \to [0,\infty)$ with parameters $\tau > 0$ and $\epsilon \geq 0$ satisfying

$$(i)\ \theta_{\tau,\epsilon}(0) = 0; \quad (ii)\ 0 < \theta_{\tau,\epsilon}(t) < t\ \forall t \in (0, \bar{t}] \text{ for some } \bar{t} > 0 \tag{56}$$

and

$$\rho\left(\left(\frac{(1+\epsilon)t^2 - (\theta_{\tau,\epsilon}(t))^2}{\tau}\right)^{1/2}\right) = t \quad \iff \quad \theta_{\tau,\epsilon}(t) = \left((1+\epsilon)t^2 - \tau\left(\rho^{-1}(t)\right)^2\right)^{1/2} \tag{57}$$

for $\tau > 0$ fixed. In the next theorem the parameter ϵ is exactly the violation in aα-fne mappings; the parameter τ is directly computed from the constant α.

In preparation for the results that follow, we will require at least one of the additional assumptions on θ.

Assumption 2. *The gauge* $\theta_{\tau,\epsilon}$ *satisfies* (56) *and at least one of the following holds.*

(a) $\theta_{\tau,\epsilon}$ *satisfies*

$$\theta^{(k)}_{\tau,\epsilon}(t) \to 0 \text{ as } k \to \infty\ \forall t \in (0, \bar{t}), \tag{58}$$

and the sequence (μ_k) *is Fejér monotone with respect to* $\operatorname{inv}\mathcal{P} \cap \mathscr{P}_2(G)$*, i.e.*

$$d_{W_{2,\mathbf{p}}}(\mu_{k+1}, \pi) \leq d_{W_{2,\mathbf{p}}}(\mu_k, \pi) \quad \forall k \in \mathbb{N}, \forall \pi \in \operatorname{inv}\mathcal{P} \cap \mathscr{P}_2(G); \tag{59}$$

(b) $\theta_{\tau,\epsilon}$ *satisfies*

$$\sum_{j=1}^{\infty} \theta^{(j)}_{\tau,\epsilon}(t) < \infty\ \forall t \in (0, \bar{t}) \tag{60}$$

where $\theta^{(j)}_{\tau,\epsilon}$ *denotes the* j*-times composition of* $\theta_{\tau,\epsilon}$*.*

In the case of linear metric subregularity this becomes

$$\rho(t) = \kappa t \quad \iff \quad \theta_{\tau,\epsilon}(t) = \left((1+\epsilon) - \frac{\tau}{\kappa^2}\right)^{1/2} t \quad (\kappa \geq \sqrt{\tfrac{\tau}{(1+\epsilon)}}).$$

The condition $\kappa \geq \sqrt{\frac{\tau}{(1+\epsilon)}}$ is not a real restriction since, if (55) is satisfied for some $\kappa' > 0$, then it is satisfied for all $\kappa \geq \kappa'$. The conditions in (56) in this case simplify to $\theta_{\tau,\epsilon}(t) = \gamma t$ where

$$0 < \gamma := 1 + \epsilon - \frac{\tau}{\kappa^2} < 1 \quad \iff \quad \sqrt{\tfrac{\tau}{(1+\epsilon)}} \leq \kappa \leq \sqrt{\tfrac{\tau}{\epsilon}}. \tag{61}$$

In other words, $\theta_{\tau,\epsilon}(t)$ satisfies Assumption 2(b). The weaker Assumption 2(a) is used to characterize sublinear convergence.

Theorem 7 (convergence rates)**.** *Let $G \subset \mathcal{E}$ be compact. Let $T_i : G \to G$ satisfy the assumptions of Theorem 1 for all $i \in \mathbb{I}$. Assume furthermore that there is at least one $\pi \in \operatorname{inv}\mathcal{P} \cap \mathscr{P}_2(G)$ where $\mathcal{P}$ is the Markov operator associated with T_i. If, in addition, Ψ satisfies* (55) *with gauge ρ given implicitly by* (57) *in terms of $\theta_{\tau,\epsilon}$ where $\tau = (1-\overline{\alpha})/\overline{\alpha}$, $\epsilon = \overline{p\epsilon}$ as in Theorem 1, then for any $\mu_0 \in \mathscr{P}_2(G)$ the distributions μ_k of the iterates of Algorithm 1 satisfy*

$$d_{W_{2,\mathbf{p}}}(\mu_{k+1}, \operatorname{inv}\mathcal{P}) \le \theta_{\tau,\epsilon}\left(d_{W_{2,\mathbf{p}}}(\mu_k, \operatorname{inv}\mathcal{P})\right) \quad \forall k \in \mathbb{N}. \tag{62}$$

In addition, let τ and ϵ be such that $\theta_{\tau,\epsilon}$ satisfies (56) *where $t_0 := d_{W_{2,\mathbf{p}}}(\mu_0, \operatorname{inv}\mathcal{P}) < \bar{t}$ for all $\mu_0 \in \mathscr{P}_2(G)$, and let at least one of the conditions in Assumption 2 hold. Then $\mu_k \to \pi^{\mu_0} \in \operatorname{inv}\mathcal{P} \cap \mathscr{P}_2(G)$ in the $d_{W_{2,\mathbf{p}}}$ metric with rate $O\left(\theta_{\tau,\epsilon}^{(k)}(t_0)\right)$ in case Assumption 2*(a) *and with rate $O(s_k(t_0))$ for $s_k(t) := \sum_{j=k}^{\infty} \theta_{\tau,\epsilon}^{(j)}(t)$ in case Assumption 2*(b)*.*

Proof. First we note that, since G is assumed to be compact, $\mathcal{P}$ is a nonempty self-mapping on $\mathscr{P}_2(G)$ and $\mathscr{P}_2(G)$ is locally compact ([1, Remark 7.19]). By Theorem 1, the update function Φ is aα-fne in expectation with respect to the weighted norm $\|\cdot\|_{\mathbf{p}}$ with constant $\overline{\alpha}$ and violation $\overline{p\epsilon}$. The statement is an extension of [20, Theorem 2.6], which establishes (62) and convergence under Assumption 2(b).

To establish convergence under Assumption 2(a), we show first that $d_{W_{2,\mathbf{p}}}(\mu_k, S) \to 0$ where, to reduce notational clutter we define $S := \operatorname{inv}\mathcal{P} \cap \mathscr{P}_2(G)$. Indeed, let $\pi \in S$ and define $d_k^\pi := d_{W_{2,\mathbf{p}}}(\mu_k, \pi)$. Since $d_{k+1}^\pi \le d_k^\pi$ for all k, this establishes that the sequence $(d_k^\pi)_{k\in\mathbb{N}}$ is bounded and monotone non-increasing, therefore convergent. Noting that $d_{W_{2,\mathbf{p}}}(\mu_k, S) \le d_k^\pi$ for all k and any fixed $\pi \in S$, this also shows that $d_{W_{2,\mathbf{p}}}(\mu_k, S)$ converges. The inequality (62) only requires assumption (56), and this together with assumption (58) yields

$$d_{W_{2,\mathbf{p}}}(\mu_k, S) \le \theta^{(k)}(t_0) \to 0 \text{ as } k \to \infty.$$

Since $\mathscr{P}_2(G)$ is locally compact and $\mathcal{P}$ is Feller since T_i is continuous for all i, $\operatorname{inv}\mathcal{P}$ is closed [16]; so for every $k \in \mathbb{N}$ the infimum in $d_{W_{2,\mathbf{p}}}(\mu_k, S)$ is attained at some π_k. Now, for such a π_k we have, again by Fejér monotonicity, that

$$d(\mu_l, \mu_k) \le d(\mu_l, \overline{x}^k) + d(\mu_k, \overline{x}^k) \le d(\mu_{l-1}, \overline{x}^k) + d(\mu_k, \overline{x}^k) \le \cdots \le 2d(\mu_k, S).$$

Since the right hand side converges to 0 as $k \to \infty$ this shows that the sequence is a Cauchy sequence on $(\mathscr{P}_2(G), W_2)$ – a separable complete metric space [42, Theorem 6.9] – and therefore convergent to some probability measure $\pi^{\mu_0} \in \mathscr{P}_2(G)$. The Markov operator $\mathcal{P}$ is Feller and when a Feller Markov chain converges in distribution, it does so to an invariant measure: $\pi^{\mu_0} \in \operatorname{inv}\mathcal{P}$ (see [16, Theorem 1.10]). □

Note that

$$\forall \overline{\epsilon} > \epsilon, \forall t \in [0, t_0], \quad \theta_{\tau,\overline{\epsilon}}(t) > \theta_{\tau,\epsilon}(t).$$

It is common in optimization algorithms to encounter mappings whose violation ϵ can be controlled by choosing a step length parameter small enough; the gradient descent operator is just such a mapping. This means that, if condition (56) and at least one of (a) or (b) in Assumption 2 is satisfied for *some* ϵ, and the violation of the fixed point mappings T_i can be made arbitrarily small, then Theorem 7 guarantees convergence with rate given by either $O(\theta_{\tau,\epsilon}^{(k)})$ in case (a) or $O(s_k(t_0))$ in case (b) for small enough step sizes on small enough neighborhoods of a fixed point.

4.2.1 Special Case: consistent stochastic feasibility

Recall that, when $\operatorname{inv}\mathcal{P} = \mathscr{C}$ defined by (47) (which, by Theorem 5 holds when $\mathcal{P}$ is a paracontraction in measure) the relation (49) holds with equality, so condition (55) simplifies to

$$\begin{aligned} d_{W_{2,\mathbf{p}}}(\mu, \operatorname{inv}\mathcal{P}) = d_{W_{2,\mathbf{p}}}(\mu, \Psi^{-1}(0)) \quad &\le \quad \rho(\Psi(\mu)) \quad \forall \mu \in \mathscr{P}_2(G) \\ &\Longleftrightarrow \\ \left(\int_G \inf_{z\in C} \|x - z\|_{\mathbf{p}}^2 \mu(dx)\right)^{1/2} \quad &\le \quad \rho\left(\left(\int_G \|x - T_1 x\|^2 \mu(dx)\right)^{1/2}\right) \qquad \forall \mu \in \mathscr{P}_2(G). \end{aligned} \tag{63}$$

Writing this pointwise (i.e., for $\mu = \delta_x$) reduces the expression to

$$\inf_{z\in C} \|x - z\|_{\mathbf{p}} \le \rho\left(\|x - T_1 x\|\right) \qquad \forall x \in G, \tag{64}$$

whereby, recalling that $\overline{p} = \max_j\{p_j\}$, (55) yields

$$\tfrac{1}{\sqrt{\overline{p}}} d(x, C) \le \inf_{z\in C} \|x - z\|_{\mathbf{p}} \le \rho\left(\|x - T_1(x)\|\right) \qquad \forall x \in G. \tag{65}$$

This is recognizable as a slight generalization of the error bound studied by Luo and Tseng [28].

The next result shows that, for paracontractions, metric subregularity is *automatically* satisfied by Markov chains that are *gauge monotone* with respect to $\operatorname{inv}\mathcal{P}$. Let $(X_k)_{k\in\mathbb{N}}$ be a sequence of random variables on the closed subset $G \subset \mathcal{E}$ generated by Algorithm 1, and let $(\mu_k)_{k\in\mathbb{N}}$ be the corresponding sequence of distributions. Let $\operatorname{inv}\mathcal{P}$ be nonempty and let the continuous mapping $\theta : \mathbb{R}_+ \to \mathbb{R}_+$ satisfy

$$(i)\ \theta(0) = 0; \quad (ii)\ 0 < \theta(t) \le t\ \forall t \in (0, \overline{t}) \text{ for some } \overline{t} > 0. \tag{66}$$

This is obviously the same as (56) but without the parameters since in this case $\epsilon = 0$ and τ is just some scaling. For $t_0 := d_{W_{2,\mathbf{p}}}(\mu_0, \operatorname{inv}\mathcal{P})$, the sequence $(\mu_k)_{k\in\mathbb{N}}$ is said to be *gauge monotone relative to* $\operatorname{inv}\mathcal{P}$ *with rate* θ whenever

$$d_{W_{2,\mathbf{p}}}(\mu_{k+1}, \operatorname{inv}\mathcal{P}) \le \theta\left(d_{W_{2,\mathbf{p}}}(\mu_k, \operatorname{inv}\mathcal{P})\right)\ \forall k \in \mathbb{N} \tag{67}$$

where θ satisfies (66) with $t_0 < \overline{t}$. The sequence $(\mu_k)_{k\in\mathbb{N}}$ is said to be *linearly monotone relative to* $\operatorname{inv}\mathcal{P}$ with rate c if (67) is satisfied for $\theta(t) \le c\cdot t$ for all $t \in [0, t_0]$ and some constant $c \in [0, 1]$.

A Markov chain $(X_k)_{k\in\mathbb{N}}$ that converges to some law $\pi^{\mu_0} \in \mathscr{P}_2(G)$ is said to converge *gauge monotonically* in distribution whenever the corresponding sequence of distributions $(\mu_k)_{k\in\mathbb{N}}$ is gauge monotone with gauge θ satisfying (66) with $d_{W_{2,\mathbf{p}}}(\mu_0, \operatorname{inv}\mathcal{P}) \le \overline{t}$.

Proposition 8 (gauge monotonic paracontractions in measure converge to invariant measures)**.** *Let $G \subset \mathcal{E}$ be compact. Let the Markov operator corresponding to Algorithm* (1)*, $\mathcal{P} : \mathscr{P}_2(G) \to \mathscr{P}_2(G)$, be a paracontraction with respect to the metric $d_{W_{2,\mathbf{p}}}$. For a fixed $\mu_0 \in \mathscr{P}_2(G)$, let the sequence of measures $(\mu_k)_{k\in\mathbb{N}}$ corresponding to the iterates of Algorithm 1 be gauge monotone relative to* $\operatorname{inv}\mathcal{P}$ *with rate θ satisfying* (66) *where $t_0 := d_{W_{2,\mathbf{p}}}(\mu_0, \operatorname{inv}\mathcal{P}) < \overline{t}$. Suppose furthermore that at least one of the conditions* (a) *or* (b) *of Assumption 2 are satisfied (replacing $\theta_{\tau,\epsilon}$ with θ). Then $(\mu_k)_{k\in\mathbb{N}}$ converges gauge monotonically with respect to $d_{W_{2,\mathbf{p}}}$ to some $\pi^{\mu_0} \in \operatorname{inv}\mathcal{P} \cap \mathscr{P}_2(G)$ with rate $O(\theta^{(k)}(t_0))$ if Assumption 2*(a) *holds, and in the case of Assumption 2*(b) *with rate $O(s_k(t_0))$ for $s_k(t) := \sum_{j=k}^{\infty} \theta^{(j)}(t)$ and $t_0 := d_{W_{2,\mathbf{p}}}(\mu_0, \operatorname{inv}\mathcal{P})$. Moreover, $\operatorname{supp}\pi^{\mu_0} \subset C$ for C defined by* (42)*.*

Proof. In both cases, the proof of convergence with the respective rates follows exactly the proof of convergence in Theorem 7. For the last statement, Theorem 5(ii) establishes that $\operatorname{supp}\pi^{\mu_0} \subset C$, which completes the proof. □

The following is a generalization of [18, Theorem 3.15].

Theorem 9 (necessity of metric subregularity for monotone sequences)**.** *Let $G \subset \mathcal{E}$ be compact. Let the Markov operator corresponding to Algorithm* (1)*, $\mathcal{P} : \mathscr{P}_2(G) \to \mathscr{P}_2(G)$, be a paracontraction with respect to the weighted Wasserstein metric $d_{W_{2,\mathbf{p}}}$. Suppose all sequences $(\mu_k)_{k\in\mathbb{N}}$ corresponding to Algorithm 1 and initialized in $\mathscr{P}_2(G)$ are gauge monotone relative to* $\operatorname{inv}\mathcal{P}$ *with rate θ satisfying* (66) *and at least one of the conditions in Assumption 2. Suppose, in addition, that* $(\operatorname{Id} - \theta)^{-1}(\cdot)$ *is continuous on $\mathbb{R}_+$, strictly increasing, and* $(\operatorname{Id} - \theta)^{-1}(0) = 0$*. Then Ψ defined by* (41) *is gauge metrically subregular for* 0 *relative to $\mathscr{P}_2(G)$ on $\mathscr{P}_2(G)$ with gauge $\rho(\cdot) = (\operatorname{Id} - \theta)^{-1}(\cdot)$, i.e. Ψ satisfies* (55)*.*

Proof. If the sequence $(\mu_k)_{k\in\mathbb{N}}$ is gauge monotone relative to $\operatorname{inv}\mathcal{P}$ with rate θ satisfying (66) and at least one of the conditions in Assumption 2, then by the triangle inequality

$$\begin{aligned} d_{W_{2,\mathbf{p}}}(\mu_{k+1}, \mu_k) &\ge d_{W_{2,\mathbf{p}}}(\mu_k, \overline{\mu}_{k+1}) - d_{W_{2,\mathbf{p}}}(\mu_{k+1}, \overline{\mu}_{k+1}) \\ &\ge d_{W_{2,\mathbf{p}}}(\mu_k, \operatorname{inv}\mathcal{P}) - d_{W_{2,\mathbf{p}}}(\mu_{k+1}, \operatorname{inv}\mathcal{P}) \\ &\ge d_{W_{2,\mathbf{p}}}(\mu_k, \operatorname{inv}\mathcal{P}) - \theta\left(d_{W_{2,\mathbf{p}}}(\mu_k, \operatorname{inv}\mathcal{P})\right) \ge 0 \quad \forall k \in \mathbb{N}, \end{aligned} \tag{68}$$

where $\overline{\mu}_{k+1}$ is a metric projection of μ_{k+1} onto $\operatorname{inv}\mathcal{P}$ (exists since $\operatorname{inv}\mathcal{P}$ is closed in $\mathscr{P}_2(G)$). On the other hand, by Theorem 5(ii), inequality (49) is tight, so $\Psi^{-1}(0) = \operatorname{inv}\mathcal{P}$ and

$$\begin{aligned}\Psi(\mu_k) = \left(\int_G \|x - T_1 x\|^2 \mu_k(dx)\right)^{1/2} &\geq \inf_{\gamma \in C(\mu_k \mathcal{P}, \mu_k)} \left(\int_{G\times G} \|x-y\|_{\mathbf{p}}^2 \gamma(dx,dy)\right)^{1/2} \\ &= d_{W_{2,\mathbf{p}}}(\mu_{k+1}, \mu_k) \quad \forall k \in \mathbb{N}. \end{aligned} \tag{69}$$

Combining (68) and (69) yields

$$d(0, \Psi(\mu_k)) = \Psi(\mu_k) \geq d_{W_{2,\mathbf{p}}}(\mu_k, \Psi^{-1}(0)) - \theta\left(d_{W_{2,\mathbf{p}}}(\mu_k, \Psi^{-1}(0))\right) \quad \forall k \in \mathbb{N}. \tag{70}$$

By assumption $(\mathrm{Id} - \theta)^{-1}(\cdot)$ is continuous on $\mathbb{R}_+$, strictly increasing, and $(\mathrm{Id} - \theta)^{-1}(0) = 0$, so

$$(\mathrm{Id} - \theta)^{-1}\left(d(0, \Psi(\mu_k))\right) \geq d_{W_{2,\mathbf{p}}}(\mu_k, \Psi^{-1}(0)) \quad \forall k \in \mathbb{N}. \tag{71}$$

Since this holds for *any* sequence $(\mu_k)_{k\in\mathbb{N}}$ initialized in $\mathscr{P}_2(G)$ and these converge by Proposition 8 to points in $\operatorname{inv}\mathcal{P} \cap \mathscr{P}_2(G)$, we conclude that Ψ is metrically subregular for 0 on $\mathscr{P}_2(G)$ with gauge $\rho = (\mathrm{Id} - \theta)^{-1}$. □

5 Block-Stochastic Splitting for Composite Optimization

We return now to stochastic blockwise methods for solving (1). It is already understood that the critical points of $f + \sum_{j=1}^m g_j$, denoted $\operatorname{crit}(f + \sum_{j=1}^m g_j)$, are fixed points of the deterministic, non-block versions of Algorithms 2 and 3; and fixed points of the deterministic, non-block versions of these algorithms are invariant distributions corresponding to iterates of these same stochastic blockwise algorithms. When $\operatorname{inv}\mathcal{P} = \mathscr{C}$ defined by (47), then in fact any $\overline{x} \in C := \{x \mid \mathbb{P}(x \in \mathsf{Fix}\, T_\xi) = 1\}$ is almost surely at least a stationary point. This leads to the following elementary observations.

Lemma 10. *Let T_i defined by either* (17) *(if f is differentiable) or* (19) *be single-valued on $\mathcal{E}$ and let $\mathcal{P}$ be the Markov operator with update function T_i. Then $\operatorname{crit}(f + \sum_{j=1}^m g_j) \subset S := \bigcup_{\pi \in \operatorname{inv}\mathcal{P}} \operatorname{supp}\pi$, and if f and g_j $(j = 1, \ldots, m)$ are convex, then $\overline{x} \in C$ if and only if $\overline{x} \in \operatorname{crit}(f + \sum_{j=1}^m g_j)$ almost surely.*

5.1 Regularity

In this section we determine the regularity of the blockwise mappings T_i for the two cases (17) and (19). In Theorem 1, the regularity constants ϵ_i and α_i are bounded above by the constants of T_1, which is the mapping including all of the blocks. It suffices, then, to determine the regularity of T_1 for the two cases (17) and (19).

Proposition 11 (regularity of partial resolvents)**.** *For $j = 1, 2, \ldots, m$, for each vector of parameters $x \in G \subset \mathcal{E}$, let $f_j(\cdot; x) : G_j \subset \mathcal{E}_j \to (-\infty, +\infty]$ defined by* (16) *be subdifferentially regular with subdifferentials satisfying*

$$\begin{aligned} \exists \tau_{f_j} \geq 0 :\ & \forall x \in G, \forall u_j, v_j \in G_j,\ \forall z_j \in t_j \partial f_j(u_j; x), w_j \in t_j \partial f_j(v_j; x), \\ & -\tfrac{\tau_{f_j}}{2} \|(u_j + z_j) - (v_j + w_j)\|^2 \\ & \qquad \leq \langle z_j - w_j,\ u_j - v_j \rangle . \end{aligned} \tag{72}$$

For $f_t(u; x) := \sum_{j=1}^m t_j f_j(u_j; x)$, the resolvent $J_{\partial f_t, 1}$ is aα-fne with constant $\alpha_f = 1/2$ and violation $\tau_f = \max_j\{\tau_{f_j}\}$ on G. If f_j is convex on $\mathcal{E}_j$ for each $j = 1, 2, \ldots, m$, then $J_{\partial f_t, 1}$ is α-fne with constant $\alpha_f = 1/2$ and no violation on $\mathcal{E}$.

Condition (72) generalizes the notion of hypomonotonicity [38] and is satisfied by any *prox-regular* function.

Proof. By [27, Proposition 2.3(iv)], condition (72) is equivalent to $J_{\partial f_j, t_j}$ being aα-fne on G_j with constant $\alpha_{f_j} = 1/2$ and violation τ_{f_j}. Extending this, for $f_t(x) := \sum_{j=1}^m t_j f_j(u_j; x)$ we have $\partial f_t(u; x) = \left[[t_1 \partial_{u_1} f_1(u_1; x)]^T, \ldots, [t_m \partial_{u_m} f_m(u_m; x)]^T \right]^T$ and

$$\begin{aligned}
&\forall v, u \in\subset G, \text{ for } z := \partial f_t(u; x), w := \partial f_t(v; x),\\
&\langle z - w,\ u - v \rangle = \sum_{j=1}^m \langle z_j - w_j,\ u_j - v_j \rangle_{G_j}\\
&\qquad\qquad \geq \sum_{j=1}^m \tfrac{-\tau_{f_j}}{2} \|(u_j + z_j) - (v_j + w_j)\|^2_{G_j}\\
&\qquad\qquad \geq \frac{-\max_j\{\tau_{f_j}\}}{2} \sum_{j=1}^m \|(u_j + z_j) - (v_j + w_j)\|^2_{G_j}\\
&\qquad\qquad = \tfrac{-\tau_f}{2} \|(u + z) - (v + w)\|^2 .
\end{aligned}$$

Application of [27, Proposition 2.3(iv)] to f_t establishes the claim. The convex statement follows from monotonicity of the gradient. □

The following corollary is just the specialization of Proposition 11 to the case that $f_j(\cdot; x)$ is independent of the parameter x.

Corollary 12 (regularity of resolvents of block separable functions)**.** *In the setting of Proposition 11 let $h_j(\cdot) : G_j \subset \mathcal{E}_j \to (-\infty, +\infty]$ satisfy*

$$\begin{aligned}
&\exists \tau_{h_j} \geq 0 :\ \forall x_j, y_j \in \mathcal{E}_j,\ \forall z_j \in t_j \partial h_j(x_j), w_j \in t_j \partial h_j(y_j),\\
&\qquad -\tfrac{\tau_{h_j}}{2} \|(x_j + z_j) - (y_j + w_j)\|^2 \leq \langle z_j - w_j,\ x_j - y_j \rangle .
\end{aligned} \tag{73}$$

Then for $h_t(x) := \sum_{j=1}^m t_j h_j(x_j)$, the resolvent $J_{\partial h_t, 1}$ is aα-fne with constant $\alpha_h = 1/2$ and violation $\tau_h = \max_j\{\tau_{h_j}\}$ on G. If h_j is convex on $\mathcal{E}_j$ for each $j = 1, 2, \ldots, m$, then $J_{\partial h_t, 1}$ is α-fne with constant $\alpha_h = 1/2$ and no violation on $\mathcal{E}$.

Proposition 13 (regularity of gradient descent)**.** *Let $f : \mathcal{E} \to \mathbb{R}$ be continuously differentiable with blockwise Lipschitz and hypomonotone gradient, that is f satisfies*

$$\forall j = 1, 2, \ldots, m,\ \exists L_j > 0 :\quad \sum_{j=1}^m \|\nabla_{x_j} f(x) - \nabla_{x_j} f(y)\|^2 \leq \sum_{j=1}^m L_j^2 \|x_j - y_j\|^2 \quad \forall x, y \in \mathcal{E}, \tag{74a}$$

and

$$\forall j = 1, 2, \ldots, m,\ \exists \tau_{f_j} \geq 0 : \sum_{j=1}^m -\tau_{f_j} \|x_j - y_j\|^2 \leq \sum_{j=1}^m \left\langle \nabla_{x_j} f(x) - \nabla_{x_j} f(y),\ x_j - y_j \right\rangle \quad \forall x, y \in \mathcal{E}. \tag{74b}$$

Then the gradient descent mapping with blockwise heterogeneous step lengths defined by $T_{GD} := \mathrm{Id} - \bigoplus_{j=1}^m t_j \nabla_{x_j} f$ is aα-fne on $\mathcal{E}$ with violation at most

$$\epsilon_{GD} = \max_j \left\{ 2 t_j \tau_j + \tfrac{t_j^2 L_j^2}{\overline{\alpha}} \right\} < 1, \qquad \textit{with constant} \quad \overline{\alpha} = \max_j\{\alpha_j\} \tag{75a}$$

whenever the blockwise steps t_j satisfy

$$t_j \in \left(0, \frac{\overline{\alpha}\sqrt{\tau_j^2 + L_j^2} - \overline{\alpha}\tau_j}{L_j^2} \right). \tag{75b}$$

If f is convex then, with global step size $t_j = t < \frac{2\overline{\alpha}}{\overline{L}}$ ($j = 1, 2, \ldots, m$) for $\overline{\alpha} \in (0, 1)$ with $\overline{L} = \max_j\{L_j\}$, the gradient descent mapping T_{GD} is α-fne with constant $\overline{\alpha}$ (no violation).

Proof. By [27, Proposition 2.1], the claim holds if and only if $\mathrm{Id} - \frac{1}{\overline{\alpha}} \bigoplus_{j=1}^m t_j \nabla_{x_j} f$ is almost nonexpansive on $\mathcal{E}$ with violation at most

$$\epsilon' = \epsilon_{GD}/\overline{\alpha} = \tfrac{1}{\overline{\alpha}} \max_j \left\{ 2t_j\tau_j + \tfrac{t_j^2 L_j^2}{\overline{\alpha}} \right\}. \tag{76}$$

To see this latter property, since f satisfies (74) we have

$$\begin{aligned}
&\left\| \left(x - \tfrac{1}{\overline{\alpha}} \bigoplus_{j=1}^m t_j \nabla_{x_j} f(x) \right) - \left(y - \tfrac{1}{\overline{\alpha}} \bigoplus_{j=1}^m t_j \nabla_{x_j} f(y) \right) \right\|^2 \\
&\quad = \|x-y\|^2 - \tfrac{2}{\overline{\alpha}} \sum_{j=1}^m t_j \left\langle x_j - y_j,\ \nabla_{x_j} f(x) - \nabla_{x_j} f(y) \right\rangle + \tfrac{1}{\overline{\alpha}^2} \sum_{j=1}^m t_j^2 \left\| \nabla_{x_j} f(x) - \nabla_{x_j} f(y) \right\|^2 \\
&\quad \le \|x-y\|^2 + \tfrac{2}{\overline{\alpha}} \sum_{j=1}^m t_j\tau_j \|x_j - y_j\|^2 + \tfrac{1}{\overline{\alpha}^2} \sum_{j=1}^m t_j^2 L_j^2 \|x_j - y_j\|^2 \\
&\quad \le \left(1 + \tfrac{1}{\overline{\alpha}} \max_j \left\{ 2t_j\tau_j + \tfrac{t_j^2 L_j^2}{\overline{\alpha}} \right\} \right) \|x-y\|^2
\end{aligned} \tag{77}$$

for all $x, y \in \mathcal{E}$. A simple calculation shows that the violation does not exceed 1 whenever the step t_j is bounded by (75b). This proves the result for the nonconvex setting.

If f is convex, then [21, Proposition 3.4] shows that a different bound on the steps is possible. Note that by [3, Corollaire 10]

$$\tfrac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2 \le \langle \nabla f(x) - \nabla f(y),\ x - y \rangle$$

Let $\overline{\alpha} = \max_j\{\alpha_j\}$ with $\alpha_j \in (0,1)$ and $\overline{L} = \max_j\{L_j\}$. For $t = \frac{2\overline{\alpha}}{\overline{L}}$ we have $2t = \frac{t^2\overline{L}}{\overline{\alpha}}$ and

$$\begin{gathered}
\tfrac{t^2\overline{L}}{\overline{\alpha}} \tfrac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2 \le 2t \langle \nabla f(x) - \nabla f(y),\ x - y \rangle \\
\Longleftrightarrow \\
\tfrac{1}{\overline{\alpha}} \|t\nabla f(x) - t\nabla f(y)\|^2 \le 2 \langle t\nabla f(x) - t\nabla f(y),\ x - y \rangle \\
\Longleftrightarrow \\
\|x-y\|^2 + \left(1 + \tfrac{1-\overline{\alpha}}{\overline{\alpha}}\right) \|t\nabla f(x) - t\nabla f(y)\|^2 \\
\le 2 \langle t\nabla f(x) - t\nabla f(y),\ x - y \rangle + \|x-y\|^2 \\
\Longleftrightarrow \\
\|(x - t\nabla f(x)) - (y - t\nabla f(y))\|^2 \\
\le \|x-y\|^2 - \tfrac{1-\overline{\alpha}}{\overline{\alpha}} \|t\nabla f(x) - t\nabla f(y)\|^2 \\
\Longleftrightarrow \\
\left\| \left(x - \bigoplus_{j=1}^m t\nabla_{x_j} f(x) \right) - \left(y - \bigoplus_{j=1}^m t\nabla_{x_j} f(y) \right) \right\|^2 \\
\le \|x-y\|^2 - \sum_{j=1}^m \tfrac{1-\overline{\alpha}}{\overline{\alpha}} \|t\nabla_j f(x) - t\nabla_j f(y)\|^2 \\
\Longleftrightarrow \\
\|T_{GD}x - T_{GD}y\|^2 \le \|x-y\|^2 - \tfrac{1-\overline{\alpha}}{\overline{\alpha}} \psi(x, y, T_{GD}x, T_{GD}y)
\end{gathered}$$

where the last implication follows from (26) with blockwise step $t_j = t$ for all j in T_{GD} .
□

Remark 1 The violation in the nonconvex case can be controlled by choosing a smaller blockwise step t_j. In the convex setting, larger step sizes are possible, but these are limited by the global Lipschitz constant $\overline{L}$ and the constant $\overline{\alpha}$. Note that the upper bound on the step length suggested by Proposition 13 is consistent with the upper bound on the steps in Example 2.

Proposition 14 (blockwise composite mappings). *Let $G \subset \mathcal{E}$ with $G_j \subset \mathcal{E}_j$ for $j = 1, 2 \ldots, m$.*

(i) ***Fully nonconvex.*** *For all $j \in \{1,2,\ldots,m\}$ let $f : G \to \mathbb{R}$ be subdifferentially regular with subdifferential satisfying* (72) *and let $h_j : G_j \to (-\infty,+\infty]$ be proper, l.s.c., and subdifferentially regular satisfying* (73).

(a) *The partial blockwise Douglas-Rachford mapping T_i^{DR} defined by* (19) *($j \in M_i$) is aα-fne on $G_{M_i} \bigoplus \{z\}_{M_i^\circ}$ for any fixed $z_{M_i^\circ} \in G_{M_i^\circ}$ with respective constant and violation*

$$\alpha_{DR} = \frac{2}{3}, \text{ and } \quad \epsilon_{DR} \le \tau_f + \tau_h + \tau_f\tau_h \tag{78}$$

where $\tau_h := \max_j\{\tau_{h_j}\}$ and $\tau_f := \max_j\{\tau_{f_j}\}$.

(b) *If f is continuously differentiable on $\mathcal{E}$ and satisfies* (74)*, the partial blockwise forward-backward mapping T_i^{FB} defined by* (17) *with step lengths t_j satisfying* (75b) *($j \in M_i$) is aα-fne on affine subspaces $G_{M_i} \bigoplus \{z\}_{M_i^\circ}$ for any fixed $z_{M_i^\circ} \in G_{M_i^\circ}$ with respective constant and violation*

$$\alpha_{FB} := \frac{2}{1 + \frac{1}{\max\{\frac{1}{2},\ \overline{\alpha}\}}}, \text{ and } \quad \epsilon_{FB} \le \epsilon_{GD} + \tau_h + \epsilon_{GD}\tau_h \tag{79}$$

where $\overline{\alpha} := \max_j\{\alpha_j\}$, $\tau_h := \max_j\{\tau_{h_j}\}$ and ϵ_{GD} is no larger than (75a).

(ii) ***Partially nonconvex.*** *For all $j \in \{1,2,\ldots,m\}$ let $f : \mathcal{E} \to \mathbb{R}$ be continuously differentiable with gradient satisfying* (74) *and let the functions h_j be convex on G_j ($j = 1,2,\ldots,m$). Then for all $i \in \mathbb{I}$, T_i^{FB} is aα-fne on $G_{M_i} \bigoplus \{z\}_{M_i^\circ}$ for any $z_{M_i^\circ} \in G_{M_i^\circ}$ with constant α_{FB} given by* (79)*, violation ϵ_{FB} at most ϵ_{GD}, and this can be made arbitrarily small by choosing the step lengths t_i small enough.*

(iii) ***Convex.*** *If f and h_j are convex on $\mathcal{E}$ ($j = 1,2,\ldots,m$), then*

(a) *T_i^{DR} is α-fne on $\mathcal{E}_{M_i} \bigoplus \{z\}$ with constant $\alpha_{DR} = 2/3$ and no violation;*

(b) *if f is continuously differentiable and ∇f satisfies* (74a)*, T_i^{FB} with global step size $t < \frac{2\overline{\alpha}}{\overline{L}}$ for $\overline{L} = \max_j\{L_j\}$ is α-fne on $\mathcal{E}_{M_i} \bigoplus \{z\}$ with constant α_{FB} given by* (79) *and no violation.*

Proof. Part (i). By Theorem 1, the respective regularity constants ϵ_i and α_i are bounded above by the respective constants of T_1^{FB} and T_1^{DR}, which are the mappings including all of the blocks. It suffices, then, to determine the regularity of T_1^{FB} and T_1^{DR}. Part (ia). By Proposition 11 and Corollary 12 $J_{\partial f_t}$ and $J_{\partial h_t}$ are aα-fne with constant $\alpha_{h_t} = 1/2$ and violation $\tau_f = \max_j\{\tau_{f_j}\}$ (respectively $\tau_h = \max_j\{\tau_{h_j}\}$) on G. Then by [27, Proposition 2.4] T_1^{DR} is aα-fne with constant $\alpha_{DR} = 2/3$ and (maximal) violation given by (78) on G.

Part (ib). By Proposition 13, T_{GD} is aα-fne on G with violation ϵ_{GD} no larger than (75a) and constant $\overline{\alpha} = \max_j\{\alpha_j\}$. By Corollary 12 $J_{\partial h_t}$ is aα-fne with constant $\alpha_{h_t} = 1/2$ and violation $\tau_h = \max_j\{\tau_{h_j}\}$ on $\mathcal{E}$. Then by [27, Proposition 2.4/Proposition 3.7] T_1^{FB} is aα-fne with constant α_{FB} and (maximal) violation given by (79) on G.

Parts (ii)-(iii) follow immediately from part (i) and Propositions 11-13. □

Corollary 15. *For $G \subset \mathcal{E}$, let $\Phi : G \times \mathbb{I} \to G$ be the update function given by $\Phi(x,i) = T_i x$ where T_i is either T_i^{FB} or T_i^{DR} defined respectively by* (17) *and* (19).

(i) ***Fully nonconvex.*** *Under the assumptions of Proposition 14*(i)*, that is both f and h in* (17) *are nonconvex, the corresponding update function $\Phi(x,i)$ is aα-fne in expectation with respect to the weighted norm $\|\cdot\|_{\mathbf{p}}$ with regularity constants $\overline{p}\epsilon_{DR}$ and α_{DR} (respectively $\overline{p}\epsilon_{FB}$ and $\alpha_{FB} = 2/3$) corresponding to* (78) *(respectively* (79)*).*

(ii) ***Partially nonconvex.*** *Under the assumptions of Proposition 14*(ii)*, that is f smooth nonconvex with Lipschitz and hypomonotone gradient and h_j convex in* (17)*, $\Phi(x,i) = T_i^{FB}(x)$ is aα-fne in expectation with respect to the weighted norm $\|\cdot\|_{\mathbf{p}}$ with constant α_{FB} as above and violation at most $\overline{p}\epsilon_{GD}$ with ϵ_{GD} given by* (75a)*; this violation can be made arbitrarily small by choosing the step lengths t_i small enough.*

(iii) ***Convex.*** *If both f and h_j are convex on $\mathcal{E}$ ($j = 1,2,\ldots,m$), then $T_i^{DR}(x)$ is α-fne in expectation with respect to the weighted norm $\|\cdot\|_{\mathbf{p}}$ (no violation) and constant $\alpha_{DR} = 2/3$ on $\mathcal{E}$. In the case of T^{FB}, if ∇f satisfies* (74a) *and the global step size is bounded by $t < \frac{2\overline{\alpha}}{\overline{L}}$ for $\overline{L} = \max_j\{L_j\}$, $T_i^{FB}(x)$ is α-fne in expectation with respect to the weighted norm $\|\cdot\|_{\mathbf{p}}$ (no violation) and constant α_{GD} on $\mathcal{E}$.*

Proof. This is an immediate consequence of Proposition 14 and Theorem 1 □

Before presenting the convergence results it is worthwhile pointing out that the partial blockwise forward-backward mappings T_i^{FB} and T_i^{DR} have common fixed points, and these are critical points of (1). In other words, the stochastic fixed point problem is consistent. As shown in Section 4.2.1, in this case the metric subregularity condition (55) simplifies to (65) when $\Psi(\mu) = 0$ if and only if $\mu \in \operatorname{inv} \mathcal{P}$ and $\operatorname{supp} \mu \subset C$. In the convex setting we have the following correspondence between invariant measures of the stochastic block iterations and minima of (1).

Proposition 16. *Let $\mathcal{P}$ be the Markov operator associated with either Algorithm 2 or 3. In the setting of Lemma 10, if f and g_j (for all $j = 1, \dots, m$) are convex, then $\operatorname{inv} \mathcal{P} = \{\pi \mid \operatorname{supp} \pi \subset C\}$ and whenever $x \in C$ then almost surely $x \in \operatorname{argmin} \{f + \sum_{j=1}^m g_j\}$.*

Proof. When f and g_j (for all $j = 1, \dots, m$) are convex, the corresponding mappings T_i defined by either (17) or (19) are single-valued self-mappings on $\mathcal{E}$ and α-fne on $\mathcal{E}_{M_i} \oplus \{z\}_{M_i^\circ}$ for every $z \in M_i^\circ$ as long as the step size t_i is small enough ([27, Propositions 3.7 and 3.10] specialized to the convex case). Then by Corollary 4 the mappings T_i are paracontractions in expectation on $\mathcal{E}$. The claim then follows from Theorem 5(ii) and Lemma 10 since in this case $\operatorname{crit} (f + \sum_{j=1}^m g_j) = \operatorname{argmin} \{f + \sum_{j=1}^m g_j\}$. □

The final result of this study collects all of these facts in the context of the Markov chain underlying Algorithm 2 and 3.

Proposition 17. *Let $\mathcal{P}$ be the Markov operator associated with the S-BFBS Algorithm 2 or the S-BDRS Algorithm 3 and let $(\mu_k)_{k\in\mathbb{N}}$ be the corresponding sequence of measures initialized by any $\mu_0 \in \mathscr{P}_2(G)$, where G is a closed subset of $\mathcal{E}$. Assume that $G \supset \operatorname{crit} \left(f + \sum_{j=1}^m g_j\right) \neq \emptyset$ and $\operatorname{inv} \mathcal{P} = \mathscr{C}$ defined by (47). Let Ψ given by (41) be such that $\Psi(\mu) = 0$ if and only if $\mu \in \operatorname{inv} \mathcal{P}$. Additionally, let the mappings T_i be self-mappings on G where Ψ satisfies (65) with gauge ρ given by (57) with $\tau = (1-\alpha_*)/\alpha_*$, $\epsilon = \overline{p}\epsilon_*$ for constants α_* and violation ϵ_* given by either (79) or (78) (depending on the algorithm), and $\theta_{\tau,\epsilon}$ satisfying (56).*

(1) ***Fully nonconvex.*** *Under the assumptions of Proposition 14(i), the sequence (μ_k) satisfies*

$$d_{W_{2,\mathbf{p}}} (\mu_{k+1}, \operatorname{inv} \mathcal{P}) \le \theta_{\tau,\epsilon} \left(d_{W_{2,\mathbf{p}}} (\mu_k, \operatorname{inv} \mathcal{P})\right) \quad \forall k \in \mathbb{N}.$$

If τ and ϵ are such that at least one of the conditions in Assumption 2 holds, then $\mu_k \to \pi^{\mu_0} \in \operatorname{inv} \mathcal{P} \cap \mathscr{P}_2(G)$ in the $d_{W_{2,\mathbf{p}}}$ metric with rate $O\left(\theta_{\tau,\epsilon}^{(k)}(t_0)\right)$ in the case of Assumption 2(a) where $t_0 = d_{W_{2,\mathbf{p}}} \left(\mu^0, \operatorname{inv} \mathcal{P}\right)$, and with rate $O(s_k(t_0))$ for $s_k(t) := \sum_{j=k}^\infty \theta_{\tau,\epsilon}^{(j)}(t)$ in the case of Assumption 2(b). Moreover, $\operatorname{supp} \pi^{\mu_0} \subset C := \{x \in G \mid \mathbb{P}\left(T_\xi x = x\right) = 1\}$.

(2) ***Partially nonconvex.*** *Under the assumptions of Proposition 14(ii), that is f smooth nonconvex with Lipschitz and hypomonotone gradient and h_j convex in (17), if there exist τ and ϵ such that at least one of the conditions in Assumption 2 holds, then for all step lengths t_i small enough in (17) and any initial distribution μ_0 close enough to $\operatorname{inv} \mathcal{P}$, the sequence $\mu_k \to \pi^{\mu_0} \in \operatorname{inv} \mathcal{P} \cap \mathscr{P}_2(G)$ in the $d_{W_{2,\mathbf{p}}}$ metric with rate at least $O\left(\theta_{\tau,\epsilon}^{(k)}(t_0)\right)$ in the case of Assumption 2(a), and with rate at least $O(s_k(t_0))$ in the case of Assumption 2(b); moreover, $\operatorname{supp} \pi^{\mu_0} \subset C$.*

(3) ***Convex.*** *If f and h_j are convex on $\mathcal{E}$, and there exists τ such that at least one of the conditions in Assumption 2 holds when $\epsilon = 0$, then the sequence (μ_k) corresponding to Algorithm 3 initialized from any $\mu_0 \in \mathscr{P}_2(\mathcal{E})$, converges in the metric $d_{W_{2,\mathbf{p}}}$ to an invariant distribution with rate at least $O\left(\theta_{\tau,\epsilon}^{(k)}(t_0)\right)$ in the case of Assumption 2(a), and with rate at least $O(s_k(t_0))$ in the case of Assumption 2(b). Moreover $\operatorname{supp} \pi^{\mu_0} \subset C := \operatorname{argmin} \left(f + \sum_{j=1}^m g_j\right)$. If f is continuously differentiable and satisfies (74a), then the stated convergence in the case of Algorithm 2 holds for the global step length $t < \frac{2\overline{\alpha}}{\overline{L}}$.*

6 Final Remarks

There a several open technicalities lurking between the lines above, and one rather obvious challenge hiding in plain sight. To the hidden technicalities belong the question of whether metric subregularity is necessary for quantitative convergence in some appropriate metric of Markov operators that are not paracontractions in measure. We conjecture that this is true. Another open technical issue concerns the statement of asymptotic regularity in Proposition 6. This result is incomplete without some extension to a weak type of convergence in distribution. For consistent stochastic fixed point problems, if each of the update functions T_i were α-fne, then almost sure weak convergence of the iterates is guaranteed [18, Theorem 3.9]; at issue here is whether this holds when T_i is pointwise α-fne *in expectation* at invariant measures of the corresponding Markov operator. We expect that there should be a counterexample to this claim. Characterization of the supports of invariant measures in the inconsistent case is quite challenging and essential for meaningfully connecting the limiting distributions of the algorithms to solutions to the underlying optimization problem. Finally, the restriction of the study to single-valued mappings does not allow one to capture the full extent of behavior one sees with nonconvex problems. Projection methods for sparse affine feasibility, for instance, have the property that the projection onto a sparsity constraint can be multi-valued on all neighborhoods of a solution (see [17, Lemma III.2]). An extension of the analysis presented here to multi-valued mappings, is required.

The most difficult challenge to all of this is the task of numerically monitoring convergence in distribution of random variables. To do this completely one needs first of all efficient means for computing the Wasserstein distance between measures; in other words, one needs to solve optimal transport problems efficiently. Again, for consistent stochastic feasibility when convergence of the iterates can be guaranteed almost surely, optimal transport is not needed; more generally, however, this machinery is essential. Secondly, one needs to numerically estimate the distributions whose distances are to be computed. These are significant challenges worthy of attention.

7 Funding and/or Conflicts of interests/Competing interests

This work was supported in part by a grant from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 432680300 – SFB 1456. The manuscript has not been submitted to any other journal for simultaneous consideration. The author has no financial or non-financial interests that are directly or indirectly related to the work submitted for publication.

References

1. L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures.* Basel: Birkhäuser, 1st edition, 2005.
2. J. B. Baillon, R. E. Bruck, and S. Reich. On the asymptotic behavior of nonexpansive mappings and semigroups in Banach spaces. *Houston J. Math.*, 4(1):1–9, 1978.
3. J.-B. Baillon and G. Haddad. Quelques propriétés des opérateurs angle-bornes et n-cycliquement monotones. *Isr. J. Math.*, 26:137–150, 1977.
4. P. Billingsley. *Convergence of probability measures. 2nd ed.* Chichester: Wiley, 2nd ed. edition, 1999.
5. D. Butnariu. The expected-projection method: Its behavior and applications to linear operator equations and convex optimization. *J. Appl. Anal.*, 1(1):93–108, 1995.
6. D. Butnariu, Y. Censor, and S. Reich. Iterative averaging of entropic projections for solving stochastic convex feasibility problems. *Computational Optimization and Applications*, 8:21–39, 1997.
7. D. Butnariu and S. D. Flåm. Strong convergence of expected-projection methods in Hilbert spaces. *Numer. Funct. Anal. and Optim.*, 16(5&6):601–636, 1995.
8. L. M. Briceño-Arias, G. Chierchia, E. Chouzenoux, and J.-C. Pesquet. A random block-coordinate Douglas–Rachford splitting method with low computational complexity for binary logistic regression. *Computational Optimization and Applications*, (72):707–726, 2019.
9. R. E. Bruck and S. Reich. Nonexpansive projections and resolvents of accretive operators in Banach spaces. *Houston J. Math.*, 3(4):459–470, 1977.
10. P. L. Combettes and J.-C. Pesquet. Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. *SIAM J. Optim.*, 25(2):1221–1248, 2015.
11. J. Eckstein. *Splitting Methods for Monotone Operators with Applications to Parallel Optimization.* PhD thesis, MIT, Cambridge, MA, 1989.
12. M. Edelstein. A remark on a theorem of M. A. Krasnoselski. *Amer. Math. Monthly*, 73(5):509–510, May 1966.
13. O. Fercoq and P. Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM J. Optim.*, 25(4):1997–2023, 2015.
14. D. Gabay. *Augmented Lagrangian Methods: Applications to the Solution of Boundary- Value Problems*, chapter Applications of the method of multipliers to variational inequalities, pages 299–331. North-Holland, Amsterdam, 1983.
15. R. Glowinski and A. Marroco. Sur l'approximation, par elements finis d'ordre un, et las resolution, par penalisation-dualitè, d'une classe de problemes de dirichlet non lineares. *Revue Francais d'Automatique, Informatique et Recherche Opérationelle*, 9(R-2):41–76, 1975.

16. M. Hairer. Convergence of Markov processes. *Lecture notes, University of Warwick*, 2021.
17. R. Hesse, D. R. Luke and P. Neumann Alternating projections and Douglas–Rachford for sparse affine feasibility. *IEEE Trans. Signal. Process.* 62(18):4868–4881, 2014.
18. N. Hermer, D. R. Luke, and A. Sturm. Random function iterations for consistent stochastic feasibility. *Numer. Funct. Anal. Opt.*, 40(4):386–420, 2019.
19. N. Hermer, D. R. Luke, and A. Sturm. Nonexpansive Markov operators and random function iterations for stochastic fixed point problems. *J. Conv. Anal.*, 30(4):1073–1114, 2023. https://arxiv.org/abs/2205.15897.
20. N. Hermer, D. R. Luke, and A. Sturm. Rates of convergence for chains of expansive Markov operators. https://arxiv.org/abs/2206.05213, 2022.
21. I. Kartamyschew. Random forward-backward algorithm in the context of random function iteration. Master's thesis, Universität Göttingen, August 2020.
22. M. A. Krasnoselski. Two remarks on the method of successive approximations. *Math. Nauk. (N.S.)*, 63(1):123–127, 1955. (Russian).
23. Z. Lu and L. Xiao. On the complexity analysis of randomized block-coordinate descent methods. *Math. Program.*, 152(1-2 (A)):615–642, 2015.
24. D. R. Luke and Y. Malitsky. Block-coordinate primal-dual method for the nonsmooth minimization over linear constraints. In P. Giselsson and A. Rantzer, editors, *Distributed and Large-Scale Optimization.* Springer Verlag, 2018.
25. D. R. Luke, S. Sabach, and M. Teboulle. Optimization on spheres: Models and proximal algorithms with computational performance comparisons, *SIAM J. on Math. of Data Sci.*, 1(3), 408-445, 2019.
26. D. R. Luke, M. Teboulle, and N. H. Thao. Necessary conditions for linear convergence of iterated expansive, set-valued mappings. *Math. Program.*, 180:1–31, 2018.
27. D. R. Luke, N. H. Thao, and M. K. Tam. Quantitative convergence analysis of iterated expansive, set-valued mappings. *Math. Oper. Res.*, 43(4):1143–1176, 2018.
28. Z.-Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Ann. of Oper. Res.*, 46(1):157–178, 1993.
29. W. R. Mann. Mean value methods in iterations. *Proc. Amer. Math. Soc.*, 4:506–510, 1953.
30. J. J. Moreau. Proximité et dualité dans un espace Hilbertian. *Bull. de la Soc. Math. de France*, 93(3):273–299, 1965.
31. I. Necoara and D. Clipici. Parallel random coordinate descent method for composite minimization: convergence analysis and error bounds. *SIAM J. Optim.*, 26(1):197–226, 2016.
32. A. Nedić. Random algorithms for convex minimization problems. *Math. Program.*, 129(2):225–253, Oct 2011.
33. Yu. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM J. Optim.*, 22(2):341–362, 2012.
34. J.-C. Pesquet and A. Repetti. A class of randomized primal-dual algorithms for distributed optimization. *J. Nonlinear Convex Anal.*, 16(12):2453–2490, 2015.
35. Z. Qu and P. Richtárik. Coordinate descent with arbitrary sampling. I: Algorithms and complexity. *Optim. Methods Softw.*, 31(5):829–857, 2016.
36. P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Math. Program.*, 144(1-2 (A)):1–38, 2014.
37. P. Richtárik and M. Takáč. Distributed coordinate descent method for learning with big data. *J. Mach. Learn. Res.*, 17(75):1–25, 2016.
38. R. T. Rockafellar and R. J. Wets. *Variational Analysis.* Grundlehren Math. Wiss. Springer-Verlag, Berlin, 3 edition, 2009.
39. S. Salzo and S. Villa. Parallel random block-coordinate forward-backward algorithm: a unified convergence analysis. *Math. Program.*, pages 1436–4646, 2021.
40. D.W. Stroock. *Probability Theory: An Analytic View.* Cambridge University Press, 2010.
41. T. Szarek. Feller processes on nonlocally compact spaces. *Ann. Probab.*, 34(5):1849–1863, 2006.
42. C. Villani. *Optimal transport: Old and New.*, volume 338. Berlin: Springer, 2009.
43. S. J. Wright. Coordinate descent algorithms. *Math. Program.*, 151(1 (B)):3–34, 2015.