Nonexpansive Markov Operators and Random Function Iterations for Stochastic Fixed Point Problems

Neal Hermer*, D. Russell Luke†and Anja Sturm‡

March 1, 2023

Abstract

We study the convergence of random function iterations for finding an invariant measure of the corresponding Markov operator. We call the problem of finding such an invariant measure the *stochastic fixed point problem*. This generalizes earlier work studying the *stochastic feasibility problem*, namely, to find points that are, with probability 1, fixed points of the random functions [31]. When no such points exist, the stochastic feasibility problem is called *inconsistent*, but still under certain assumptions, the more general stochastic fixed point problem has a solution and the random function iteration converges to an invariant measure for the corresponding Markov operator. We show how common structures in deterministic fixed point theory can be exploited to establish existence of invariant measures and convergence in distribution of the Markov chain. This framework specializes to many applications of current interest including, for instance, stochastic algorithms for large-scale distributed computation, and deterministic iterative procedures with computational error. The theory developed in this study provides a solid basis for describing the convergence of simple computational methods without the assumption of infinite precision arithmetic or vanishing computational errors.

2010 Mathematics Subject Classification: Primary 60J05, 46N10, 46N30, 65C40, 49J55 Secondary 49J53, 65K05.

Keywords: Averaged mappings, nonexpansive mappings, stochastic feasibility, inconsistent stochastic fixed point problem, iterated random functions, convergence of Markov chain

1. Introduction

Stochastic algorithms have emerged as a major approach to solving large-scale optimization problems. The analysis of these algorithms is for the most part restricted to ergodic results,

*Institute for Numerical and Applied Mathematics, University of Goettingen, 37083 Goettingen, Germany. NH was supported by Deutsche Forschungsgemeinschaft Research Training Grant 2088 TP-B5. E-mail: n.hermer@math.uni-goettingen.de

†Institute for Numerical and Applied Mathematics, University of Goettingen, 37083 Goettingen, Germany. DRL was supported in part by Deutsche Forschungsgemeinschaft Research Training Grant 2088 TP-B5. E-mail: r.luke@math.uni-goettingen.de

‡Institute for Mathematical Stochastic, University of Goettingen, 37077 Goettingen, Germany. AS was supported in part by Deutsche Forschungsgemeinschaft Research Training Grant 2088 TP-B5. E-mail: asturm@math.uni-goettingen.de

that is, convergence of the average of the iterates to a single fixed point. The approach we present in this paper aims to provide a convergence theory for the entire distribution behind the iterates, not just their mean. In order to keep the already technical proofs as simple as possible, the underlying setting is relatively benign: we consider random selections of nonexpansive self-mappings on a Euclidean space, denoted $\mathcal{E}$; the self-mappings, $T_i : \mathcal{E} \to \mathcal{E}$ where i indexes a possibly uncountable collection $\{T_i\}_{i\in I}$, are understood as actions taken by an algorithm on an iterate X_k to produce the next iterate X_{k+1}. The procedure is formally a *random function iteration* (RFI)[24].

The present study is a continuation of a development begun in [31] which was confined to the assumption that the self-mappings have common fixed points. The deterministic analog to this situation is classical. The deterministic result that tracks most closely to our assumptions is [2, Theorem 4.1] where the authors study finite compositions of *firmly nonexpansive* self-mappings that have common fixed points, and which have *boundedly compact images*. We are not aware of any convergence result for a deterministic fixed point iteration that does not require compactness of some sort, though the requirement of common fixed points can be dropped in Hilbert space settings. Our stochastic results are most closely anticipated by Butnariu [14] who studied the variant where the mappings T_i are projectors onto convex sets C_i; the collection of sets in that study did not need to be countable nor have common points.

We are concerned in this paper with, to a lesser extent, (i) *existence of invariant distributions* of the Markov operators associated with the random function iterations, and, our principal focus, (ii) *convergence* of the Markov chain to an invariant distribution. Rates of convergence and attendant stopping rules are postponed for a follow-up work. The existence theory is already well developed and is surveyed in Section 3.1 below. We show how existence is guaranteed when, for instance, the image is compact for some non-negligible collection of operators T_i (Proposition 3.2) or when the expectation of the random variables X_k is finite (Proposition 3.3). Beyond these preparatory results, our main focus is convergence, and here there are two principal contributions: first, we provide convergence results without the assumption that the (possibly uncountably infinite collection of) self mappings possess common fixed points; secondly, we prove convergence without any compactness assumption. While the case of consistent stochastic feasibility can be analyzed with the standard tools from deterministic fixed point theory [31], dropping the assumption of common fixed points requires the full extent of elementary probability theory; the dividend for this is a vastly expanded range of applications. That we do not require compactness for convergence shows that randomization of the usual deterministic strategies compensates for the absence of compactness, and indicates a way forward for other results that rely on this assumption. The contribution to the theory of Markov chains is the extension to applications where the operators are not contractive, quasi- or otherwise.

There are many more applications than one could reasonably list, but to reach the broadest possible audience and inspired by [14], a simple example from first semester numerical analysis is illustrative. Consider the underdetermined linear system of equations

$$Ax = b, \quad A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m, \ m < n.$$

Equivalent to this problem is the problem of finding the intersection of the hyperplanes defined by the single equations $\langle a_j, x\rangle = b_j$ $(j = 1, 2, \ldots m)$ where a_j is the jth row of the matrix A:

$$\text{Find} \quad \bar{x} \in \cap_{j=1}^m \{x \mid \langle a_j, x\rangle = b_j\}. \tag{1}$$

An intuitive technique to solve this problem is the method of cyclic projections: Given an initial guess x_0, construct the sequence (x_k) via

$$x_{k+1} = P_m P_{m-1} \cdots P_1 x_k, \tag{2}$$

where P_j is the orthogonal projection onto the jth hyperplane above. This method was proposed by von Neumann in [59] where he also proved that, without numerical error, the iterates converge to the projection of the initial point x_0 onto the itersection.

The projectors have a closed form representation, and the algorithm is easily implemented. The results of one implementation for a randomly generated matrix A and vector b with $m = 50$ and $n = 60$ yields the following graph shown in Figure 1(a). As the figure shows, the method

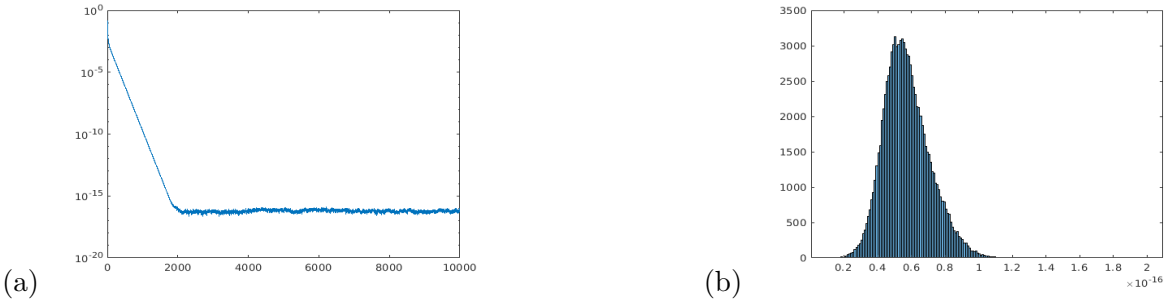

Figure 1: (a) The residual $x_{k+1} - x_k$ of iterates of the cyclic projections algorithm for solving the linear system $Ax = b$ for $A \in \mathbb{R}^{50\times 60}$, and $b \in \mathbb{R}^{50}$ randomly generated. (b) A histogram of the residual sizes over the last 8000 iterations.

performs as predicted by the theory, up to the numerical precision of the implementation. After that point, the iterates behave more like random variables with distribution indicated by the histogram shown in Figure 1(b). The theory developed in this study provides a solid basis for describing the convergence of simple computational methods without the assumption of infinite precision arithmetic or vanishing computational errors [52, 54]. This particular situation could be analyzed in the stability framework of perturbed convergent fixed point iterations with unique fixed points developed in [18]; our approach captures their results and opens the way to a much broader range of applications. An analysis of nonmonotone fixed point iterations with error can be found already in [33], though the precision is assumed to increase quickly to exact evaluation.

The main object of interest in our approach is the Markov operator on a space of probability measures with the appropriate metric. We take for granted much of the basic theory of Markov chains, which interested readers can find, for instance, in [32] or [43]. We are indebted to the work of Butnariu and collaborators who studied stochastic iterative procedures for solving infinite dimensional linear operator equations in [14–17]. Another important application motivating our analytical strategy involves stochastic implementations of deterministic algorithms for large-scale optimization problems [11, 21, 27, 45, 53]. Such stochastic algorithms are central to distributed computation with applications in machine learning [6, 22, 25, 29, 36, 50]. Here each T_{ξ_k} represents a randomly selected, low-dimensional update mechanism in an iterative procedure.

As with classical fixed point iterations, the limit – or more accurately, *limiting distribution* – of the Markov chain, if it exists, will in general depend on the initialization. Uniqueness of invariant measures of the Markov operator is not a particular concern for feasibility problems where *any* feasible point will do. The notation and necessary background is developed in Section 2, which we conclude with the main statements of this study (Section 2.5). Section 3 contains the technical details, starting with existence theory in Section 3.1, general ergodic theory in Section 3.2 with gradually increasing regularity assumptions on the Markov operators, equicontinuity in Section 3.3 and finally Markov operators generated by nonexpansive mappings in Section 3.4. The assumptions on the mappings generating the Markov operators are commonly employed in the analysis of deterministic algorithms in continuous optimization. Our first main result,

Theorem 2.8, establishes convergence for Markov chains that are generated from nonexpansive mappings in $\mathcal{E}$ and follows easily in Section 3.5 upon establishing tightness of the sequence of measures. Section 3.6 collects further facts needed for the second main result of this study, Theorem 2.9, which establishes convergence in the Prokhorov-Lèvy metric of Markov chains to an invariant measure (assuming this exists) when the Markov operators are constructed from *averaged mappings* in $\mathcal{E}$ (Definition 2.7). We conclude this study with Section 4 where we focus on applications to optimization on measure spaces and (inconsistent) feasibility.

2. Random Function Iterations and the Stochastic Fixed Point Problem

In this section we give a rigorous formulation of the RFI, then interpret this as a Markov chain and define the corresponding Markov operator. We then formulate modes of convergence of these Markov chains to invariant measures for the Markov operators and formulate the stochastic feasibility and stochastic fixed point problems. At the end of this section we present the main results of this article. The proofs of these results are developed in Section 3.

Our notation is standard. As usual, $\mathbb{N}$ denotes the natural numbers *including* 0. For the Euclidean space $\mathcal{E}$ the Borel σ-algebra is denoted by $\mathcal{B}(\mathcal{E})$ and $(\mathcal{E}, \mathcal{B}(\mathcal{E}))$ is the corresponding measure space. We denote by $\mathscr{P}(\mathcal{E})$ the set of all probability measures on $\mathcal{E}$. The *support of the probability measure* μ is the smallest closed set A, for which $\mu(A) = 1$ and is denoted by $\operatorname{supp} \mu$.

There is a lot of overlapping notation in probability theory. Where possible we will try to stick to the simplest conventions, but the context will make certain notation preferable. The notation $X \sim \mu \in \mathscr{P}(\mathcal{E})$ 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. All of these different ways of indicating a measure μ will be used.

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 to a set $A \subset \mathcal{E}$ is denoted by $d(x, A) := \inf_{w \in A} \|x - w\|$. For the ball of radius r around a subset of points $A \subset \mathcal{E}$, we write $\mathbb{B}(A, r) := \bigcup_{x \in A} \mathbb{B}(x, r)$. The 0-1-indicator function of a set A is given by

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{else.} \end{cases}$$

Continuing with the development initiated in the introduction, we will consider a collection of mappings $T_i : \mathcal{E} \to \mathcal{E}$, $i \in I$ where I is an arbitrary index set. The measure space of indexes is denoted by $(I, \mathcal{I})$, and ξ is an I-valued random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The pairwise independence of two random variables ξ and η is denoted $\xi \perp\!\!\!\perp \eta$. 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 having the same distribution as ξ ($\xi_k \overset{\mathrm{d}}{=} \xi$). The method of random function iteration is formally presented in Algorithm 1.

Algorithm 1: Random Function Iteration (RFI)

Initialization: Set $X_0 \sim \mu_0 \in \mathscr{P}(\mathcal{E})$, $\xi_k \sim \xi \quad \forall k \in \mathbb{N}$.

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

  $X_{k+1} = T_{\xi_k} X_k$

We will use the notation

$$X_k^{X_0} := T_{\xi_{k-1}} \dots T_{\xi_0} X_0 \tag{3}$$

to denote the sequence of the RFI initialized with $X_0 \sim \mu_0$. This is particularly helpful when characterizing sequences initialized with the delta distribution of a point, where X_k^x denotes the RFI sequence initialized with $X_0 \sim \delta_x$. The following assumptions will be employed throughout.

Assumption 2.1. (a) $X_0, \xi_0, \xi_1, \dots, \xi_k$ are independent for every $k \in \mathbb{N}$, where ξ_k are i.i.d. for all k with the same distribution as ξ.

(b) The function $\Phi : \mathcal{E} \times I \to \mathcal{E}$, $(x, i) \mapsto T_i x$ is measurable.

2.1. RFI as a Markov chain

Markov chains are conveniently defined in terms of *transition kernels*. A transition kernel is a mapping $p : \mathcal{E} \times \mathcal{B}(\mathcal{E}) \to [0, 1]$ that is measurable in the first argument and is a probability measure in the second argument; that is, $p(\cdot, A)$ is measurable for all $A \in \mathcal{B}(\mathcal{E})$ and $p(x, \cdot)$ is a probability measure for all $x \in \mathcal{E}$.

Definition 2.2 (Markov chain)**.** A sequence of random variables (X_k), $X_k : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{E}, \mathcal{B}(\mathcal{E}))$ is called Markov chain with transition kernel p if for all $k \in \mathbb{N}$ and $A \in \mathcal{B}(\mathcal{E})$ $\mathbb{P}$-a.s. the following hold:

(i) $\mathbb{P}(X_{k+1} \in A \,|\, X_0, X_1, \dots, X_k) = \mathbb{P}(X_{k+1} \in A \,|\, X_k)$;

(ii) $\mathbb{P}(X_{k+1} \in A \,|\, X_k) = p(X_k, A)$.

Proposition 2.3. *Under Assumption 2.1, the sequence of random variables (X_k) generated by Algorithm 1 is a Markov chain with transition kernel p given by*

$$(x \in \mathcal{E})(A \in \mathcal{B}(\mathcal{E})) \qquad p(x, A) := \mathbb{P}(\Phi(x, \xi) \in A) = \mathbb{P}(T_\xi x \in A) \tag{4}$$

for the measurable update function $\Phi : \mathcal{E} \times I \to \mathcal{E}$, $(x, i) \mapsto T_i x$.

Proof. It follows from [34, Lemma 1.26] that the mapping $p(\cdot, A)$ defined by (4) is measurable for all $A \in \mathcal{B}(\mathcal{E})$, and it is immediate from the definition that $p(x, \cdot)$ is a probability measure for all $x \in \mathcal{E}$. So p defined by (4) is a transition kernel. The remainder of the statement is an immediate consequence of the disintegration theorem (see, for example, [56]). □

The Markov operator $\mathcal{P}$ is defined pointwise for a measurable function $f : \mathcal{E} \to \mathbb{R}$ via

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

when the integral exists. Note that

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

The Markov operator $\mathcal{P}$ is *Feller* if $\mathcal{P}f \in C_b(\mathcal{E})$ whenever $f \in C_b(\mathcal{E})$, where $C_b(\mathcal{E})$ is the set of bounded and continuous functions from $\mathcal{E}$ to $\mathbb{R}$. This property is central to the theory of existence of invariant measures introduced below. The next fundamental result establishes the relation of the Feller property of the Markov operator to the generating mappings T_i.

Proposition 2.4 (Theorem 4.22 in [8])**.** *Under Assumption 2.1, if T_i is continuous for all $i \in I$, then the Markov operator $\mathcal{P}$ is Feller.*

Let $\mu \in \mathscr{P}(\mathcal{E})$. In a slight abuse of notation we denote the dual Markov operator $\mathcal{P}^* : \mathscr{P}(\mathcal{E}) \to \mathscr{P}(\mathcal{E})$ acting on a measure μ by action on the right by $\mathcal{P}$ via

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

This notation allows easy identification of the distribution of the k-th iterate of the Markov chain generated by Algorithm 1: $\mathcal{L}(X_k) = \mu_0\mathcal{P}^k$.

2.2. The Stochastic Fixed Point Problem

As studied in [31], the *stochastic feasibility* problem is stated as follows:

$$\text{Find } x^* \in C := \{x \in \mathcal{E} \,|\, \mathbb{P}(x = T_\xi x) = 1\}\,. \tag{5}$$

A point x such that $x = T_i x$ is a *fixed point* of the operator T_i; the set of all such points is denoted by

$$\operatorname{Fix} T_i := \{x \in \mathcal{E} \,|\, x = T_i x\}\,.$$

In [31] it was assumed that $C \neq \emptyset$. If $C = \emptyset$ this is called the *inconsistent stochastic feasibility* problem.

Inconsistent stochastic feasibility is far from exotic. Take, for example, the not unusual assumption of additive noise: define $T_\eta(x) := f(x) + \eta$ where $f : \mathbb{R}^n \to \mathbb{R}^n$ and η is noise. Then $\mathbb{P}(T_\eta(x) = x) = \mathbb{P}(\eta = x - f(x)) = 0$. More concretely, let $f = \mathrm{Id} - t\nabla F$ where $F : \mathbb{R}^n \to \mathbb{R}$ is a differentiable, strongly convex function and t is some appropriately small stepsize. This yields the noisy gradient descent method

$$T_\eta(x) = x - t\nabla F(x) + \eta.$$

Though this has no fixed point, the additive noise η can be constructed so that the resulting Markov chain is ergodic and its (unique!) invariant distribution concentrates around the unique global minimum of F. This example is purely for illustration. The stochastic gradient or stochastic approximation technique traces its origins to the seminal paper of Robbins and Monro [51]. This has been studied as a purely stochastic method for stochastic optimization in [26, 47]; more recent studies include [25]. We emphasize here that our results *do not* address the issue of convergence of limiting distributions for the RFI as the step size vanishes, $t \to 0$.

The inconsistency of the problem formulation is an artifact of asking the wrong question. A fixed point of the (dual) Markov operator $\mathcal{P}$ is called an *invariant* measure, i.e. $\pi \in \mathscr{P}(\mathcal{E})$ is invariant whenever $\pi\mathcal{P} = \pi$. The set of all invariant probability measures is denoted by $\operatorname{inv}\mathcal{P}$. We are interested in the following generalization of (5):

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

We refer to this as the *stochastic fixed point problem.*

2.3. Modes of convergence

In [31], we considered almost sure convergence of the sequence (X_k) to a random variable X:

$$X_k \to X \text{ a.s.} \quad \text{as } k \to \infty.$$

Almost sure convergence is commonly encountered in the studies of stochastic algorithms in optimization, and can be guaranteed for consistent stochastic feasibility under most of the

regularity assumptions on T_i considered here (see [31, Theorem 3.8 and 3.9]) though this does not require the full power of the theory of general Markov processes. In fact, the next result shows that almost sure convergence is possible only for consistent stochastic feasibility. The following statement first appeared in Lemma 3.2.1 of [30].

Proposition 2.5 (a.s. convergence implies consistency)**.** *Let $T_i : \mathcal{E} \to \mathcal{E}$ be continuous for all $i \in I$. Let $\pi \in \operatorname{inv}\mathcal{P} \neq \emptyset$ and $X_0 \sim \pi$. Generate the sequence $(X_k^{X_0})_{k\in\mathbb{N}}$ via Algorithm 1 where $X_0 \perp\!\!\!\perp \xi_k$ for all k. If the sequence $(X_k^{X_0})$ converges almost surely, then the stochastic feasibility problem is consistent. Moreover,* $\operatorname{supp}\pi \subset C$.

Before proceeding to the proof, note that the *measure* remains the same for each iterate $X_k^{X_0}$ - the issue here is when the *iterates* converge (almost surely).

Proof. In preparation for our argument, which is by contradiction, choose any $x \in \operatorname{supp}\pi$ where $\pi \in \operatorname{inv}\mathcal{P}$, and define

$$I^\epsilon := \{i \in I \,|\, \|T_i x - x\| > \epsilon\}$$

for $\epsilon \geq 0$. Note that $I^\epsilon \supseteq I^{\epsilon_0}$ whenever $\epsilon \leq \epsilon_0$. Define the set

$$J_\delta^\epsilon := \left\{ i \in I \;\middle|\; \|T_i x - T_i y\| \leq \epsilon, \quad \forall y \in \overline{\mathbb{B}}\,(x,\delta) \right\}.$$

These sets satisfy $J_{\delta_1}^\epsilon \subset J_{\delta_2}^\epsilon$ whenever $\delta_1 \geq \delta_2$ and, since T_i is continuous for all $i \in I$, we have that for each $\epsilon > 0$, $J_\delta^\epsilon \uparrow I$ as $\delta \to 0$. A short argument shows that I^ϵ and J_δ^ϵ are measurable for each δ and $\epsilon > 0$.

Suppose now, to the contrary, that $C = \emptyset$. Then $\mathbb{P}(T_\xi x = x) < 1$ and hence $\mathbb{P}(\|T_\xi x - x\| > 0) > 0$. Since $I^\epsilon \supseteq I^{\epsilon_0}$ for $\epsilon \leq \epsilon_0$ we have $\mathbb{P}^\xi(I^{\epsilon_0}) \leq \mathbb{P}^\xi(I^\epsilon)$ whenever $\epsilon \leq \epsilon_0$. In particular, there must exist an ϵ_0 such that $0 < \mathbb{P}^\xi(I^{\epsilon_0})$. On the other hand, there is a constant $\delta > 0$ such that $\delta < \epsilon_0/2$ and $\mathbb{P}^\xi(K_\delta^{\epsilon_0}) > 0$ where $K_\delta^{\epsilon_0} := I^{\epsilon_0} \cap J_\delta^{\epsilon_0/2}$. This construction then yields

$$(\forall i \in K_\delta^{\epsilon_0}) \qquad \|T_i y - x\| \geq \|T_i x - x\| - \|T_i y - T_i x\| \geq \frac{\epsilon_0}{2} > \delta \quad \forall y \in \overline{\mathbb{B}}(x,\delta).$$

Next, we claim that $X_k^{X_0} \sim \pi$ for all $k \in \mathbb{N}$. Indeed, for any $Y \sim \pi$, if ξ is independent of Y, then $T_\xi Y \sim \pi$. To see this, note that For $A \in \mathcal{B}(\mathcal{E})$ Fubini's Theorem and disintegration yield

$$\begin{aligned}\mathbb{P}(T_\xi Y \in A) &= \mathbb{E}[\mathbb{E}\left[\mathbb{1}_A(T_\xi Y) \,|\, \xi\right]] = \mathbb{E}\int \mathbb{1}_A(T_\xi y)\pi(\mathrm{d}y) = \int\int \mathbb{1}_A(z)\mathbb{P}^{T_\xi y}(\mathrm{d}z)\pi(\mathrm{d}y) \\ &= \int\int \mathbb{1}_A(z)p(y,\mathrm{d}z)\pi(\mathrm{d}y) = \pi\mathcal{P}(A) = \pi(A) = \mathbb{P}(Y \in A).\end{aligned}$$

It follows that $\operatorname{supp}\mathcal{L}(Y) = \operatorname{supp}\mathcal{L}(T_\xi Y)$, and since ξ_k are i.i.d, $X_k^Y \sim \pi$ for all $k \in \mathbb{N}$. This establishes the claim.

The independence of ξ_k and X_0 for all k implies the independence of ξ_k and $X_k^{X_0}$ for all k. Moreover, $\mathbb{P}\left(X_k^{X_0} \in B_\delta\right) = \pi(B_\delta) > 0$ for all $k \in \mathbb{N}$, where to avoid clutter we denote $B_\delta := \mathbb{B}(x,\delta)$. This yields

$$(\forall k \in \mathbb{N}) \quad \mathbb{P}\left(X_k^{X_0} \in B_\delta, X_{k+1}^{X_0} \notin B_\delta\right) \geq \mathbb{P}\left(X_k^{X_0} \in B_\delta, \xi_k \in K_\delta^{\epsilon_0}\right) = \pi(B_\delta)\mathbb{P}^\xi(K_\delta^{\epsilon_0}) > 0.$$

Thus, we also have

$$\begin{aligned}(\forall k \in \mathbb{N}) \quad \mathbb{P}\left(X_k^{X_0} \in B_\delta, X_{k+1}^{X_0} \in B_\delta\right) &= \mathbb{P}\left(X_k^{X_0} \in B_\delta\right) - \mathbb{P}\left(X_k^{X_0} \in B_\delta, X_{k+1}^{X_0} \notin B_\delta\right) \\ &\leq \pi(B_\delta) - \pi(B_\delta)\mathbb{P}^\xi(K_\delta^{\epsilon_0}) \\ &= \pi(B_\delta)(1 - \mathbb{P}^\xi(K_\delta^{\epsilon_0})) < \pi(B_\delta).\end{aligned}$$

However, by assumption, $X_k^{X_0} \to X_*$ a.s. for some random variable X_* with $\mathbb{P}(X_* \in B_\delta) = \pi(B_\delta) > 0$. If $X_* \in B_\delta$ then due to the a.s. convergence there exists a (random) k_* such that $X_k^{X_0} \in B_\delta$ for all $k \geq k_*$. This implies that

$$\mathbb{P}\left(X_k^{X_0} \in B_\delta, X_{k+1}^{X_0} \in B_\delta, X_* \in B_\delta\right) \to \mathbb{P}(X_* \in B_\delta) = \pi(B_\delta).$$

which is a contradiction since by the above

$$\mathbb{P}\left(X_k^{X_0} \in B_\delta, X_{k+1}^{X_0} \in B_\delta, X_* \in B_\delta\right) \leq \mathbb{P}\left(X_k^{X_0} \in B_\delta, X_{k+1}^{X_0} \in B_\delta\right) < \pi(B_\delta).$$

So it must be true that $\mathbb{P}(\|T_\xi x - x\| > 0) = 0$. In other words, $\mathbb{P}(T_\xi x = x) = 1$, hence $C \neq \emptyset$. Moreover, since the point x was any arbitrary point in $\operatorname{supp}\pi$, we conclude that $\operatorname{supp}\pi \subset C$. □

For inconsistent feasibility, or more general stochastic fixed point problems that are the aim of the RFI, Algorithm 1, we focus on *convergence in distribution*. Let (ν_k) be a sequence of probability measures on $\mathcal{E}$. The sequence (ν_k) is said to converge in distribution to ν whenever $\nu \in \mathscr{P}(\mathcal{E})$ and for all $f \in C_b(\mathcal{E})$ it holds that $\nu_k f \to \nu f$ as $k \to \infty$, where $\nu f := \int f(x)\nu(\mathrm{d}x)$. Equivalently a sequence of random variables (X_k) is said to converge in distribution if their laws $(\mathcal{L}(X_k))$ do.

We now consider two modes of convergence in distribution for the corresponding sequence of measures $(\mathcal{L}(X_k)))_{k\in\mathbb{N}}$ on $\mathscr{P}(\mathcal{E})$:

1. convergence in distribution of the Cesàro averages of the sequence $(\mathcal{L}(X_k))$ to a probability measure $\pi \in \mathscr{P}(\mathcal{E})$, i.e. for any $f \in C_b(\mathcal{E})$

$$\nu_k f := \frac{1}{k}\sum_{j=1}^{k} \mathcal{L}(X_j) f = \mathbb{E}\left[\frac{1}{k}\sum_{j=1}^{k} f(X_j)\right] \to \pi f, \qquad \text{as } k \to \infty;$$

2. convergence in distribution of the sequence $(\mathcal{L}(X_k))$ to a probability measure $\pi \in \mathscr{P}(\mathcal{E})$, i.e. for any $f \in C_b(\mathcal{E})$

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

Clearly, the second mode of convergence implies the first. This is used in Section 3.5 and Section 3.7.

An elementary fact from the theory of Markov chains (Proposition 3.1) is that, if the Markov operator $\mathcal{P}$ is Feller and π is a cluster point of (ν_k) with respect to convergence in distribution then π is an invariant probability measure. Existence of invariant measures for a Markov operator then amounts to verifying that the operator is Feller (by Proposition 2.4, automatic if the T_i are continuous) and that cluster points exist (guaranteed by *tightness* – or compactness with respect to the topology of convergence in distribution – of the sequence, see [10, Section 5]. In particular, this means that there exists a convergent subsequence (ν_{k_j}) with

$$(\forall f \in C_b(\mathcal{E})) \qquad \nu_{k_j} f = \mathbb{E}\left[\frac{1}{k_j}\sum_{i=1}^{k_j} f(X_i)\right] \to \pi f, \qquad \text{as } j \to \infty.$$

Convergence of the whole sequence, i.e. $\nu_k \to \pi$, amounts then to showing that π is the unique cluster point of (ν_k) (see Proposition A.1).

Common metrics for spaces of measures are the *Prokhorov-Lèvy distance* and the *Wasserstein metric*.

Definition 2.6 (Prokhorov-Lèvy&Wasserstein distance)**.** Let (G, d) be a separable complete metric space and let $\mu, \nu \in \mathscr{P}(G)$.

(i) The Prokhorov-Lèvy distance, denoted by d_P, is defined by

$$d_P(\mu,\nu) = \inf\left\{\epsilon > 0 \,|\, \mu(A) \le \nu(\mathbb{B}(A,\epsilon)) + \epsilon,\ \nu(A) \le \mu(\mathbb{B}(A,\epsilon)) + \epsilon \quad \forall A \in \mathcal{B}(G)\right\}. \quad (7)$$

(ii) For $p \ge 1$ let

$$\mathscr{P}_p(G) = \left\{\mu \in \mathscr{P}(G) \,\middle|\, \exists\, x \in G \,:\, \int d^p(x,y)\mu(\mathrm{d}y) < \infty\right\}. \quad (8)$$

The Wasserstein p-metric on $\mathscr{P}_p(G)$, denoted W_p, is defined by

$$W_p(\mu,\nu) := \left(\inf_{\gamma \in C(\mu,\nu)} \int_{G\times G} d^p(x,y)\gamma(dx,dy)\right)^{1/p} \qquad (p \ge 1) \quad (9)$$

where $C(\mu,\nu)$ is the set of couplings of μ and ν (measures on the product space $G \times G$ whose marginals are μ and ν respectively - see (44)).

For probability measures on separable metric spaces the Prokhorov-Lèvy distance metrizes weak convergence (convergence in distribution). Convergence in Wasserstein distance is stronger than convergence in distribution (the p-th moments of the measures also converge in the p-Wasserstein metric). Indeed, it can be seen from the metrics that

$$d_P(\mu,\nu)^2 \le W_p(\mu,\nu)^p \quad (p \ge 1).$$

2.4. Regularity

Our main results concern convergence of Markov chains under common regularity assumptions on the mappings $\{T_i\}$. The regularity of T_i is dictated by the application, and our primary interest is to follow this through to the regularity of the corresponding Markov operator. In [41] a framework was developed for a quantitative convergence analysis of set-valued mappings T_i that are *calm* (one-sided Lipschitz continuous in the sense of set-valued-mappings) with Lipschitz constant possibly greater than 1. Since we restrict our attention to nonexpansive operators (Lipschitz constant 1), we do not need the added complication of set-valued mappings, but this theory can be extended to such settings.

The definitions below are simplified versions of the analogous properties in more general settings (see [9, 38].)

Definition 2.7 (pointwise nonexpansive and averaged mappings in $\mathcal{E}$)**.** Let $F : D \to \mathcal{E}$ where $D \subset \mathcal{E}$.

(i) The mapping F is said to be *pointwise nonexpansive at* $x_0 \in D$ *on* D whenever

$$\|F(x) - F(x_0)\| \le \|x - x_0\|, \qquad \forall x \in D. \quad (10)$$

When the above inequality holds for all $x_0 \in D$ then F is said to be *nonexpansive on* D.

(ii) The mapping F is said to be *pointwise averaged at* $x_0 \in D$ *on* D whenever

$$\exists \alpha \in (0,1) : \quad \|F(x) - F(x_0)\|^2 \le \|x - x_0\|^2 - \tfrac{1-\alpha}{\alpha}\psi(x, x_0, F(x), F(x_0)) \quad \forall x \in D$$

where the *transport discrepancy* ψ of F at x, x_0, $F(x)$ and $F(x_0)$ is defined by

$$\psi(x, x_0, F(x), F(x_0)) := \|(F(x) - x) - (F(x_0) - x_0)\|^2. \tag{11}$$

When the above inequality holds for all $x_0 \in D$ then F is said to be *averaged on* D.

Our definition of averaged mappings differs from the standard definition in that it places the transport discrepancy ψ at the center. Equation (11) shows, at least in a Hilbert space setting, that ψ is nonnegative. Consequently, any pointwise averaged mapping on $\mathcal{E}$ is also pointwise nonexpansive. This is not the case in more general metric spaces.

Our definition of averaged mappings with $\alpha = 1/2$ is equivalent to the definition of *firmly contractive* mappings given in [13, Definition 6]. We conform to the dominant terminology in linear spaces originating with [4]. We note, however, that the descriptor "averaging" is a misnomer since the same notion plays a central role in general metric spaces where the concept of averaging is vacuous; in metric space settings such mappings are called α*-firmly nonexpansive* [1, 9, 38].

In normed linear spaces, Baillon and Bruck [3] showed that nonexpansive mappings whose orbits are bounded under convex relaxations are *asymptotically regular* with a universal rate constant. Precisely: let $T : D \to D$ be nonexpansive, where D is a convex subset of a normed linear space, and define x_m recursively by $x_m = T_\lambda x_{m-1} := ((1-\lambda)\,\mathrm{Id} + \lambda T)$ for $\lambda \in (0,1)$ and $x_0 \in D$. If $\|x - TT_\lambda^k x\| \leq 1$ for any $\lambda \in (0,1)$ and for all $0 \leq k \leq m$, then [3, Main Result]

$$\|x_m - Tx_m\| < \frac{\operatorname{diam} D}{\sqrt{\pi m \lambda (1-\lambda)}}. \tag{12}$$

Cominetti, Soto and Vaisman [23] recently confirmed a conjecture of Baillon and Bruck that a universal rate constant also holds for nonexpansive mappings with arbitrary relaxation in $(0,1)$ chosen at each iteration; in particular, that

$$\|x_m - Tx_m\| \leq \frac{\operatorname{diam} D}{\sqrt{\pi \sum_{k=1}^m \lambda_k (1-\lambda_k)}},$$

where x_m is defined recursively by $x_m = T_{\lambda_m} x_{m-1} := ((1-\lambda_m)\,\mathrm{Id} + \lambda_m T)\, x_{m-1}$ for $\lambda_m \in (0,1)$ $(m = 1, 2, \dots)$. The operators T_{λ_m} all have the same set of fixed points (namely $\operatorname{Fix} T$), so these results are complementary to [31] where it was shown [31, Theorem 3.5] that sequences of random variables on compact metric spaces generated by Algorithm 1 with *paracontractions* such as T_{λ_m} above converge *almost surely* to a random variable in $\operatorname{Fix} T$, assuming that this is nonempty (see Proposition 2.5 in this context). Necessary and sufficient conditions for linear convergence of the *iterates* were also determined in a more limited setting in [31, Theorems 3.11 and 3.15]. The results of [3, 23, 31], however, do not apply to inconsistent stochastic feasibility considered here.

2.5. Main Results

All of our main results concern Markov operators $\mathcal{P}$ with update function $\Phi(x, i) = T_i(x)$ and transition kernel p given by (4) for self mappings $T_i : \mathcal{E} \to \mathcal{E}$. For any $\mu_0 \in \mathscr{P}_2(\mathcal{E})$, we denote the distributions of the iterates of Algorithm 1 by $\mu_k = \mu_0 \mathcal{P}^k = \mathcal{L}(X_k)$, and we denote $d_{W_2}(\mu_k, \operatorname{inv} \mathcal{P}) := \inf_{\pi' \in \operatorname{inv} \mathcal{P}} W_2(\mu_k, \pi')$.

In most of our main results, it will be assumed that $\operatorname{inv} \mathcal{P} \neq \emptyset$. The existence theory is already well developed and is surveyed in Section 3.1 below. The main convergence result for nonexpansive mappings follows from a fundamental result of Worm [61, Theorem 7.3.13].

Theorem 2.8 (convergence of Cesàro average in $\mathcal{E}$)**.** *Let $T_i : \mathcal{E} \to \mathcal{E}$ be nonexpansive ($i \in I$) and assume $\operatorname{inv}\mathcal{P} \neq \emptyset$. Let $\mu \in \mathscr{P}(\mathcal{E})$ and $\nu_k = \frac{1}{k}\sum_{j=1}^{k} \mu\mathcal{P}^j$, then this sequence converges in the Prokhorov-Lèvy metric to an invariant probability measure for $\mathcal{P}$, i.e. $\nu_k \to \pi^\mu$ where*

$$\pi^\mu = \int_{\operatorname{supp}\mu} \pi^x \mu(\mathrm{d}x), \tag{13}$$

where for each $x \in \operatorname{supp}\mu \subset \mathbb{R}^n$ there exists the limit of (ν_k^x) and it is denoted by the invariant measure π^x.

When the mappings are averaged, we obtain the following stronger result. It is worth pointing interested readers to an analogous metric space result of [9, Theorem 27] in which it is shown that, on p-uniformly convex spaces, sequences generated by fixed point iterations of compositions of pointwise averaged mappings T_i converge in a weak sense whenever $\cap_i \operatorname{Fix} T_i$ is nonempty. When the composition is boundedly compact, then the fixed point iteration converges strongly to a fixed point.

Theorem 2.9 (convergence for averaged mappings on $\mathcal{E}$)**.** *Let $T_i : \mathcal{E} \to \mathcal{E}$ be averaged with constant $\alpha_i \leq \alpha < 1$ ($i \in I$). Assume $\operatorname{inv}\mathcal{P} \neq \emptyset$. For any initial distribution $\mu_0 \in \mathscr{P}(\mathcal{E})$ the distributions μ_k of the iterates generated by Algorithm 1 converge in the Prokhorov-Lèvy metric to an invariant probability measure for $\mathcal{P}$.*

The proof of this result is very different than the strategy applied to the analogous metric space result of [9, Theorem 27].

3. Background Theory and Proofs

In this section we prepare tools to prove the main results from Section 2.5. We start by establishing convergence results on the supports of ergodic measures on a general Polish space G (a complete separable metric space), and then, for global convergence analysis of Algorithm 1, we restrict ourselves to $\mathcal{E}$. We begin with existence of invariant measures. We then analyze properties of (and convergence of the RFI on) subsets of G, called ergodic sets. Then we turn our attention to the global convergence analysis.

3.1. Existence of Invariant Measures

A sequence of probability measures (ν_k) is called *tight* if for any $\epsilon > 0$ there exists a compact $K \subset G$ with $\nu_k(K) > 1 - \epsilon$ for all $k \in \mathbb{N}$. By Prokhorov's theorem (see, for instance, [10]), a sequence $(\nu_k) \subset \mathscr{P}(G)$, for G a Polish space, is tight if and only if (ν_k) is compact in $\mathscr{P}(G)$, i.e. any subsequence of (ν_k) has a subsequence that converges in distribution (see, for instance [10]).

A basic building block is the existence of invariant measures proved by Lasota and T. Szarek [37, Proposition 3.1]. Based on this, we show how existence can be verified easily. But first, we show how to obtain existence constructively.

Proposition 3.1 (construction of an invariant measure)**.** *Let $\mu \in \mathscr{P}(G)$ and $\mathcal{P}$ be a Feller Markov operator. Let $(\mu\mathcal{P}^k)_{k\in\mathbb{N}}$ be a tight sequence of probability measures on a Polish space G, and let $\nu_k = \frac{1}{k}\sum_{j=1}^{k} \mu\mathcal{P}^j$. Any cluster point of the sequence $(\nu_k)_{k\in\mathbb{N}}$ is an invariant measure for $\mathcal{P}$.*

Proof. Our proof follows [28, Theorem 1.10]. Tightness of the sequence $(\mu\mathcal{P}^k)$ implies tightness of the sequence (ν_k) and therefore by Prokhorov's Theorem there exists a convergent subsequence (ν_{k_j}) with limit $\pi \in \mathscr{P}(G)$. By the Feller property of $\mathcal{P}$ one has for any continuous and bounded $f : G \to \mathbb{R}$ that also $\mathcal{P}f$ is continuous and bounded, and hence

$$\begin{aligned} |(\pi\mathcal{P})f - \pi f| &= |\pi(\mathcal{P}f) - \pi f| \\ &= \lim_j \left|\nu_{k_j}(\mathcal{P}f) - \nu_{k_j} f\right| \\ &= \lim_j \frac{1}{k_j}\left|\mu\mathcal{P}^{k_j+1} f - \mu\mathcal{P}f\right| \\ &\leq \lim_j \frac{2\|f\|_\infty}{k_j} \\ &= 0. \end{aligned}$$

Now, $\pi f = (\pi\mathcal{P})f$ for all $f \in C_b(G)$ implies that $\pi = \pi\mathcal{P}$. □

When a Feller Markov chain converges in distribution (i.e. $\mu\mathcal{P}^k \to \pi$), it does so to an invariant measure (since $\mu\mathcal{P}^{k+1} \to \pi\mathcal{P}$). A Markov operator need not possess a unique invariant probability measure or any invariant measure at all. Indeed, for the case that $T_i = P_i$, $i \in I$ is a projector onto a nonempty closed and convex set $C_i \subset \mathcal{E}$. A sufficient condition for the deterministic Alternating Projections Method to converge in the inconsistent case to a limit cycle for convex sets is that one of the sets is compact (this is an easy consequence of [20, Theorem 4]). Translating this into the present setting, a sufficient condition for the existence of an invariant measure for $\mathcal{P}$ is the existence of a compact set $K \subset \mathcal{E}$ and $\epsilon > 0$ such that $p(x, K) \geq \epsilon$ for all $x \in \mathcal{E}$. This holds, for instance, when there are only finitely many sets with one of them, say $C_{\bar{i}}$, compact and $\mathbb{P}(\xi = \bar{i}) = \epsilon$, since $p(x, C_{\bar{i}}) = \mathbb{P}(P_\xi x \in C_{\bar{i}}) \geq \mathbb{P}(P_\xi x \in C_{\bar{i}}, \xi = \bar{i}) = \mathbb{P}(\xi = \bar{i}) = \epsilon$ for all $x \in \mathcal{E}$. More generally, we have the following result.

Proposition 3.2 (existence of invariant measures for finite collections of continuous mappings)**.** *Let G be a Polish space and let $T_i : G \to G$ be continuous for $i \in I$, where I is a finite index set. If for one index $i \in I$ it holds that $\mathbb{P}(\xi = i) > 0$ and $T_i(G) \subset K$, where $K \subset G$ is compact, then there exists an invariant measure for $\mathcal{P}$.*

Proof. We have from $T_i(G) \subset K$ that $\mathbb{P}(T_\xi x \in K) \geq \mathbb{P}(\xi = i)$ and hence for the sequence (X_k) generated by Algorithm 1 for an arbitrary initial probability measure

$$\mathbb{P}(X_{k+1} \in K) = \mathbb{E}[\mathbb{P}\left(T_{\xi_k} X_k \in K \,|\, X_k\right)] \geq \mathbb{P}(\xi = i) \qquad \forall k \in \mathbb{N}.$$

The assertion follows now immediately from [37, Proposition 3.1] since $\mathbb{P}(\xi = i) > 0$ and $\mathcal{P}$ is Feller by continuity of T_j for all $j \in I$. □

Next we mention an existence result which requires that the RFI sequence (X_k) possess a uniformly bounded expectation.

Proposition 3.3 (existence in $\mathcal{E}$, RFI)**.** *Let $T_i : \mathcal{E} \to \mathcal{E}$ $(i \in I)$ be continuous. Let (X_k) be the RFI sequence (generated by Algorithm 1) for some initial measure. Suppose that for all $k \in \mathbb{N}$ it holds that $\mathbb{E}\left[\|X_k\|\right] \leq M$ for some $M \geq 0$. Then there exists an invariant measure for the RFI Markov operator $\mathcal{P}$ given by (4).*

Proof. For any $\epsilon > M$ Markov's inequality implies that

$$\mathbb{P}(\|X_k\| \geq \epsilon) \leq \frac{\mathbb{E}\left[\|X_k\|\right]}{\epsilon} \leq \frac{M}{\epsilon} < 1$$

Hence,

$$\limsup_{k\to\infty} \mathbb{P}(\|X_k\| \leq \epsilon) \geq \limsup_{k\to\infty} \mathbb{P}(\|X_k\| < \epsilon) \geq 1 - \frac{M}{\epsilon} > 0.$$

Existence of an invariant measure then follows from [37, Proposition 3.1] since closed balls in $\mathcal{E}$ with finite radius are compact, $\mathbb{P}(X_k \in \cdot) = \mu\mathcal{P}^k$ and continuity of T_i yields the Feller property for $\mathcal{P}$. □

3.2. Ergodic theory of general Markov Operators

Recall that an invariant probability measure π of a Markov operator $\mathcal{P}$ is called *ergodic*, if any *p-invariant set*, i.e. $A \in \mathcal{B}(G)$ with $p(x, A) = 1$ for all $x \in A$, has π-measure 0 or 1. In this section we study the properties of the RFI Markov chain when it is initialized by a distribution in the support of any ergodic measure for $\mathcal{P}$. The convergence properties for these points can be much stronger than the convergence properties of Markov chains initialized by measures with support outside the support of the ergodic measures.

The consistent stochastic feasibility problem was analyzed in [31] without the need of the notion of convergence of measures since, as shown in Proposition 2.5, for consistent stochastic feasibility convergence of sequences defined by (3) is almost sure, if they converge at all. More general convergence of measures is more challenging as the next example illustrates.

Example 3.4 (nonexpansive mappings, negative result)**.** For nonexpansive mappings in general, one cannot expect that the sequence $(\mathcal{L}(X_k))_{k\in\mathbb{N}}$ converges to an invariant probability measure. Consider the nonexpansive operator $T := T_1 x := -x$ on $\mathbb{R}$ and set, in the RFI setup, $\xi = 1$ and $I = \{1\}$. Then $X_{2k} = x$ and $X_{2k+1} = -x$ for all $k \in \mathbb{N}$, if $X_0 \sim \delta_x$. This implies for $x \neq 0$ that $(\mathcal{L}(X_k))$ does not converge to the invariant distribution $\pi_x = \frac{1}{2}(\delta_x + \delta_{-x})$ (depending on x), since $\mathbb{P}(X_{2k} \in B) = \delta_x(B)$ and $\mathbb{P}(X_{2k+1} \in B) = \delta_{-x}(B)$ for $B \in \mathcal{B}(\mathbb{R})$. Nevertheless the Cesàro average $\nu_k := \frac{1}{k}\sum_{j=1}^{k} \mathbb{P}^{X_j}$ converges to π_x.

As Example 3.4 shows, meaningful notions of ergodic convergence are possible (in our case, convergence of the Cesàro average) even when convergence in distribution can not be expected. We start by collecting several general results for Markov chains on Polish spaces. In the next section we restrict ourselves to equicontinuous and Feller Markov operators.

The following decomposition theorem on Polish spaces is key to our development. Two measures π_1, π_2 are called *mutually singular* when there is $A \in \mathcal{B}(G)$ with $\pi_1(A^c) = \pi_2(A) = 0$. For more detail see, for instance, [60].

Proposition 3.5. *Denote by $\mathcal{I}$ the set of all invariant probability measures for $\mathcal{P}$ and by $\mathscr{E} \subset \mathcal{I}$ the set of all those that are ergodic. Then, $\mathcal{I}$ is convex and $\mathscr{E}$ is precisely the set of its extremal points. Furthermore, for every invariant measure $\pi \in \mathcal{I}$, there exists a probability measure q_π on $\mathscr{E}$ such that*

$$\pi(A) = \int_{\mathscr{E}} \nu(A) q_\pi(\mathrm{d}\nu).$$

In other words, every invariant measure is a convex combination of ergodic invariant measures. Finally, any two distinct elements of $\mathscr{E}$ are mutually singular.

Remark 3.6: If there exists only one invariant probability measure of $\mathcal{P}$, we know by Proposition 3.5 that it is ergodic. If there exist more invariant probability measures, then there exist uncountably many invariant and at least two ergodic probability measures.

Proposition 3.7. *Let π be an ergodic invariant probability measure for $\mathcal{P}$, let $(G, \mathcal{G})$ be a measurable space, and let $f : G \to \mathbb{R}$ be measurable, bounded and satisfy $\pi|f|^p < \infty$ for $p \in [1, \infty]$. Then*

$$\nu_k^x f := \frac{1}{k} \sum_{j=1}^{k} p^j(x, f) \to \pi f \qquad \text{as } k \to \infty \qquad \text{for } \pi\text{-a.e. } x \in G,$$

where $p^j(x, f) := \delta_x \mathcal{P}^j f = \mathbb{E}\left[f(X_j) \,|\, X_0 = x\right]$ for the sequence (X_k) generated by Algorithm 1 with $X_0 \sim \pi$.

Proof. This is a direct consequence of Birkhoff's ergodic theorem, [34, Theorem 9.6]. □

For fixed x in Proposition 3.7, we want the assertion to be true for all $f \in C_b(G)$. This issue is addressed in the next section by restricting our attention to equicontinuous Markov operators. The results above do not require any explicit structure on the mappings T_i that generate the transition kernel p and hence the Markov operator $\mathcal{P}$, however the assumption that the initial random variable X_0 has the same distribution as the invariant measure π is very strong. For Markov operators generated from discontinuous mappings T_i, the support of an invariant measure may not be invariant under T_ξ. To see this, let

$$Tx := \begin{cases} x, & x \in \mathbb{R} \setminus \mathbb{Q} \\ -1, & x \in \mathbb{Q}. \end{cases}$$

The transition kernel is then $p(x, A) = \mathbb{1}_A(Tx)$ for $x \in \mathbb{R}$ and $A \in \mathcal{B}(\mathbb{R})$. Let μ be the uniform distribution on $[0, 1]$, then, since λ-a.s. $T = \mathrm{Id}$ (where λ is the Lebesgue measure on $\mathbb{R}$), we have that $\mu\mathcal{P}^k = \mu$ for all $k \in \mathbb{N}$. Consequently, $\pi = \mu$ is invariant and $\operatorname{supp} \pi = [0, 1]$, but $T([0, 1]) = \{-1\} \cup [0, 1] \cap (\mathbb{R} \setminus \mathbb{Q})$, which is not contained in $[0, 1]$.

The next result shows, however, that invariance of the the support of invariant measures under *continuous mappings* T_i is guaranteed.

Lemma 3.8 (invariance of the support of invariant measures)**.** *Let G be a Polish space and let $T_i : G \to G$ be continuous for all $i \in I$. For any invariant probability measure $\pi \in \mathscr{P}(G)$ of $\mathcal{P}$ it holds that $T_\xi S_\pi \subset S_\pi$ a.s. where $S_\pi := \operatorname{supp} \pi$.*

Proof. By Fubini's Theorem, for any $A \in \mathcal{B}(G)$ it holds that

$$\begin{aligned} \pi(A) = \int_{S_\pi} p(x, A)\pi(\mathrm{d}x) &= \int_\Omega \int_{S_\pi} \mathbb{1}_A(T_\xi x)\pi(\mathrm{d}x)\,\mathrm{d}\mathbb{P} \\ &= \int_\Omega \int_{S_\pi} \mathbb{1}_{T_{\xi(\omega)}^{-1} A}(x)\pi(\mathrm{d}x)\mathbb{P}(\mathrm{d}\omega) \\ &= \mathbb{E}\left[\pi(T_\xi^{-1} A \cap S_\pi)\right] = \mathbb{E}\left[\pi(T_\xi^{-1} A)\right]. \end{aligned}$$

From $1 = \pi(S_\pi) = \mathbb{E}\left[\pi(T_\xi^{-1} S_\pi)\right]$ and $\pi(\cdot) \leq 1$, it follows that $\pi(T_\xi^{-1} S_\pi) = 1$ a.s.

Note that $T_i^{-1} S_\pi$ is closed for all $i \in I$ due to continuity of T_i and closedness of S_π. We show that $S_\pi \subset T_\xi^{-1} S_\pi$ a.s. which then yields the claim. To see this, let $S \subset G$ be any closed set with $\pi(A \cap S) = \pi(A)$ for all $A \in \mathcal{B}(G)$, and let $x \in S_\pi$. Then $\pi(\mathbb{B}(x, \epsilon) \cap S) > 0$ for all $\epsilon > 0$, i.e. $\mathbb{B}(x, \epsilon) \cap S \neq \emptyset$ for all $\epsilon > 0$. Now consider $x_k \in \mathbb{B}(x, \epsilon_k) \cap S$, where $\epsilon_k \to 0$ as $k \to \infty$. Then since S is closed, $x_k \to x \in S$, from which we conclude that $S_\pi \subset S$. Specifically, let $S = T^{-1} S_\pi$ and note that $T^{-1} S_\pi = D \setminus G$ for some D with $\pi(D) = 0$. For any $A \in \mathcal{B}(G)$ it holds that $\pi(A \cap S) = \pi(A) - \pi(A \cap D) = \pi(A)$. From the argument above, we conclude that $S_\pi \subset S = T^{-1} S_\pi$ as claimed. □

The above result means that, if the random variable X_k enters S_π for some k, then it will stay in S_π forever. This can be interpreted as a mode of convergence, i.e. convergence to the set S_π, which is closed under application of T_ξ a.s. Equality $T_\xi S_\pi = S_\pi$ a.s. cannot be expected in general. For example, let $I = \{1, 2\}$, $G = \mathbb{R}$ and $T_1 x = -1$, $T_2 x = 1$, $x \in \mathbb{R}$ and $\mathbb{P}(\xi = 1) = 0.5 = \mathbb{P}(\xi = 2)$, then $\pi = \frac{1}{2}(\delta_{-1} + \delta_1)$ and $S_\pi = \{-1, 1\}$. So $T_1 S_\pi = \{-1\}$ and $T_2 S_\pi = \{1\}$.

3.3. Ergodic convergence theory for equicontinuous Markov operators

As shown by Szarek [57] and Worm [61], equicontinuity of Markov operators and their generalizations give a nice structure to the set of ergodic measures. We collect some results here which will be used heavily in the subsequent analysis.

Definition 3.9 (equicontinuity)**.** A Markov operator is called equicontinuous if $(\mathcal{P}^k f)_{k\in\mathbb{N}}$ is equicontinuous for all bounded and Lipschitz continuous $f : G \to \mathbb{R}$.

In the following we consider the union of supports of all ergodic measures defined by

$$S := \bigcup_{\pi\in\mathscr{E}} \operatorname{supp} \pi, \tag{14}$$

where $\mathscr{E} \subset \operatorname{inv} \mathcal{P}$ denotes the set of ergodic measures.

Proposition 3.10 (tightness of $(\delta_s \mathcal{P}^k)$)**.** *Let G be a Polish space. Let $\mathcal{P}$ be equicontinuous. Suppose there exists an invariant measure for $\mathcal{P}$. Then the sequence $(\delta_s \mathcal{P}^k)_{k\in\mathbb{N}}$ is tight for all $s \in S$ defined by (14).*

Proof. In the proof of [57, Proposition 2.1] it is shown that, under the assumption that $\mathcal{P}$ is equicontinuous and

$$(\exists s, x \in G) \quad \limsup_{k\to\infty} \nu_k^x(\mathbb{B}(s, \epsilon)) > 0 \quad \forall \epsilon > 0, \tag{15}$$

then the sequence $(\delta_s \mathcal{P}^k)$ is tight. It remains to demonstrate (15). To see this, let $f = \mathbb{1}_{\mathbb{B}(s,\epsilon)}$ for some $s \in S_\pi$, where $\pi \in \mathscr{E}$ and $\epsilon > 0$ in Proposition 3.7. Then for π-a.e. $x \in G$ and $\nu_k^x := \frac{1}{k}\sum_{j=1}^k \delta_x \mathcal{P}^j$ we have

$$\limsup_{k\to\infty} \nu_k^x(\mathbb{B}(s, \epsilon)) = \lim_k \nu_k^x(\mathbb{B}(s, \epsilon)) = \pi(\mathbb{B}(s, \epsilon)) > 0.$$

This completes the proof. □

Remark 3.11 (tightness of (ν_k^s))**:** Note that the sequence (ν_k^s) is tight for $s \in S$, since by Proposition 3.10, for all $\epsilon > 0$, there is a compact subset $K \subset G$ such that $p^k(s, K) > 1 - \epsilon$ for all $k \in \mathbb{N}$, and hence also $\nu_k^s(K) > 1 - \epsilon$ for all $k \in \mathbb{N}$.

The next result due to Worm (Theorems 5.4.11 and 7.3.13 of [61])) concerns Cesáro averages for equicontinuous Markov operators.

Proposition 3.12 (convergence of Cesàro averages [61])**.** *Let $\mathcal{P}$ be Feller and equicontinuous, let G be a Polish space and let $\mu \in \mathscr{P}(G)$. Then the sequence (ν_k^μ) is tight ($\nu_k^\mu := \frac{1}{k}\sum_{j=1}^k \mu\mathcal{P}^j$) if and only if (ν_k^μ) converges to a $\pi^\mu \in \operatorname{inv} \mathcal{P}$. In this case*

$$\pi^\mu = \int_{\operatorname{supp}\mu} \pi^x \mu(\mathrm{d}x),$$

where for each $x \in \operatorname{supp}\mu \subset G$ there exists the limit of (ν_n^x) and it is denoted by the invariant measure π^x.

For the case the initial measure μ is supported in $\bigcup_{\pi\in\operatorname{inv}\mathcal{P}}\operatorname{supp}\pi$, we have the following.

Proposition 3.13 (ergodic decomposition)**.** *Let G be a Polish space and let $\mathcal{P}$ be Feller and equicontinuous. Then*

$$S = \bigcup_{\pi\in\operatorname{inv}\mathcal{P}} \operatorname{supp}\pi,$$

where S is defined in (14). Moreover S is closed, and for any $\mu \in \mathscr{P}(S)$ it holds that $\nu_k^\mu \to \pi^\mu$ as $k \to \infty$ with

$$\pi^\mu = \int_S \pi^x \mu(\mathrm{d}x),$$

where π^x is the unique ergodic measure with $x \in \operatorname{supp}\pi^x$.

Proof. This is a consequence of [61, Theorem 7.3.4] and [61, Theorem 5.4.11], Remark 3.11 and Proposition 3.12. □

Proposition 3.13 only establishes convergence of the Markov chain when it is initialized with a measure in the support of an invariant measure; moreover, it is only the average of the distributions of the iterates that converges.

Remark 3.14 (convergence of (ν_k^s) on Polish spaces)**:** Let π be an ergodic invariant probability measure for $\mathcal{P}$. Then for all $s \in \operatorname{supp}\pi$ the sequence $\nu_k^s \to \pi$ as $k \to \infty$, where $\nu_k^s = \frac{1}{k}\sum_{j=1}^k p^j(s,\cdot)$.

By Proposition 3.5 any invariant measure can be decomposed into a convex combination of ergodic invariant measures; in particular two ergodic measures π_1, π_2 are mutually singular. Note that it still could be the case that $\operatorname{supp}\pi_1 \cap \operatorname{supp}\pi_2 \neq \emptyset$. But Remark 3.14 above establishes that for a Feller and equicontinuous Markov operator $\mathcal{P}$ this is not possible, so the singularity of ergodic measures extends to their support. This leads to the following corollary.

Corollary 3.15. *Under the assumptions of Proposition 3.13 for two ergodic measures $\pi, \tilde{\pi}$, the intersection $(\operatorname{supp}\pi) \cap (\operatorname{supp}\tilde{\pi}) = \emptyset$ if and only if $\pi \neq \tilde{\pi}$.*

3.4. Ergodic theory for nonexpansive mappings

We now specialize to the case that the family of mappings $\{T_i\}_{i\in I}$ are nonexpansive operators. The next technical lemma implies that every point in the support of an ergodic measure is reached infinitely often starting from any other point in this support.

Lemma 3.16 (positive transition probability for ergodic measures)**.** *Let G be a Polish space and let $T_i : G \to G$ be nonexpansive, $i \in I$. Let π be an ergodic invariant probability measure for $\mathcal{P}$. Then for any $s, \tilde{s} \in \operatorname{supp}\pi$ it holds that*

$$\forall \epsilon > 0\, \exists \delta > 0,\ \exists (k_j)_{j\in\mathbb{N}} \subset \mathbb{N} : p^{k_j}(s, \mathbb{B}(\tilde{s},\epsilon)) \geq \delta \quad \forall j \in \mathbb{N}.$$

Proof. Given $\tilde{s} \in \operatorname{supp}\pi$ and $\epsilon > 0$, find a continuous and bounded function $f = f_{\tilde{s},\epsilon} : G \to [0,1]$ with the property that $f = 1$ on $\mathbb{B}(\tilde{s}, \frac{\epsilon}{2})$ and $f = 0$ outside $\mathbb{B}(\tilde{s},\epsilon)$. For $s \in \operatorname{supp}\pi$ let $X_0 \sim \delta_s$ and (X_k) generated by Algorithm 1. By Remark 3.14 the sequence (ν_k) converges to π as $k \to \infty$, where $\nu_k := \frac{1}{k}\sum_{j=1}^k p^j(s,\cdot)$. So in particular $\nu_k f \to \pi f \geq \pi(\mathbb{B}(\tilde{s}, \frac{\epsilon}{2})) > 0$ as $k \to \infty$. Hence, for k large enough there is $\delta > 0$ with

$$\nu_k f = \frac{1}{k}\sum_{j=1}^k p^j(s,f) \geq \delta.$$

Now, we can extract a sequence $(k_j) \subset \mathbb{N}$ with $p^{k_j}(s,f) \geq \delta$, $j \in \mathbb{N}$ and hence

$$p^{k_j}(s, \mathbb{B}(\tilde{s}, \epsilon)) \geq p^{k_j}(s,f) \geq \delta > 0. \qquad \square$$

Lemma 3.17. *Let G be a Polish space. Let $T_i : G \to G$ be nonexpansive, $i \in I$ and let $\mathcal{P}$ denote the Markov operator that is induced by the transition kernel in (4).*

(i) $\mathcal{P}$ is Feller.

(ii) $\mathcal{P}$ is equicontinuous.

Proof. (i) The mapping T_i for $i \in I$ is 1-Lipschitz continuous, so in particular it is continuous. Proposition 2.4 yields the assertion.

(ii) Let $\epsilon > 0$ and $x, y \in G$ with $d(x,y) < \epsilon / \|f\|_{\mathrm{Lip}}$, then, using Jensen's inequality, Lipschitz continuity of f and nonexpansivity of T_i, we get

$$\begin{aligned} \left| \delta_x \mathcal{P}^k f - \delta_y \mathcal{P}^k f \right| &= \left| \mathbb{E}[f(X_k^x)] - \mathbb{E}[f(X_k^y)] \right| \\ &\leq \mathbb{E}[|f(X_k^x) - f(X_k^y)|] \\ &\leq \|f\|_{\mathrm{Lip}} \mathbb{E}[d(X_k^x, X_k^y)] \\ &\leq \|f\|_{\mathrm{Lip}} \mathbb{E}[d(x,y)] < \epsilon \end{aligned}$$

for all $k \in \mathbb{N}$. $\square$

A very helpful fact used later on is that the distance between the supports of two ergodic measures is attained; moreover, any point in the support of the one ergodic measure has a nearest neighbor in the support of the other ergodic measure.

Lemma 3.18 (distance of supports is attained)**.** *Let G be a Polish space and $T_i : G \to G$ be nonexpansive, $i \in I$. Suppose $\pi, \tilde{\pi}$ are ergodic probability measures for $\mathcal{P}$. Denote the support of a measure π by $S_\pi := \operatorname{supp} \pi$. Then for all $s \in S_\pi$ there exists $\tilde{s} \in S_{\tilde{\pi}}$ with $d(s,\tilde{s}) = \operatorname{dist}(s, S_{\tilde{\pi}}) = \operatorname{dist}(S_\pi, S_{\tilde{\pi}})$.*

Proof. First we show, that $\operatorname{dist}(S_\pi, S_{\tilde{\pi}}) = \operatorname{dist}(s, S_{\tilde{\pi}})$ for all $s \in S_\pi$. Therefore, recall the notation $X_k^x = T_{\xi_{k-1}} \cdots T_{\xi_0} x$ and note that by nonexpansivity of T_i, $i \in I$ and Lemma 3.8 it holds a.s. that

$$\operatorname{dist}(X_{k+1}^x, S_\pi) \leq \operatorname{dist}(X_{k+1}^x, T_{\xi_k} S_\pi) = \inf_{s \in S_\pi} d(T_{\xi_k} X_k^x, T_{\xi_k} s) \leq \operatorname{dist}(X_k^x, S_\pi)$$

for all $x \in G$, $\pi \in \operatorname{inv} \mathcal{P}$ and $k \in \mathbb{N}$. Suppose now there would exist an $\hat{s} \in S_\pi$ with $\operatorname{dist}(\hat{s}, S_{\tilde{\pi}}) < \operatorname{dist}(s, S_{\tilde{\pi}})$. Then by Lemma 3.16 for all $\epsilon > 0$ there is a $k \in \mathbb{N}$ with $\mathbb{P}(X_k^{\hat{s}} \in \mathbb{B}(s,\epsilon)) > 0$ and hence

$$\operatorname{dist}(s, S_{\tilde{\pi}}) \leq d(s, X_k^{\hat{s}}) + \operatorname{dist}(X_k^{\hat{s}}, S_{\tilde{\pi}}) \leq \epsilon + \operatorname{dist}(\hat{s}, S_{\tilde{\pi}})$$

with positive probability for all $\epsilon > 0$, which is a contradiction. So, it holds that $\operatorname{dist}(\hat{s}, S_{\tilde{\pi}}) = \operatorname{dist}(s, S_{\tilde{\pi}})$ for all $s, \hat{s} \in S_\pi$.

For $s \in S_\pi$ let $(\tilde{s}_m) \subset S_{\tilde{\pi}}$ be a minimizing sequence for $\operatorname{dist}(s, S_{\tilde{\pi}})$, i.e. $\lim_m d(s, \tilde{s}_m) = \operatorname{dist}(s, S_{\tilde{\pi}})$. Now define a probability measure γ_k^m on $G \times G$ via

$$\gamma_k^m f := \mathbb{E}\left[\frac{1}{k} \sum_{j=1}^{k} f(X_j^s, X_j^{\tilde{s}_m}) \right]$$

for measurable $f : G \times G \to \mathbb{R}$. Then $\gamma_k^m \in C(\nu_k^s, \nu_k^{\tilde{s}_m})$ where $C(\nu_k^s, \nu_k^{\tilde{s}_m})$ is the set of all couplings for ν_k^s and $\nu_k^{\tilde{s}_m}$ (see (44)). Also, by Lemma A.5 and Remark 3.14 the sequence $(\gamma_k^m)_{k\in\mathbb{N}}$ is tight for fixed $m \in \mathbb{N}$ and there exists a cluster point $\gamma^m \in C(\pi, \tilde{\pi})$. The sequence $(\gamma^m) \subset C(\pi, \tilde{\pi})$ is again tight by Lemma A.5. Thus for any cluster point $\gamma \in C(\pi, \tilde{\pi})$ and the bounded and continuous function $(x, y) \mapsto f^M(x, y) = \min(M, d(x, y))$ this yields

$$\gamma_k^m d = \gamma_k^m f^M \searrow \gamma^m f^M \qquad \text{as } k \to \infty$$

for all $M \geq d(s, \tilde{s}_m)$, $m \in \mathbb{N}$. Since by the Monotone Convergence Theorem $\gamma^m f^M \nearrow \gamma^m d$ as $m \to \infty$, it follows that $\gamma^m f^M = \gamma^m d$ for all $M \geq d(s, \tilde{s}_m)$. The same argument holds for $M \geq d(s, \tilde{s}_1)$ and a subsequence (γ^{m_j}) with limit γ such that $\gamma d = \gamma f^M$. Hence,

$$\gamma d = \gamma f^M = \lim_j \gamma^{m_j} f^M = \lim_j \gamma^{m_j} d \leq \lim_j d(s, \tilde{s}_{m_j}) = \operatorname{dist}(s, S_{\tilde{\pi}}).$$

In particular for γ-a.e. $(x, y) \in S_\pi \times S_{\tilde{\pi}}$ it holds that $d(x, y) = \operatorname{dist}(S_\pi, S_{\tilde{\pi}})$, because $d(x, y) \geq \operatorname{dist}(S_\pi, S_{\tilde{\pi}})$ on $S_\pi \times S_{\tilde{\pi}}$. Taking the closure of these (x, y) in $G \times G$, we see that for any $s \in S_\pi$ there is $\tilde{s} \in S_{\tilde{\pi}}$ with $d(s, \tilde{s}) = \operatorname{dist}(S_\pi, S_{\tilde{\pi}})$ by Lemma A.4. □

3.5. Convergence for nonexpansive mappings in $\mathcal{E}$

By Proposition 3.12 tightness of a sequence of Cesàro averages is equivalent to convergence of said sequence. So our focus is on tightness in the Euclidean space setting.

Lemma 3.19 (tightness of $(\mu\mathcal{P}^k)$ in $\mathcal{E}$)**.** *Let $T_i : \mathcal{E} \to \mathcal{E}$ be nonexpansive for all $i \in I$, and let $\operatorname{inv}\mathcal{P} \neq \emptyset$ for the corresponding Markov operator. The sequence $(\mu\mathcal{P}^k)_{k\in\mathbb{N}}$ is tight for any $\mu \in \mathscr{P}(\mathcal{E})$.*

Proof. First, let $\mu = \delta_x$ for $x \in \mathcal{E}$. We know that the sequence $(\delta_s\mathcal{P}^k)$ is tight for $s \in S$ by Proposition 3.10. So for $\epsilon > 0$ there is a compact $K \subset \mathcal{E}$ with $p^k(s, K) \geq 1 - \epsilon$ for all $k \in \mathbb{N}$. Recall the definition of X_k^x in (3). Since a.s. $\|X_k^x - X_k^s\| \leq \|x - s\|$, we have that $p^k(x, \overline{\mathbb{B}}(K, \|x - s\|)) = \mathbb{P}(X_k^x \in \overline{\mathbb{B}}(K, \|x - s\|)) \geq p^k(s, K) \geq 1 - \epsilon$ for all $k \in \mathbb{N}$. Hence $(\delta_x\mathcal{P}^k)$ is tight.

Now consider the initial random variable $X_0 \sim \mu$ for any $\mu \in \mathscr{P}(\mathcal{E})$. For given $\epsilon > 0$ there is a compact $K_\epsilon^\mu \subset \mathcal{E}$ with $\mu(K_\epsilon^\mu) > 1 - \epsilon$. From the special case established above, there exists a compact $K_\epsilon \subset \mathcal{E}$ with $p^k(0, K_\epsilon) > 1 - \epsilon$ for all $k \in \mathbb{N}$. Let $M > 0$ such that $K_\epsilon^\mu \subset \overline{\mathbb{B}}(0, M)$ and let $x \in \overline{\mathbb{B}}(0, M)$. We have that $p^k(x, \overline{\mathbb{B}}(K_\epsilon, M)) > 1 - \epsilon$ for all $x \in \overline{\mathbb{B}}(0, M)$, since $\left\|X_k^x - X_k^{X_0}\right\| \leq \|x\| \leq M$. Hence $\mu\mathcal{P}^k(\overline{\mathbb{B}}(K_\epsilon, M)) > (1 - \epsilon)^2$, which implies tightness of the sequence $(\mu\mathcal{P}^k)$. □

Remark 3.20 (tightness of (ν_k^μ) in $\mathcal{E}$)**:** The tightness of the sequence (ν_k^μ) for any $\mu \in \mathscr{P}(\mathcal{E})$ follows immediately from tightness of $(\mu\mathcal{P}^k)$ as in Remark 3.11.

We are now in a position to prove the first main result.

Proof of Theorem 2.8. By Lemma 3.17 the Markov operator $\mathcal{P}$ is Feller and equicontinuous. By Lemma 3.19 the sequence $\left(\mu\mathcal{P}^k\right)$ is tight, and so the sequence of of Cesáro averages (ν_k^μ) is also tight (see Remark 3.20). Hence by Proposition 3.12 $\nu_k^\mu \to \pi^\mu$ with π^μ given by (13). □

3.6. More properties of the RFI for nonexpansive mappings

This section is devoted to the preparation of some tools used in Section 3.7 to prove convergence of the distributions of the iterates of the RFI. When the Markov chain is initialized with a point

not supported in S, i.e. when $\operatorname{supp}\mu \setminus S \neq \emptyset$, the convergence results on general Polish spaces are much weaker than for the ergodic case in the previous section. One problem is that the sequences $(\nu_k^x)_{k\in\mathbb{N}}$ for $x \in G \setminus S$ need not be tight anymore. The right-shift operator $\mathcal{R}$ on l^2, for example, with the initial distribution δ_{e_1}, generates the sequence $\mathcal{R}^k e_1 = e_k$, $k = 1, 2, \ldots$. Examples of spaces on which we can always guarantee tightness are, of course, Euclidean spaces as seen in the previous section, and compact metric spaces – since then $(\mathscr{P}(G), d_P)$ is compact.

For the case that the sequence of Cesàro averages does not necessarily converge, we have the following result.

Lemma 3.21 (convergence of with nonexpansive mappings)**.** *Let (G, d) be a separable complete metric space and let $T_i : G \to G$ be nonexpansive for all $i \in I$. Suppose* $\operatorname{inv}\mathcal{P} \neq \emptyset$. *Let $X_0 \sim \mu \in \mathscr{P}(G)$ and let (X_k) be the sequence generated by Algorithm 1. Denote the support of any measure μ by S_μ, and denote $\nu_k := \frac{1}{k}\sum_{j=1}^k \mu\mathcal{P}^j$.*

(i)
$$\forall \pi \in \operatorname{inv}\mathcal{P}, \quad \operatorname{dist}(X_{k+1}, S_\pi) \leq \operatorname{dist}(X_k, S_\pi) \text{ a.s.} \quad \forall k \in \mathbb{N}.$$

(ii) If the sequence (ν_k) has a cluster point $\pi \in \operatorname{inv}\mathcal{P}$, then,

(a) $\operatorname{dist}(X_k, S_\pi) \to 0$ a.s. as $k \to \infty$;

(b) all cluster points of the sequence (ν_k) have the same support;

(c) cluster points of the sequence $(\mu\mathcal{P}^k)$ have support in S_π (if they exist).

Proof. (i). By Lemma 3.8, the sets on which $T_{\xi_k} S_\pi$ is not a subset of S_π are $\mathbb{P}$-null sets and their union is also a $\mathbb{P}$-null set. This yields

$$(\forall s \in S_\pi) \quad \operatorname{dist}(X_{k+1}, S_\pi) \leq d(X_{k+1}, T_{\xi_k} s) = d(T_{\xi_k} X_k, T_{\xi_k} s) \leq d(X_k, s) \quad \text{a.s.},$$

and hence

$$\operatorname{dist}(X_{k+1}, S_\pi) \leq \operatorname{dist}(X_k, S_\pi) \quad \text{a.s.}$$

(iia) Define the function $f = \min(M, \operatorname{dist}(\cdot, S_\pi))$ for some $M > 0$. Since this is bounded and continuous, we have for a subsequence (ν_{k_j}) converging to π, that $\nu_{k_j} f = \frac{1}{k_j}\sum_{n=1}^{k_j} \mu\mathcal{P}^n f \to \pi f = 0$ as $j \to \infty$. Now part (i) and the identity

$$\mu\mathcal{P}^{n+1} f = \mathbb{E}[\min(M, \operatorname{dist}(X_{n+1}, S_\pi))] \leq \mathbb{E}[\min(M, \operatorname{dist}(X_n, S_\pi))] = \mu\mathcal{P}^n f$$

yield $\mu\mathcal{P}^n f = \mathbb{E}[\min(M, \operatorname{dist}(X_n, S_\pi))] \to 0$ as $n \to \infty$. Again by part (i)

$$Y := \lim_{n\to\infty} \min(M, \operatorname{dist}(X_n, S_\pi))$$

exists and is nonnegative; so by Lebesgue's dominated convergence theorem it follows that $Y = 0$ a.s., since otherwise $\mathbb{E}[Y] > 0 = \lim_{n\to\infty} \mu\mathcal{P}^n f$ would yield a contradiction.

(iib) Let π_1, π_2 be two cluster points of (ν_k) with support S_1, S_2 respectively, then these probability measures are invariant for $\mathcal{P}$ by Proposition 3.1. By Corollary 3.15 the intersection $S_1 \cap S_2$ must be nonempty. Suppose now w.l.o.g. $\exists y \in S_1 \setminus S_2$. Then there is an $\epsilon > 0$ with $\mathbb{B}(y, 2\epsilon) \cap S_2 = \emptyset$. Let $f : G \to [0, 1]$ be a continuous function that takes the value 1 on $\mathbb{B}(y, \frac{\epsilon}{2})$ and 0 outside of $\mathbb{B}(y, \epsilon)$. Then $\pi_1 f > 0$ and $\pi_2 f = 0$. But there are two subsequences of (ν_k) with $\nu_{k_j} f \to \pi_1 f$ and $\nu_{\tilde{k}_j} f \to \pi_2 f$ as $j \to \infty$. For the former sequence we have, for j large enough,

$$\exists \delta > 0 : \frac{1}{k_j}\sum_{n=1}^{k_j} \mu\mathcal{P}^n f \geq \delta > 0.$$

So, one can from this extract a sequence $(m_k)_{k\in\mathbb{N}} \subset \mathbb{N}$ with $\mu\mathcal{P}^{m_k} f \geq \delta$, $k \in \mathbb{N}$. Note that $\mathbb{P}(X_{m_k} \in \mathbb{B}(y,\epsilon)) \geq \mu\mathcal{P}^{m_k} f \geq \delta > 0$. This implies $\operatorname{dist}(X_{m_k}, S_2) \geq \epsilon$ with $\mathbb{P} \geq \delta$ and hence $\mathbb{E}[\operatorname{dist}(X_{m_k}, S_2)] \geq \delta\epsilon$, in contradiction to (iia). So there cannot be such y which yields $S_1 = S_2$, as claimed.

(iic) Let ν be a cluster point of the sequence $(\mu\mathcal{P}^k)$, which is assumed to exist, and assume there is $s \in \operatorname{supp}\nu \setminus S_\pi$ and $\epsilon > 0$ such that $\operatorname{dist}(s, S_\pi) > 2\epsilon$. Let $f : G \to [0,1]$ be a continuous function, that takes the value 1 on $\mathbb{B}(s, \frac{\epsilon}{2})$ and 0 outside of $\mathbb{B}(s,\epsilon)$. With (iia) we find, that

$$0 < \nu f = \lim_j \mathbb{P}^{X_{k_j}} f \leq \lim_j \mathbb{P}(X_{k_j} \in \mathbb{B}(s,\epsilon)) = 0.$$

Were $\mathbb{P}(X_{k_j} \in \mathbb{B}(s,\epsilon)) \geq \delta > 0$ for j large enough, then this would imply that

$$\mathbb{E}[\operatorname{dist}(X_{k_j}, S_\pi)] \geq \delta\epsilon,$$

which is a contradiction. We conclude that there is no such s, which completes the proof. □

We now prepare some tools to handle convergence of the distributions of the iterates of the RFI for averaged mappings in Section 3.7. We restrict ourselves to Polish spaces with *finite dimensional metric* (see Definition 3.24) in order to apply a differentiation theorem. We begin with the next technical fact.

Lemma 3.22 (characterization of balls in $(\mathscr{E}, d_P)$)**.** *Let (G,d) be a separable complete space and $T_i : G \to G$ be nonexpansive for all $i \in I$. Let $\mathscr{E}$ denote the (convex) set of ergodic measures associated to the Markov operator $\mathcal{P}$, which is induced by the family of mappings $\{T_i\}_{i\in I}$ and the marginal probability law of the random variables ξ_k. Let $\pi, \tilde{\pi} \in \mathscr{E}$ and denote the support of the measure π by S_π (and similarly for $\tilde{\pi}$). Then*

$$\tilde{\pi} \in \overline{\mathbb{B}}(\pi,\epsilon) \qquad \Longleftrightarrow \qquad S_{\tilde{\pi}} \subset \overline{\mathbb{B}}(S_\pi,\epsilon)$$

for $\epsilon \in (0,1)$, where $\overline{\mathbb{B}}(\pi,\epsilon)$ is the closed ϵ-ball with respect to the Prokhorov-Lèvy metric d_P.

Proof. By Lemma 3.18 there exist $s \in S_\pi$ and $\tilde{s} \in S_{\tilde{\pi}}$ such that $d(s,\tilde{s}) = \operatorname{dist}(S_\pi, S_{\tilde{\pi}})$. First note that, if $\pi \neq \tilde{\pi}$, then $S_\pi \cap S_{\tilde{\pi}} = \emptyset$ by Corollary 3.15, and hence $d(s,\tilde{s}) = \operatorname{dist}(S_\pi, S_{\tilde{\pi}}) > 0$. Recall the notation $X_k^x := T_{\xi_{k-1}} \cdots T_{\xi_0} x$ for $x \in G$ and note that by Lemma A.4(i) and Lemma 3.8, $\operatorname{supp}\mathcal{L}(X_k^s) \subset S_\pi$ and $\operatorname{supp}\mathcal{L}(X_k^{\tilde{s}}) \subset S_{\tilde{\pi}}$. So it holds that $d(X_k^s, X_k^{\tilde{s}}) \geq \operatorname{dist}(S_\pi, S_{\tilde{\pi}})$ a.s. for all $k \in \mathbb{N}$. Since T_i $(i \in I)$ is nonexpansive we have that $d(X_k^s, X_k^{\tilde{s}}) \leq d(s,\tilde{s})$ a.s. for all $k \in \mathbb{N}$. So, both inequalities together imply the equality

$$d(X_k^s, X_k^{\tilde{s}}) = d(s,\tilde{s}) \qquad \text{a.s. } \forall k \in \mathbb{N}. \tag{16}$$

Now, letting $c := \min(1, d(s,\tilde{s}))$, we show that $d_P(\pi,\tilde{\pi}) = c$, where d_P denotes the Prokhorov-Lèvy metric (see Lemma A.6). Indeed, take $(X,Y) \in C(\mathcal{L}(X_k^s), \mathcal{L}(X_k^{\tilde{s}}))$. Again, by Lemma A.4(i) and Lemma 3.8 $\operatorname{supp}\mathcal{L}(X) \subset S_\pi$ and $\operatorname{supp}\mathcal{L}(Y) \subset S_{\tilde{\pi}}$ and hence $d(X,Y) \geq \operatorname{dist}(S_\pi, S_{\tilde{\pi}}) = d(s,\tilde{s})$ a.s. We have, thus

$$\mathbb{P}(d(X,Y) > c - \delta) \geq \mathbb{P}(d(X,Y) > d(s,\tilde{s}) - \delta) = 1 \qquad \forall \delta > 0,$$

which implies $d_P(\mathcal{L}(X_k^s), \mathcal{L}(X_k^{\tilde{s}})) \geq c$ by Lemma A.6(i). In particular, for $c = 1$ it follows that $d_P(\mathcal{L}(X_k^s), \mathcal{L}(X_k^{\tilde{s}})) = 1$, since d_P is bounded by 1. Now, let $c < 1$, i.e. $c = d(s,\tilde{s}) < 1$. We have by (16)

$$\inf_{(X,Y)\in C(\mathcal{L}(X_k^s),\mathcal{L}(X_k^{\tilde{s}}))} \mathbb{P}\left(d(X,Y) > c\right) \leq \mathbb{P}\left(d(X_k^s, X_k^{\tilde{s}}) > c\right) = 0 \leq c.$$

Altogether we find that $d_P(\mathcal{L}(X_k^s), \mathcal{L}(X_k^{\tilde{s}})) = c$, again by Lemma A.6(i). Since also $\operatorname{supp} \nu_k^s \subset S_\pi$ and $\operatorname{supp} \nu_k^{\tilde{s}} \subset S_{\tilde{\pi}}$, where $\nu_k^x = \frac{1}{k}\sum_{j=1}^k \mathcal{L}(X_j^x)$ for any $x \in G$, it follows that

$$c \le d_P(\nu_k^s, \nu_k^{\tilde{s}}) \le \max_{j=1,\dots,k} d_P(\mathcal{L}(X_j^s), \mathcal{L}(X_j^{\tilde{s}})) = c \tag{17}$$

by Lemma A.6(v). Now taking the limit $k \to \infty$ of (17) and using Remark 3.14, it follows that $d_P(\pi, \tilde{\pi}) = c$. This proves the assertion. □

Definition 3.23 (Besicovitch family)**.** A family $\mathcal{B}$ of closed balls $B = \overline{\mathbb{B}}(x_B, \epsilon_B)$ with $x_B \in G$ and $\epsilon_B > 0$ on the metric space (G, d) is called a *Besicovitch family* of balls if

(i) for every $B \in \mathcal{B}$ one has $x_B \notin B' \in \mathcal{B}$ for all $B' \neq B$, and

(ii) $\bigcap_{B\in\mathcal{B}} B \neq \emptyset$.

Definition 3.24 (σ-finite dimensional metric)**.** Let (G, d) be a metric space. We say that d is *finite dimensional* on a subset $D \subset G$ if there exist constants $K \ge 1$ and $0 < r \le \infty$ such that $\operatorname{Card} \mathcal{B} \le K$ for every Besicovitch family $\mathcal{B}$ of balls in (G, d) centered on D with radius $< r$. We say that d is *σ-finite dimensional* if G can be written as a countable union of subsets on which d is finite dimensional.

Proposition 3.25 (differentiation theorem, [48])**.** *Let (G, d) be a separable complete metric space. For every locally finite Borel regular measure λ over (G, d), it holds that*

$$\lim_{r\to 0} \frac{1}{\lambda(\overline{\mathbb{B}}(x, r))} \int_{\overline{\mathbb{B}}(x,r)} f(y)\lambda(\mathrm{d}y) = f(x) \qquad \text{for } \lambda\text{-a.e. } x \in G,\ \forall f \in L^1_{\mathrm{loc}}(G, \lambda) \tag{18}$$

if and only if d is σ-finite dimensional.

Proposition 3.26 (Besicovitch covering property in $\mathscr{E}$)**.** *Let (G, d) be separable complete metric space with finite dimensional metric d and let $T_i : G \to G$ be nonexpansive, $i \in I$. The cardinality of any Besicovitch family of balls in $(\mathscr{E}, d_P)$ is bounded by the same constant that bounds the cardinality of Besicovitch families in G.*

Proof. Let $\mathcal{B}$ be a Besicovitch family of closed balls $B = \overline{\mathbb{B}}(\pi_B, \epsilon_B)$ in $(\mathscr{E}, d_P)$, where $\pi_B \in \mathscr{E}$ and $\epsilon_B > 0$. Note that if $\epsilon_B \ge 1$, then $|\mathcal{B}| = 1$, since in that case $B = \mathscr{E}$ since d_P is bounded by 1. So let $|\mathcal{B}| > 1$, that implies $\epsilon_B < 1$ for all $B \in \mathcal{B}$.

The defining properties of a Besicovitch family translate then with help of Lemma 3.22 into

$$\pi_B \notin B', \quad \forall B' \in \mathcal{B} \setminus \{B\} \qquad \Longleftrightarrow \qquad S_{\pi_B} \cap \overline{\mathbb{B}}(S_{\pi_{B'}}, \epsilon_{B'}) = \emptyset, \qquad \forall B' \in \mathcal{B} \setminus \{B\}, \tag{19}$$

and

$$\bigcap_{B\in\mathcal{B}} B \neq \emptyset \qquad \Longleftrightarrow \qquad \bigcap_{B\in\mathcal{B}} \overline{\mathbb{B}}(S_{\pi_B}, \epsilon_B) \neq \emptyset. \tag{20}$$

Now fix π in the latter intersection in (20) and let $s \in S_\pi$. Also fix for each $B \in \mathcal{B}$ a point $s_B \in S_{\pi_B}$ with the property that $s_B \in \operatorname{argmin}_{\tilde{s}\in S_{\pi_B}} d(s, \tilde{s})$ (possible by Lemma 3.18). Then the family $\mathcal{C}$ of balls $\overline{\mathbb{B}}(s_B, \epsilon_B) \subset G$, $B \in \mathcal{B}$ is also a Besicovitch family: We have $s_B \notin B'$ for $B \neq B'$ due to (19) and by the choice of s_B one has $s \in \bigcap_{B\in\mathcal{C}} B$. Since the cardinality of any Besicovitch family in G is bounded by a uniform constant, it follows, that also the cardinality of $\mathcal{B}$ is uniformly bounded. □

Remark 3.27: The cardinality of any Besicovitch family in $\mathcal{E}$ is uniformly bounded depending on $\dim \mathcal{E}$ [42, Lemma 2.6].

Lemma 3.28 (equality around the support of ergodic measures implies equality of measures)**.** *Let (G, d) be a separable complete metric space with the finite dimensional metric d and let $T_i : G \to G$ be nonexpansive ($i \in I$). If $\pi_1, \pi_2 \in \operatorname{inv} \mathcal{P}$ satisfy*

$$\pi_1(\overline{\mathbb{B}}(S_\pi, \epsilon)) = \pi_2(\overline{\mathbb{B}}(S_\pi, \epsilon)) \tag{21}$$

for all $\epsilon > 0$ and all $\pi \in \mathscr{E}$, then $\pi_1 = \pi_2$.

Proof. From Proposition 3.5 follows the existence of probability measures q_1, q_2 on the set $\mathscr{E}$ of ergodic measures for $\mathcal{P}$ such that one has

$$\pi_j(A) = \int_{\mathscr{E}} \pi(A) q_j(\mathrm{d}\pi), \qquad A \in \mathcal{B}(G),\ j = 1, 2.$$

If we set $q = \frac{1}{2}(q_1 + q_2)$, then by the Radon-Nikodym theorem, there are densities $f_1, f_2 \geq 0$ on $\mathscr{E}$ with $q_j = f_j \cdot q$ and hence

$$\pi_j(A) = \int_{\mathscr{E}} \pi(A) f_j(\pi) q(\mathrm{d}\pi), \qquad A \in \mathcal{B}(G),\ j = 1, 2.$$

For q-measurable subsets $E \subset \mathscr{E}$, one can define a probability measure on $\mathscr{E}$ via

$$\tilde{\pi}_j(E) := \int_{\mathscr{E}} \mathbb{1}_E(\pi) f_j(\pi) q(\mathrm{d}\pi), \qquad j = 1, 2. \tag{22}$$

One then has for $\epsilon > 0$ and $\pi \in \mathscr{E}$ that

$$\pi_j(\overline{\mathbb{B}}(S_\pi, \epsilon)) = \tilde{\pi}_j(\overline{\mathbb{B}}(\pi, \epsilon)), \qquad j = 1, 2, \tag{23}$$

where $\overline{\mathbb{B}}(\pi, \epsilon) := \{\tilde{\pi} \in \mathscr{E} \mid d_P(\tilde{\pi}, \pi) \leq \epsilon\}$. This is due to Lemma 3.22, from which follows

$$\tilde{\pi}(\overline{\mathbb{B}}(S_\pi, \epsilon)) = \begin{cases} 1, & \tilde{\pi} \in \overline{\mathbb{B}}(\pi, \epsilon) \\ 0, & \text{else} \end{cases}.$$

With the above characterizations of π_j and $\tilde{\pi}_j$, we can use Proposition 3.25 to show that $f_1 = f_2$ q-a.s., which, together with (22), would imply that $\pi_1 = \pi_2$, as claimed. To apply Proposition 3.25 we require that d_P is finite dimensional. But this follows from Proposition 3.26. So Proposition 3.25 applied to $\tilde{\pi}_j$ with respect to q then gives q-a.s.

$$\lim_{\epsilon \to 0} \frac{\tilde{\pi}_j(\overline{\mathbb{B}}(\pi, \epsilon))}{q(\overline{\mathbb{B}}(\pi, \epsilon))} = f_j(\pi), \quad j = 1, 2. \tag{24}$$

And since $\tilde{\pi}_1(\overline{\mathbb{B}}(\pi, \epsilon)) = \tilde{\pi}_2(\overline{\mathbb{B}}(\pi, \epsilon))$ by (23) and assumption (21), we have $f_1 = f_2$ q-a.s., which completes the proof. □

Remark 3.29: In the assertion of Lemma 3.28, it is enough to claim the existence of a sequence $(\epsilon_k^\pi)_{k \in \mathbb{N}} \subset \mathbb{R}_+$ with $\epsilon_k^\pi \to 0$ as $k \to \infty$ satisfying

$$\pi_1(\overline{\mathbb{B}}(S_\pi, \epsilon_k^\pi)) = \pi_2(\overline{\mathbb{B}}(S_\pi, \epsilon_k^\pi)) \qquad \forall \pi \in \mathscr{E},\ \forall k \in \mathbb{N},$$

because from Proposition 3.25 one has the existence of the limit in (24) q-a.s.

3.7. Convergence theory for averaged mappings

Continuing the development of the convergence theory under greater regularity assumptions on the mappings T_i ($i \in I$), in this section we examine what is achievable under the assumption that the mappings T_i are averaged (Definition 2.7). We restrict ourselves to the Euclidean space $\mathcal{E}$, and begin with a technical lemma that describes properties of sequences whose relative expected distances are invariant under T_ξ.

Lemma 3.30 (constant expected separation)**.** *Let $T_i : \mathcal{E} \to \mathcal{E}$ be averaged with $\alpha_i \leq \alpha < 1$, $i \in I$. Let $\mu, \nu \in \mathscr{P}(\mathcal{E})$ and $X \sim \mu$, $Y \sim \nu$ independent of (ξ_k) satisfy*

$$\mathbb{E}\left[\left\|X_k^X - X_k^Y\right\|^2\right] = \mathbb{E}\left[\|X - Y\|^2\right] \qquad \forall k \in \mathbb{N},$$

where $X_k^x := T_{\xi_{k-1}} \cdots T_{\xi_0} x$ for $x \in \mathcal{E}$ is the RFI sequence started at x. Then for $\mathbb{P}^{(X,Y)}$-a.e. $(x, y) \in \mathcal{E} \times \mathcal{E}$ we have $X_k^x - X_k^y = x - y$ $\mathbb{P}$-a.s. for all $k \in \mathbb{N}$. Moreover, if there exists an invariant measure for $\mathcal{P}$, then

$$\pi^x(\cdot) = \pi^y(\cdot - (x - y)) \qquad \mathbb{P}^{(X,Y)}\text{-a.s.}$$

for the limiting invariant measures π^x of the Cesàro average of $(\delta_x \mathcal{P}^k)$ and π^y of the Cesàro average of $(\delta_y \mathcal{P}^k)$.

Proof. By the characterization of averaged mappings (11), one has

$$\begin{aligned}
\mathbb{E}\left[\|X - Y\|^2\right] &\geq \mathbb{E}\left[\|T_{\xi_0}X - T_{\xi_0}Y\|^2\right] + \tfrac{1-\alpha}{\alpha}\mathbb{E}\left[\|(X - T_{\xi_0}X) - (Y - T_{\xi_0}Y)\|^2\right] \\
&\geq \mathbb{E}\left[\|T_{\xi_1}T_{\xi_0}X - T_{\xi_1}T_{\xi_0}Y\|^2\right] \\
&\quad + \tfrac{1-\alpha}{\alpha}\left(\mathbb{E}\left[\|(T_{\xi_0}X - T_{\xi_1}T_{\xi_0}X) - (T_{\xi_0}Y - T_{\xi_1}T_{\xi_0}Y)\|^2\right] + \mathbb{E}\left[\|(X - T_{\xi_0}X) - (Y - T_{\xi_0}Y)\|^2\right]\right) \\
&\geq \cdots \\
&\geq \mathbb{E}\left[\left\|T_{\xi_{k-1}} \cdots T_{\xi_0}X - T_{\xi_{k-1}} \cdots T_{\xi_0}Y\right\|^2\right] \\
&\quad + \tfrac{1-\alpha}{\alpha}\sum_{j=0}^{k-1}\mathbb{E}\left[\left\|(T_{\xi_{j-1}} \cdots T_{\xi_{-1}}X - T_{\xi_j} \cdots T_{\xi_0}X) - (T_{\xi_{j-1}} \cdots T_{\xi_{-1}}Y - T_{\xi_j} \cdots T_{\xi_0}Y)\right\|^2\right],
\end{aligned}$$

where we used $T_{\xi_{-1}} := \mathrm{Id}$ for a simpler representation of the sum. The assumption $\mathbb{E}\left[\left\|X_k^X - X_k^Y\right\|^2\right] = \mathbb{E}\left[\|X - Y\|^2\right]$ for all $k \in \mathbb{N}$ then implies, that for $j = 1, \ldots, k$ $\mathbb{P}$-a.s.

$$X_k^X - X_{k-1}^X = X_k^Y - X_{k-1}^Y \quad (k \in \mathbb{N}),$$

and hence by induction

$$X_k^X - X_k^Y = X - Y.$$

By disintegrating and using $(X, Y) \perp\!\!\!\perp (\xi_k)$ we have $\mathbb{P}$-a.s.

$$\begin{aligned}
0 &= \mathbb{E}\left[\left\|(X - X_k^X) - (Y - X_k^Y)\right\|^2 \,\middle|\, X, Y\right] \\
&= \int_{I^{k+1}} \|(X - T_{i_k} \cdots T_{i_0}X) - (Y - T_{i_k} \cdots T_{i_0}Y)\|^2 \mathbb{P}^\xi(\mathrm{d}i_k) \cdots \mathbb{P}^\xi(\mathrm{d}i_0).
\end{aligned}$$

Consequently, for $\mathbb{P}^{(X,Y)}$-a.e. $(x,y) \in \mathcal{E} \times \mathcal{E}$, we have

$$X_k^x - X_k^y = x - y \quad \forall k \in \mathbb{N} \quad \mathbb{P} - \text{a.s.}$$

So in particular for any $A \in \mathcal{B}(\mathcal{E})$

$$p^k(x, A) = \mathbb{P}(X_k^x \in A) = \mathbb{P}(X_k^y \in A - (x-y)) = p^k(y, A - (x-y))$$

and hence, denoting $f_h = f(\cdot + h)$ and $\nu_k^x = \frac{1}{k}\sum_{j=1}^k p^j(x, \cdot)$, one also has for $f \in C_b(\mathcal{E})$ by Theorem 2.8

$$\nu_k^y f_{x-y} \to \pi^y f_{x-y} = \pi_{x-y}^y f \quad \textit{and} \quad \nu_k^x f \to \pi^x f \text{ as } k \to \infty,$$

where $\pi_{x-y}^y := \pi^y(\cdot - (x-y))$. So from $\nu_k^y f_{x-y} = \nu_k^x f$ for any $f \in C_b(\mathcal{E})$ and $k \in \mathbb{N}$ it follows that $\pi_{x-y}^y = \pi^x$. □

We can now give the proof of the second main result. For a given $h : \mathcal{E} \times \mathcal{E} \to \mathbb{R}$ we will define sequences of functions $(\overline{h}_k)$ on $\mathcal{E} \times \mathcal{E}$ via

$$\overline{h}_k(x,y) := \mathbb{E}\left[h(X_k^x, X_k^y)\right], \quad X_k^z := T_{\xi_{k-1}} \cdots T_{\xi_0} z \text{ for any } z \in \mathcal{E} \quad (k \in \mathbb{N}).$$

Note that, by continuity of T_i, $i \in I$ and Lebesgue's dominated convergence theorem, $\overline{h}_k \in C_b(\mathcal{E} \times \mathcal{E})$ for all $k \in \mathbb{N}$ whenever $h \in C_b(\mathcal{E} \times \mathcal{E})$.

Proof of Theorem 2.9. Let $x, y \in \mathcal{E}$, define $F(x,y) := \|x-y\|^2$ and the corresponding sequence of functions

$$\overline{F}_k(x,y) := \mathbb{E}\left[F(X_k^x, X_k^y)\right], \quad X_k^z := T_{\xi_{k-1}} \cdots T_{\xi_0} z \text{ for any } z \in \mathcal{E} \quad (k \in \mathbb{N}).$$

By the remarks preceding this proof, $\overline{F}_k \in C_b(\mathcal{E} \times \mathbb{R}^n)$ for all $k \in \mathbb{N}$. From the regularity of T_i, $i \in I$ and the characterization (11), we get that a.s. for all $k \in \mathbb{N}$

$$\|X_k^x - X_k^y\|^2 \geq \left\|X_{k+1}^x - X_{k+1}^y\right\|^2 + \tfrac{1-\alpha}{\alpha}\left\|(X_k^x - X_{k+1}^x) - (X_k^y - X_{k+1}^y)\right\|^2. \tag{25}$$

After computing the expectation, this is the same as

$$\overline{F}_k(x,y) \geq \overline{F}_{k+1}(x,y) + \tfrac{1-\alpha}{\alpha}\mathbb{E}\left[\left\|(X_k^x - X_{k+1}^x) - (X_k^y - X_{k+1}^y)\right\|^2\right].$$

We conclude that $(\overline{F}_k(x,y))$ is a monotonically nonincreasing sequence for any $x, y \in \mathcal{E}$.
Recall the notation $S_\pi := \operatorname{supp}\pi$ for some measure π. Let $s, \tilde{s} \in S_\pi$ for the ergodic invariant measure $\pi \in \mathscr{E}$ and define the sequence of measures γ_k by

$$\gamma_k f := \mathbb{E}\left[f(X_k^s, X_k^{\tilde{s}})\right]$$

for any measurable function $f : \mathcal{E} \times \mathcal{E} \to \mathbb{R}$. Note that due to nonexpansiveness the pair $(X_k^s, X_k^{\tilde{s}})$ a.s. takes values in $G_r := \{(x,y) \in \mathcal{E} \times \mathcal{E} : ||x-y||^2 \leq r\}$ for $r = ||s - \tilde{s}||^2$, so that γ_k is concentrated on this set. Since (X_k^s) is a tight sequence by Lemma 3.19, and likewise for $(X_k^{\tilde{s}})$, we know from Lemma A.5 that the sequence (γ_k) is tight as well. Let γ be a cluster point of (γ_k), which is again concentrated on $G_{||s-\tilde{s}||^2}$, and consider a subsequence (γ_{k_j}) such that $\gamma_{k_j} \to \gamma$. By Lemma A.5 we also know that $\gamma \in C(\nu_1, \nu_2)$ where ν_1 and ν_2 are the distributions of the limit in convergence in distribution of $(X_{k_j}^s)$ and $(X_{k_j}^{\tilde{s}})$. For any $f \in C_b(\mathcal{E} \times \mathbb{R}^n)$ we have $\gamma_{k_j} f \to \gamma f$. So consider the case $f = F^M$ where $F^M := \min(M, F)$ for $M \in \mathbb{R}$. Since

$\|x-y\|^2 = F(x,y) = F^M(x,y)$ almost surely (with respect to γ_{k_j} and γ) for $M \geq \|s-\tilde{s}\|^2$, we have

$$\gamma_{k_j} F = \gamma_{k_j} F^M \to \gamma F^M = \gamma F.$$

However, by the monotonicity in (25) we now also obtain convergence for the entire sequence:

$$\gamma_k F = \gamma_k F^M \searrow \gamma F^M = \gamma F.$$

Let $(X,Y) \sim \gamma$ and $(\tilde{\xi}_k) \perp\!\!\!\perp (\xi_k)$ be another i.i.d. sequence with $(X,Y) \perp\!\!\!\perp (\tilde{\xi}_k), (\xi_k)$. We use the notation $\tilde{X}_k^x := T_{\tilde{\xi}_{k-1}} \cdots T_{\tilde{\xi}_0} x$, $x \in \mathcal{E}$. Define the sequence of functions

$$\overline{F}_k^M(x,y) := \mathbb{E}\left[F^M(X_k^x, X_k^y)\right] \quad (k \in \mathbb{N}),$$

and note that $\overline{F}_k^M \in C_b(\mathcal{E} \times \mathcal{E})$. When $M \geq \|s-\tilde{s}\|^2$ this yields

$$\begin{aligned}
\gamma \overline{F}_k = \gamma \overline{F}_k^M &= \mathbb{E}\left[\min\left(M, \left\|\tilde{X}_k^X - \tilde{X}_k^Y\right\|^2\right)\right] = \lim_{j\to\infty} \gamma_{k_j} \overline{F}_k^M \\
&= \lim_{j\to\infty} \mathbb{E}\left[\min\left(M, \left\|\tilde{X}_k^{X_{k_j}^s} - \tilde{X}_k^{X_{k_j}^{\tilde{s}}}\right\|^2\right)\right] \\
&= \lim_{j\to\infty} \mathbb{E}\left[\min\left(M, \left\|X_{k+k_j}^s - X_{k+k_j}^{\tilde{s}}\right\|^2\right)\right] \\
&= \lim_{j\to\infty} \gamma_{k+k_j} F^M = \gamma F^M = \gamma F.
\end{aligned}$$

This means that for all $k \in \mathbb{N}$,

$$\mathbb{E}\left[\left\|X_k^X - X_k^Y\right\|^2\right] = \mathbb{E}[\|X-Y\|^2].$$

For $\mathbb{P}^{(X,Y)}$-a.e. (x,y) we have $x, y \in S_\pi$ and thus $\pi^x = \pi^y = \pi$ where π^x is the unique ergodic measure with $x \in S_{\pi^x}$ (see Remark 3.14). An application of Lemma 3.30 then yields $\pi(\cdot) = \pi(\cdot - (x-y))$, i.e. $x = y$. Hence $X = Y$ a.s. implying $\nu_1 = \nu_2 =: \nu$ and $\gamma F = 0$. That means

$$\gamma_k F = \mathbb{E}\left[\left\|X_k^s - X_k^{\tilde{s}}\right\|^2\right] \to 0 \qquad \text{as } k \to \infty.$$

Now Lemma A.6 yields

$$\mathbb{P}\left(\left\|X_k^s - X_k^{\tilde{s}}\right\| > \epsilon\right) \leq \frac{\mathbb{E}\left[\|X_k^s - X_k^{\tilde{s}}\|\right]}{\epsilon} \leq \frac{\mathbb{E}\left[\sqrt{\|X_k^s - X_k^{\tilde{s}}\|^2}\right]}{\epsilon} \to 0$$

as $k \to \infty$ for any $\epsilon > 0$; so this yields convergence of the corresponding probability measures $\delta_s \mathcal{P}^k$ and $\delta_{\tilde{s}} \mathcal{P}^k$ in the Prokhorov-Lèvy metric:

$$d_P(\delta_s \mathcal{P}^k, \delta_{\tilde{s}} \mathcal{P}^k) \to 0.$$

By the triangle inequality, therefore, if $\delta_s \mathcal{P}^{k_j} \to \nu$, then also $\delta_{\tilde{s}} \mathcal{P}^{k_j} \to \nu$ for any $\tilde{s} \in S_\pi$. Hence

$$d_P(\delta_{\tilde{s}} \mathcal{P}^{k_j}, \nu) \leq d_P(\delta_s \mathcal{P}^{k_j}, \delta_{\tilde{s}} \mathcal{P}^{k_j}) + d_P(\delta_s \mathcal{P}^{k_j}, \nu) \to 0, \quad \text{as } j \to \infty.$$

By Lebesgue's dominated convergence theorem we conclude that, for any $f \in C_b(\mathcal{E})$ and $\mu \in \mathscr{P}(S_\pi)$,

$$\mu\mathcal{P}^{k_j} f = \int_{S_\pi} \delta_s \mathcal{P}^{k_j} f \mu(\mathrm{d}s) \to \nu f, \quad \text{as } j \to \infty.$$

In particular, $\mu\mathcal{P}^{k_j} \to \nu$ and taking $\mu = \pi$ yields $\nu = \pi$. Thus, all cluster points of $(\delta_s \mathcal{P}^k)$ for all $s \in S_\pi$ have the same distribution π and hence, because the sequence is tight, $\delta_s \mathcal{P}^k = p^k(x, \cdot) \to \pi$.

Now, let $\mu \in \mathscr{P}(S)$, where $S = \bigcup_{\pi \in \mathscr{E}} S_\pi$. By what we have just shown we have for $x \in \operatorname{supp} \mu$, that $p^k(x, \cdot) \to \pi^x$, where π^x is unique ergodic measure with $x \in S_{\pi^x}$. Then, again by Lebesgue's dominated convergence theorem, one has for any $f \in C_b(\mathcal{E})$,

$$\mu\mathcal{P}^k f = \int f(y) p^k(x, \mathrm{d}y) \mu(\mathrm{d}x) \to \int f(y) \pi^x(\mathrm{d}y) \mu(\mathrm{d}x) =: \pi^\mu f \text{ as } k \to \infty, \tag{26}$$

and the measure π^μ is again invariant for $\mathcal{P}$ by invariance of π^x for all $x \in S$. Now, let $\mu = \delta_x$, $x \in \mathcal{E} \setminus S$. We obtain the tightness of $(\delta_x \mathcal{P}^k)$ from the tightness of $(\delta_s \mathcal{P}^k)$ for $s \in S$. Indeed, for $\epsilon > 0$ there exists a compact $K_\epsilon \subset \mathcal{E}$ with $p^k(s, K_\epsilon) > 1 - \epsilon$ for all $k \in \mathbb{N}$. This together with the fact that T_i, $i \in I$ is nonexpansive implies that $\|X_k^x - X_k^s\| \leq \|x - s\|$ for all $k \in \mathbb{N}$ hence $p^k(x, \overline{\mathbb{B}}(K_\epsilon, \|x - s\|)) > 1 - \epsilon$, where p is the transition kernel defined by (4). Tightness implies the existence of a cluster point ν of the sequence $(\delta_x \mathcal{P}^k)$. From Theorem 2.8 we know that $\nu_k^x = \frac{1}{k} \sum_{j=1}^k \delta_x \mathcal{P}^j \to \pi^x$ for some $\pi^x \in \operatorname{inv} \mathcal{P}$ with $S_{\pi^x} \subset S$. Furthermore, we have $\nu \in \mathscr{P}(S_{\pi^x}) \subset \mathscr{P}(S)$ by Lemma 3.21(iic). So by (26) there exists $\pi^\nu \in \operatorname{inv} \mathcal{P}$ with $\nu \mathcal{P}^k \to \pi^\nu$.

In order to complete the proof we have to show that $\nu = \pi^x$, i.e. π^x is the unique cluster point of $(\delta_x \mathcal{P}^k)$ and hence convergence follows by Proposition A.1. It suffices to show that $\pi^\nu = \pi^x$, since then, as $k \to \infty$

$$d_P(\nu, \pi^x) = \lim_k d_P(\delta_x \mathcal{P}^k, \pi^x) = \lim_k d_P(\delta_x \mathcal{P}^{k+j}, \pi^x) = d_P(\nu \mathcal{P}^j, \pi^x) = d_P(\nu \mathcal{P}^j, \pi^\nu) \to 0.$$

To begin, fix $\pi \in \operatorname{inv} \mathcal{P}$. For any $\epsilon > 0$ let $A_k := \{X_k^x \in \overline{\mathbb{B}}(S_\pi, \epsilon)\}$. By nonexpansivity $A_k \subset A_{k+1}$ for $k \in \mathbb{N}$, since we have by Lemma 3.8 a.s.

$$\operatorname{dist}(X_{k+1}^x, S_\pi) \leq \operatorname{dist}(X_{k+1}^x, T_{\xi_k} S_\pi) \leq \operatorname{dist}(X_k^x, S_\pi).$$

Hence $(p^k(x, \overline{\mathbb{B}}(S_\pi, \epsilon))) = (\mathbb{P}(A_k))$ is a monotonically increasing sequence and bounded from above and therefore the sequence converges to some $b_\epsilon^x \in [0, 1]$ as $k \to \infty$. It follows

$$b_\epsilon^x = \lim_k p^k(x, \overline{\mathbb{B}}(S_\pi, \epsilon)) = \lim_k \frac{1}{k} \sum_{j=1}^k p^j(x, \overline{\mathbb{B}}(S_\pi, \epsilon)). \tag{27}$$

and thus $\nu(\overline{\mathbb{B}}(S_\pi, \epsilon)) = \pi^x(\overline{\mathbb{B}}(S_\pi, \epsilon))$ for all ϵ, which make $\overline{\mathbb{B}}(S_\pi, \epsilon)$ both ν- and π^x-continuous. Note that there are at most countably many $\epsilon > 0$ for which this may fail, see [35, Chapter 3, Example 1.3]).

With the same argument used for (27) we also obtain for any $k \in \mathbb{N}$ that $\nu \mathcal{P}^k(\overline{\mathbb{B}}(S_\pi, \epsilon)) = \pi^x(\overline{\mathbb{B}}(S_\pi, \epsilon))$ with only countably many ϵ excluded, and so

$$\pi^\nu(\overline{\mathbb{B}}(S_\pi, \epsilon)) = \pi^x(\overline{\mathbb{B}}(S_\pi, \epsilon))$$

also needs to hold for all except countably many ϵ. Since $\pi^\nu \in \operatorname{inv} \mathcal{P}$, this implies that $\pi^\nu = \pi^x$ by Lemma 3.28 combined with Remark 3.29. For a general initial measure $\mu_0 \in \mathscr{P}(\mathcal{E})$, one has, yet again by Lebesgue's dominated convergence theorem, that

$$\mu_0 \mathcal{P}^k f = \int f(y) p^k(x, \mathrm{d}y) \mu_0(\mathrm{d}x) \to \int f(y) \pi^x(\mathrm{d}y) \mu_0(\mathrm{d}x) =: \pi^{\mu_0} f,$$

where π^x denotes the limit of $(\delta_x \mathcal{P}^k)$ and the measure π^{μ_0} is again invariant for $\mathcal{P}$. This completes the proof. □

Remark 3.31 (a.s. convergence)**:** If we were to choose X and Y in (25) such that $\mathcal{L}(X), \mathcal{L}(Y) \in \mathscr{P}(S_\pi)$, where $\pi \in \mathscr{E}$, then still $\gamma_k F \to \gamma F = 0$, where $\gamma \in C(\pi, \pi)$. For $(W, Z) \sim \gamma$ it still holds that $W = Z$ and hence

$$\left\| X_k^X - X_k^Y \right\| \to 0 \qquad \text{a.s.}$$

by monotonicity of $(\gamma_k F)$.

4. Examples: Stochastic Optimization and Inconsistent Convex Feasibility

To fix our attention we focus on the following optimization problem

$$\underset{\mu \in \mathscr{P}_2(\mathcal{E})}{\text{minimize}} \int_{\mathcal{E}} \mathbb{E}_\xi [f_{\xi^f}(x) + g_{\xi^g}(x)] \mu(dx). \tag{28}$$

The random variable with values on $I_f \times I_g$ are denoted $\xi = (\xi^f, \xi^g)$. This model covers deterministic composite optimization as a special case: I_f and I_g consist of single elements and the measure μ is a point mass.

The algorithms reviewed in this section rely on proximal, or simply *prox*, mappings of the functions f_i and g_i, denoted prox_{f_i} and prox_{g_i}. For proper, lsc and convex functions $f : \mathcal{E} \to \mathbb{R} \cup \{+\infty\}$, the proximal mapping [44] defined by

$$\text{prox}_f(x) := \underset{y}{\text{argmin}} \{ f(y) + \tfrac{1}{2} \|y - x\|^2 \}. \tag{29}$$

4.1. Stochastic convex forward-backward splitting

We begin with a general prescription of the forward-backward splitting algorithm together with abstract properties of the corresponding fixed point mapping, and then specialize this to more concrete instances. It is assumed throughout this subsection that $f_i : \mathcal{E} \to \mathbb{R}$ is continuously differentiable and convex for all $i \in I_f$ and that the extended-valued function $g_i : \mathcal{E} \to \mathbb{R} \cup \{+\infty\}$ is proper, lower semi-continuous (lsc) and convex for all $i \in I_g$.

Algorithm 2: Stochastic Convex Forward-Backward Splitting

Initialization: Set $X_0 \sim \mu_0 \in \mathscr{P}_2(\mathcal{E})$, $X_0 \sim \mu$, $t > 0$, and $(\xi_k)_{k \in \mathbb{N}}$ another i.i.d. sequence with values on $I_f \times I_g$ and $X_0 \perp\!\!\!\perp (\xi_k)$.

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

$$X_{k+1} = T^{FB}_{\xi_k} X_k := \text{prox}_{t g_{\xi_k^g}} \left(X_k - t \nabla f_{\xi_k^f}(X_k) \right) \tag{30}$$

When $f_{\xi^f}(x) = f(x) + \eta_{\xi^f} \cdot x$ and g_{ξ^g} is the zero function, then this is just steepest descents with linear noise discussed in Section 2.2. More generally, (30) with g_{ξ^g} the zero function models stochastic gradient descents, which is a central algorithmic template in many applications. To date, convergence results for these types of methods are limited to ergodic results or to a.s. convergence. As Proposition 2.5 shows, almost sure convergence requires the existence of a

common fixed point of all the randomly selected operators. Ergodic results on the other hand only provide access to one of the moments of the limiting distribution. Our analysis provides for information on all moments in the limit and does not require common fixed points.

For the next result it is helpful to recognize the forward-backward mapping as the composition of two mappings: $T_i^{FB} := \operatorname{prox}_{tg_i} \circ T_i^{GD_t}$ where $T_i^{GD_t} := \operatorname{Id} - t\nabla f_i$.

Proposition 4.1. *In addition to the standing assumptions, suppose that for all $i \in I_f$, ∇f_i is Lipschitz continuous with constant L on $\mathcal{E}$. Then for all $\alpha \in (0,1)$ and all step lengths $t \in \left(0, \frac{2\alpha}{L}\right]$ the following hold.*

(i) $T_i^{GD_t}$ is averaged on $\mathcal{E}$ with constant α;

(ii) T_i^{FB} is averaged on $\mathcal{E}$ with constant $\overline{\alpha} = \frac{2}{1+2/\max\{tL,1\}}$ for all $i \in I$.

(iii) Whenever there exists an invariant measure for the Markov operator $\mathcal{P}$ corresponding to (30), the distributions of the sequences of random variables converge to an invariant measure in the Prokhorov-Lèvy metric.

Proof. (i). By [5, Corollaire 10]

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

For $t = \frac{2\alpha}{L}$ we have $2t = \frac{t^2 L}{\alpha}$, and for all $x, y \in \mathcal{E}$

$$\begin{gathered}
\tfrac{t^2 L}{\alpha}\tfrac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \leq 2t\,\langle \nabla f(x) - \nabla f(y),\, x - y\rangle \\
\iff \\
\tfrac{1}{\alpha}\|t\nabla f(x) - t\nabla f(y)\|^2 \leq 2\,\langle t\nabla f(x) - t\nabla f(y),\, x - y\rangle \\
\iff \\
\|x - y\|^2 + \left(1 + \tfrac{1-\alpha}{\alpha}\right)\|t\nabla f(x) - t\nabla f(y)\|^2 \\
\leq 2\,\langle t\nabla f(x) - t\nabla f(y),\, x - y\rangle + \|x - y\|^2 \\
\iff \\
\|(x - t\nabla f(x)) - (y - t\nabla f(y))\|^2 \\
\leq \|x - y\|^2 - \tfrac{1-\alpha}{\alpha}\|t\nabla f(x) - t\nabla f(y)\|^2 \\
\iff \\
\|T_{GD_t}x - T_{GD_t}y\|^2 \leq \|x - y\|^2 - \tfrac{1-\alpha}{\alpha}\psi(x, y, T_{GD_t}x, T_{GD_t}y),
\end{gathered}$$

where the last implication follows from (11).

(ii) Since g_i is proper, convex and lsc for all i, the prox mapping is well-defined and averaged with constant $\alpha = 1/2$ on $\mathcal{E}$. The rest follows from part (i) and the calculus of compositions of averaged mappings [7, Proposition 4.32].

(iii). This follows from part (ii) and Theorem 2.9. □

Note that an upper bound on the step size t is $2/L$; the price to pay for taking such a large step is the loss of the averaging property (averaging constant $\alpha = 1$). The result also captures stochastic gradient descent as a special case: $g_i := 0$ for all i. The assumptions of Proposition 4.1 are not unusual, but weaker than the standard assumption of strong convexity [25]. The generality of global convergence and the information that this yields is new: convergence is to a distribution, not just the expectation. The result narrows the work of proving convergence of stochastic forward-backward algorithms to verifying that $\operatorname{inv}\mathcal{P}$ is nonempty. The next corollary shows how this is done for the special case of stochastic gradient descent.

Proposition 4.2 (existence of invariant measures and convergence: stochastic gradient descent)**.** *In problem (28) let $g_i(x) := 0$ for all i at each x. In addition to assumptions of Proposition 4.1, suppose that*

(i) ∇f_i is strongly monotone on $\mathcal{E}$ with the same constant for all $i \in I$:

$$\exists \tau_f > 0 : \quad \forall i \in I, \quad \tau_f \|x-y\|^2 \leq \langle \nabla f_i(x) - \nabla f_i(y), x-y\rangle \quad \forall x, y \in \mathcal{E}; \tag{31}$$

(ii) the expectation $\mathbb{E}[f_\xi(x)]$ attains a minimum at $\bar{x} \in \mathcal{E}$ with value $\mathbb{E}[f_\xi(\bar{x})] = \bar{p}$;

(iii) $\mathbb{E}\left[\|X_0 - \bar{x}\|^2\right] < \infty$ where X_0 is a random variable with distribution $\mu_0 \in \mathscr{P}_2(\mathcal{E})$.

Then for any $t \in (0, \min\{1/L, 1/\tau_f, \frac{\tau_f}{L^2}\}]$, the Markov operator corresponding to stochastic gradient descent with fixed step length t possesses invariant measures, and, when initialized with μ_0, the distributions of the iterates converge in the Prokhorov-Lèvy metric to an invariant measure.

Proof. In this case $T_i^{FB} = T_i^{GD_t} := \mathrm{Id} - t\nabla f_i$. By [41, Proposition 3.6], for any step size $t \leq \frac{\tau_f}{L^2}$, T^{GD_t} is averaged with constant $\alpha = 1/2$ on $\mathcal{E}$. Convergence to an invariant measure, whenever this exists, then follows from Theorem 2.9.

It remains to show that the corresponding Markov operator possesses invariant distributions. To establish this, note that

$$\begin{aligned}\|X_{k+1} - \bar{x}\|^2 =& \|X_k - t\nabla f_{\xi_k}(X_k) - \bar{x}\|^2 - \|X_{k+1} - X_k - t\nabla f_{\xi_k}(X_k)\|^2 \\ =& \|X_k - \bar{x}\|^2 - \|X_{k+1} - X_k\|^2 - 2t\left\langle \nabla f_{\xi_k}(X_k), X_{k+1} - X_k + X_k - \bar{x}\right\rangle. \end{aligned} \tag{32}$$

For functions with Lipschitz continuous gradients the following *growth condition* holds

$$\langle \nabla f_{\xi_k}(X_k), X_{k+1} - X_k\rangle \geq f_{\xi_k}(X_{k+1}) - f_{\xi_k}(X_k) - \frac{L}{2}\|X_{k+1} - X_k\|^2. \tag{33}$$

The assumption of strong monotonicity of the gradients implies [40, Proposition 2.2]

$$\langle \nabla f_{\xi_k}(X_k),\ X_k - \bar{x}\rangle \geq f_{\xi_k}(X_k) - f_{\xi_k}(\bar{x}) + \tfrac{\tau_f}{2}\|X_k - \bar{x}\|^2.$$

Taking the expectation and interchanging the gradient with the expectation yields

$$\langle \nabla \mathbb{E}[f_\xi(X_k)], X_k - \bar{x}\rangle \geq \mathbb{E}[f_\xi(X_k)] - \bar{p} + \frac{\tau_f}{2}\|X_k - \bar{x}\|^2. \tag{34}$$

Putting (32)-(34) together yields

$$\begin{aligned}\mathbb{E}\left[\|X_{k+1} - \bar{x}\|^2\right] &\leq (1 - t\tau_f)\mathbb{E}\left[\|X_k - \bar{x}\|^2\right] - (1 - tL)\mathbb{E}\left[\|X_{k+1} - X_k\|^2\right] - 2t\left(\mathbb{E}[f_{\xi_k}(X_{k+1})] - \bar{p}\right) \\ &\leq (1 - t\tau_f)\mathbb{E}\left[\|X_k - \bar{x}\|^2\right] + 2t\bar{p} \quad \forall t \in (0, \min\{1/L, 1/\tau_f\}].\end{aligned}$$

Thus, whenever $t \in (0, \min\{1/L, 1/\tau_f\}]$

$$\mathbb{E}\left[\|X_k - \bar{x}\|^2\right] \leq (1 - t\tau_f)^k \mathbb{E}\left[\|X_0 - \bar{x}\|^2\right] + 2t\bar{p}\sum_{i=0}^{k-1}(1 - t\tau_f)^i \leq \mathbb{E}\left[\|X_0 - \bar{x}\|^2\right] + \frac{2\bar{p}}{\tau_f}.$$

So the sequence $\left(\mathbb{E}\left[\|X_k - \bar{x}\|^2\right]\right)_{k\in\mathbb{N}}$ is bounded since, by assumption, $\mathbb{E}\left[\|X_0 - \bar{x}\|^2\right]$ is bounded. The existence of an invariant measure then follows from Theorem 3.3. □

As noted in [41, pp1171], the step t in the stochastic gradient could be taken much larger than the conservative estimates given here. A justification of this is beyond the scope of this study.

4.2. Stochastic Douglas-Rachford

Another prevalent algorithm for nonconvex problems is the Douglas-Rachford algorithm [39]. This is based on compositions of *reflected prox mappings*:

$$R_f := 2\,\mathrm{prox}_f - \mathrm{Id}\,. \tag{35}$$

For this algorithm we assume only convexity of the constituent functions. Algorithm 3 has been

Algorithm 3: Stochastic Douglas-Rachford Splitting

Initialization: Set $X_0 \sim \mu_0 \in \mathscr{P}_2(\mathcal{E})$, $X_0 \sim \mu$, and $(\xi_k)_{k\in\mathbb{N}}$ another i.i.d. sequence with $\xi_k = (\xi_k^f, \xi_k^g)$ taking values on $I_f \times I_g$ and $X_0 \perp\!\!\!\perp (\xi_k)$.

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

$$X_{k+1} = T_{\xi_k}^{DR} X_k := \frac{1}{2}\left(R_{f_{\xi_k^f}} \circ R_{g_{\xi_k^g}} + \mathrm{Id}\right)(X_k) \tag{36}$$

studied for solving large-scale, convex optimization and monotone inclusions (see for example [12, 19]).

Proposition 4.3. *Suppose that for all $i = (i_1, i_2) \in I_f \times I_g$ the extended-valued functions $f_{i_1} : \mathcal{E} \to \mathbb{R} \cup \{+\infty\}$ and $g_{i_2} : \mathcal{E} \to \mathbb{R} \cup \{+\infty\}$ are proper, lsc and convex on $\mathcal{E}$. Then whenever there exists an invariant measure for the Markov operator $\mathcal{P}$ corresponding to (36), the distributions of the sequences of random variables converge to an invariant measure in the Prokhorov-Lèvy metric.*

Proof. This follows immediately from [39, Proposition 2] (which establishes that T_i^{DR} is averaged with constant $\alpha = 1/2$ on $\mathcal{E}$ for all i) and Theorem 2.9. □

4.3. Inconsistent set feasibility

We conclude this study with our explanation for the numerical behavior observed in Fig. 1. This is an affine feasibility problem:

$$\text{Find} \quad x \in L := \cap_{j\in I}\{x \mid \langle a_j, x\rangle = b_j\}. \tag{37}$$

When the intersection is empty we say that the problem is *inconsistent*. Consistent or not, we apply the method of cyclic projections (2). Even though the projectors onto the corresponding problems have an analytic expression, this representation can only be evaluated to finite precision. The trick here is to view the algorithm not as inexact cyclic projections onto deterministic hyperplanes, but rather as *exact* projections onto randomly selected hyperplanes.

Indeed, consider the following generalized affine noise model for a single affine subspace: $H_{\overline{x}}^{(\xi,\zeta)} = \{x \in \mathbb{R}^n \mid \langle a + \xi, x - \overline{x}\rangle = \zeta\}$, where $a \in \mathbb{R}^n$ and $\overline{x}$ satisfies $A\overline{x} = b$ for a given $b \in \mathbb{R}$ and noise $(\xi, \zeta) \in \mathbb{R}^n \times \mathbb{R}$ is independent. The key conceptual distinction is that the analysis proceeds with *exact* projections onto randomly selected hyperplanes $H_{\overline{x}}^{(\xi,\zeta)}$, rather than working with inexact projections onto deterministic hyperplanes.

The main result of this section uses the following result about the more familiar contractive mappings.

Proposition 4.4. *Let $T_i : \mathcal{E} \to \mathcal{E}$ for $i \in I$ and let $\Phi : \mathcal{E} \times I \to \mathcal{E}$ be given by $\Phi(x, i) := T_i(x)$. Denote by $\mathcal{P}$ the Markov operator with update function Φ and transition kernel p defined by (4). Suppose that Φ is a* contraction in expectation *with constant $r < 1$, i.e. $\mathbb{E}[\|\Phi(x,\xi) - \Phi(y,\xi)\|^2] \leq r^2\|x-y\|^2$ for all $x, y \in \mathcal{E}$. Suppose in addition that there exists $y \in \mathcal{E}$ with $\mathbb{E}[\|\Phi(y,\xi) - y\|^2] < \infty$. Then the following hold.*

(i) There exists a unique invariant measure $\pi \in \mathscr{P}_2(\mathcal{E})$ for $\mathcal{P}$ and

$$W_2(\mu_0\mathcal{P}^n, \pi) \leq r^n W_2(\mu_0, \pi)$$

for all $\mu_0 \in \mathscr{P}_2(\mathcal{E})$; that is, the sequence (μ_k) defined by $\mu_{k+1} = \mu_k\mathcal{P}$ converges to π linearly (geometrically) from any initial measure $\mu_0 \in \mathscr{P}_2(\mathcal{E})$.

(ii) Φ is averaged in expectation with constant $\alpha = (1+r)/2$:

$$\begin{aligned} &\text{for } \alpha = (1+r)/2, \quad \forall x, y \in \mathcal{E}, \\ &\mathbb{E}\left[\|\Phi(x,\xi) - \Phi(y,\xi)\|^2\right] \leq \|x-y\|^2 - \tfrac{1-\alpha}{\alpha}\mathbb{E}\left[\psi(x, y, \Phi(x,\xi), \Phi(y,\xi))\right]. \end{aligned} \tag{38}$$

Proof. Note that for any pair of distributions $\mu_1, \mu_2 \in \mathscr{P}_2(\mathcal{E})$ and an optimal coupling $\gamma \in C_*(\mu_1, \mu_2)$ (possible by Lemma A.7) it holds that

$$\begin{aligned} W_2^2(\mu_1\mathcal{P}, \mu_2\mathcal{P}) &\leq \int_{\mathcal{E}\times\mathcal{E}} \mathbb{E}[d^2(\Phi(x,\xi), \Phi(y,\xi))]\ \gamma(dx, dy) \\ &\leq r^2 \int_{\mathcal{E}\times\mathcal{E}} d^2(x,y)\ \gamma(dx, dy) = r^2 W_2^2(\mu_1, \mu_2), \end{aligned}$$

where ξ is independent of γ. Moreover, $\mathcal{P}$ is a self-mapping on $\mathscr{P}_2(\mathcal{E})$. To see this let $\mu \in \mathscr{P}_2(\mathcal{E})$ independent of ξ and let y be a point in $\mathcal{E}$ where $\mathbb{E}[\|\Phi(y,\xi) - y\|^2] < \infty$. Then by the triangle inequality and the contraction property

$$\begin{aligned} &\int_{\mathcal{E}} \mathbb{E}[\|\Phi(x,\xi) - y\|^2]\ \mu(dx) \\ &\quad\leq 4\left(\int_{\mathcal{E}} \mathbb{E}[\|\Phi(x,\xi) - \Phi(y,\xi)\|^2]\ \mu(dx) + \mathbb{E}[\|\Phi(y,\xi) - y\|^2]\right) \\ &\quad\leq 4\left(\int_{\mathcal{E}} r^2\|x-y\|^2\ \mu(dx) + \mathbb{E}[\|\Phi(y,\xi) - y\|^2]\right) < \infty. \end{aligned}$$

Therefore $\mu\mathcal{P} \in \mathscr{P}_2(\mathcal{E})$. Altogether, this establishes that $\mathcal{P}$ is a contraction on the separable complete metric space $(\mathscr{P}_2(\mathcal{E}), W_2)$ and hence Banach's Fixed Point Theorem yields existence and uniqueness of $\operatorname{inv}\mathcal{P}$ and linear convergence of the fixed point sequence.

To see (ii), note that, by (11),

$$\begin{aligned} \mathbb{E}[\psi(x, y, T_\xi x, T_\xi y)] &= \mathbb{E}\left[\|(x - \Phi(x,\xi)) - (y - \Phi(y,\xi))\|^2\right] \\ &= \|x-y\|^2 + \mathbb{E}\left[\|\Phi(x,\xi) - \Phi(y,\xi)\|^2 - 2\langle x - y, \Phi(x,\xi) - \Phi(y,\xi)\rangle\right] \\ &\leq (1+r)^2\|x-y\|^2, \end{aligned} \tag{39}$$

where the last inequality follows from the Cauchy-Schwarz inequality and the fact that $\Phi(\cdot, \xi)$ is a contraction in expectation. Again using the contraction property and (39) we have

$$\begin{aligned} \mathbb{E}\left[\|\Phi(x,\xi) - \Phi(y,\xi)\|^2\right] &\leq \|x-y\|^2 - (1-r^2)\|x-y\|^2 \\ &\leq \|x-y\|^2 - \tfrac{1-r^2}{(1+r)^2}\mathbb{E}[\psi(x, y, T_\xi x, T_\xi y)]. \end{aligned}$$

The right hand side of this inequality is just the characterization (11) of mappings that are averaged in expectation with $\alpha = (1+r)/2$. □

The simple example of a single Euclidean projector onto an affine subspace ($I = \{1\}$, and T_1 the orthogonal projection onto an affine subspace) shows that the statement of Proposition 4.4 fails without the assumption that T is a contraction.

Proposition 4.5. *Given $a \in \mathbb{R}^n$, $b \in \mathbb{R}$, define the hyperplane $H = \{y \,|\, \langle a, y\rangle = b\}$ and fix $\overline{x} \in H$. Define the random mapping $T_{(\xi,\zeta)} : \mathbb{R}^n \to \mathbb{R}^n$ by*

$$T_{(\xi,\zeta)}x := P_{H_{\overline{x}}^{(\xi,\zeta)}}x = x - \frac{\langle a+\xi, x-\overline{x}\rangle - \zeta}{\|a+\xi\|^2}(a+\xi)$$

where $(\xi, \zeta) \in \mathbb{R}^n \times \mathbb{R}$ is a vector of independent random variables.

Algorithm (1) with this random function initialized with any $\mathbb{R}^n$-valued random variable X_0 with distribution $\mu^0 \in \mathscr{P}(\mathbb{R}^n)$ converges to an invariant distribution whenever this exists.

If (ξ, ζ) satisfy

$$d := \mathbb{E}\left[\frac{(b+\zeta)^2}{\|a+\xi\|^2}\right] < \infty, \tag{40a}$$

$$c := \inf_{z\in\mathbb{S}} \mathbb{E}\left[\frac{\langle a+\xi, z\rangle^2}{\|a+\xi\|^2}\right] > 0 \tag{40b}$$

where $\mathbb{S}$ is the set of unit vectors in $\mathbb{R}^n$ then the Markov operator $\mathcal{P}$ generated by $T_{(\xi,\zeta)}$ possesses a unique invariant distribution and Algorithm (1) initialized with any X_0 with distribution $\mu^0 \in \mathscr{P}(\mathbb{R}^n)$ converges linearly to this distribution.

Proof. Each mapping $T_{(\xi,\zeta)}$ is the orthogonal projector onto the hyperplane $H_{\overline{x}}^{(\xi,\zeta)}$, and so is averaged with constant $\alpha = 1/2$. Without regard to the assumptions on the noise, based solely on Theorem 2.9 we conclude that the iteration converges to a point in $\operatorname{inv}\mathcal{P}$ whenever this is nonempty.

Existence of invariant distributions follows from the assumptions on the noise which impliy that $T_{(\xi,\zeta)}$ is actually a contraction in expectation. To see this, an elementary calculation shows that

$$\left\|T_{(\xi,\zeta)}x - T_{(\xi,\zeta)}y\right\|^2 = \left(1 - \cos^2\left(\tfrac{a+\xi}{\|a+\xi\|}, \tfrac{x-y}{\|x-y\|}\right)\right)\|x-y\|^2.$$

Taking the expectation over (ξ, ζ) yields

$$\mathbb{E}\left[\left\|T_{(\xi,\zeta)}x - T_{(\xi,\zeta)}y\right\|^2\right] \leq (1-c)\|x-y\|^2.$$

From Proposition 4.4 we get that there exists a *unique* invariant measure π_0 for $\mathcal{P}$ (even $\pi_0 \in \mathscr{P}_2$) and that it satisfies

$$W_2^2(\mu\mathcal{P}^k, \pi_0) \leq (1-c)^k W_2^2(\mu, \pi_0).$$

Convergence is therefore linear. □

Note that the noise satisfying (40) depends implicitly on the point $\overline{x}$, which will determine the concentration of the invariant distribution of the Markov operator. This corresponds to the fact that the exact projection, while unique, depends on the point being projected. One would expect the invariant distribution to be concentrated on the exact projection.

Extending this model to finitely many distorted affine subspaces as illustrated in Fig. 1 (i.e. we are given m normal vectors $a_1, \ldots, a_m \in \mathcal{E}$ and displacement vectors $b_1, \ldots, b_m$) yields

a stochastic version of cyclic projections (2) which converges linearly (geometrically) in the Wasserstein metric to a unique invariant measure for the given noise model.

Indeed, for a collection of (not necessarily distinct) points $\overline{x}_j \in H_j := \{y \,|\, \langle a_j, y\rangle = b_j\}$ $(j = 1, 2, \ldots, m)$ denote by $P^j_{(\xi_j,\zeta_j)}$ the exact projection onto the j-th random affine subspace centered on $\overline{x}_j$, i.e.

$$P^j_{(\xi_j,\zeta_j)}x = x - \frac{\langle a_j + \xi_j, x - \overline{x}_j\rangle - \zeta_j}{\|a_j + \xi_j\|^2}(a_j + \xi_j),$$

where $(\xi_i)_{i=1}^m$ and $(\zeta_i)_{i=1}^m$ are i.i.d. and $(\xi_i) \perp\!\!\!\perp (\zeta_i)$. The stochastic cyclic projection mapping is

$$T_{(\xi,\zeta)}x = P^m_{(\xi_m,\zeta_m)} \circ \ldots \circ P^1_{(\xi_1,\zeta_1)}x, \qquad x \in \mathcal{E}$$

where $(\xi, \zeta) = ((\xi_m, \xi_{m-1}, \ldots, \xi_1), (\zeta_m, \zeta_{m-1}, \ldots, \zeta_1))$. Following the same pattern of proof as Proposition 4.5 we see that $T_{(\xi,\zeta)}$ is a contraction in expectation:

$$\mathbb{E}\left[\left\|T_{(\xi,\zeta)}x - T_{(\xi,\zeta)}y\right\|^2\right] \leq (1 - c)^m\|x - y\|^2$$

where

$$c := \min_{j=1,\ldots,m} \inf_{z\in\mathbb{S}} \mathbb{E}\left[\frac{\langle a_j + \xi_j, z\rangle^2}{\|a_j + \xi_j\|^2}\right] > 0. \tag{41}$$

Hence, there exists a unique invariant measure and $(\mu\mathcal{P}^k)$ converges geometrically to it in the W_2 metric. Note that there is no assumption of summability of the errors. In fact, for the random function iteration based on the usual additive noise model it can be shown that the Markov operator does not possess invariant distributions. The assumption of summable errors in this case is tantamount to an assumption of no noise.

A. Appendix

Proposition A.1 (Convergence with subsequences)**.** *Let (G, d) be a metric space. Let (x_k) be a sequence on G with the property that any subsequence has a convergent subsequence with the same limit $x \in G$. Then $x_k \to x$.*

Proof. Assume that $x_k \not\to x$, i.e. there exists $\epsilon > 0$ such that for all $N \in \mathbb{N}$ there is $k = k(N) \geq N$ with $d(x_k, x) \geq \epsilon$. But by assumption the subsequence $(x_{k(N)})_{N\in\mathbb{N}}$ has a convergent subsequence with limit x, which is a contradiction and hence the assumption is false. □

Remark A.2: In a compact metric space, it is enough, that all cluster points are the same, because then every subsequence has a convergent subsequence.

Lemma A.3. *Let (G, d) be a separable complete metric space and let the metric $d_\times$ on $G \times G$ satisfy*

$$d_\times\left(\begin{pmatrix} x_k \\ y_k\end{pmatrix}, \begin{pmatrix} x \\ y\end{pmatrix}\right) \to 0 \qquad \Leftrightarrow \qquad d(x_k, x) \to 0 \quad \text{and} \quad d(y_k, y) \to 0. \tag{42}$$

Then $\mathcal{B}(G \times G) = \mathcal{B}(G) \otimes \mathcal{B}(G)$.

Proof. First we note that for $A, B \subset G$ it holds that $A \times B$ is closed in $(G \times G, d_\times)$ if and only if A, B are closed in (G, d) by (42). Since the σ-algebra $\mathcal{B}(G) \otimes \mathcal{B}(G)$ is generated by the family $\mathcal{A} := \{A_1 \times A_2 \,|\, A_1, A_2 \subset G \text{ closed}\}$. One has $\mathcal{B}(G \times G) \supset \mathcal{B}(G) \otimes \mathcal{B}(G)$

For the other direction, note that any metric $d_\times$ with the property (42) has the same open and closed sets. If A is closed in $(G \times G, d_\times)$ and $\tilde{d}_\times$ is another metric on $G \times G$ satisfying (42), then for $(a_k, b_k) \in A$ with $(a_k, b_k) \to (a, b) \in G \times G$ w.r.t. $\tilde{d}_\times$ it holds that $d(a_k, a) \to 0$ and $d(b_k, b) \to 0$ and hence $d_\times((a_k, b_k), (a, b)) \to 0$ as $k \to \infty$, i.e. $(a, b) \in A$, so A is closed in $(G \times G, \tilde{d}_\times)$. It follows that all open sets in $(G \times G, d_\times)$ are the same for any metric that satisfies (42). So, without loss of generality, let

$$d_\times \left(\binom{x_1}{y_1}, \binom{x_2}{y_2} \right) = \max(d(x_1, x_2), d(y_1, y_2)). \tag{43}$$

Moreover, separability of $G \times G$ yields that any open set is the countable union of balls: there exists a sequence (u_k) on U that is dense for $U \subset G \times G$ open. We can find a sequence of constants $\epsilon_k > 0$ with $\bigcup_k \mathbb{B}(u_k, \epsilon_k) \subset U$. If there exists $x \in U$, which is not covered by any ball, then we may enlarge a ball, so that x is covered: since there exists $\epsilon > 0$ with $\mathbb{B}(x, \epsilon) \subset U$ and there exists $m \in \mathbb{N}$ with $d(x, u_m) < \epsilon/2$ by denseness, we may set $\epsilon_m = \epsilon/2$ to get $x \in \mathbb{B}(u_m, \epsilon_m) \subset \mathbb{B}(x, \epsilon) \subset U$. Now to continue the proof, let $d_\times$ be given by (43). Then for any open $U \subset G \times G$ there exists a sequence (u_k) on U and a corresponding sequence of positive constants (ϵ_k) such that $U = \bigcup_k \mathbb{B}(u_k, \epsilon_k)$. This together with the fact that

$$\mathbb{B}(u_k, \epsilon_k) = \mathbb{B}(u_{k,1}, \epsilon_k) \times \mathbb{B}(u_{k,2}, \epsilon_k) \in \mathcal{B}(G) \otimes \mathcal{B}(G) \quad (u_k = (u_{k,1}, u_{k,2}) \in G \times G)$$

yields $\mathcal{B}(G \times G) \subset \mathcal{B}(G) \otimes \mathcal{B}(G)$, which establishes equality of the σ-algebras. □

Lemma A.4 (couplings)**.** *Let G be a Polish space and let $\mu, \nu \in \mathscr{P}(G)$. Let $\gamma \in C(\mu, \nu)$, where*

$$C(\mu, \nu) := \{\gamma \in \mathscr{P}(G \times G) \,|\, \gamma(A \times G) = \mu(A),\ \gamma(G \times A) = \nu(A) \quad \forall A \in \mathcal{B}(G)\}, \tag{44}$$

then

(i) $\operatorname{supp} \gamma \subset \operatorname{supp} \mu \times \operatorname{supp} \nu$,

(ii) $\overline{\{x \,|\, (x, y) \in \operatorname{supp} \gamma\}} = \operatorname{supp} \mu$.

Proof. We let the product space be equipped with the metric in (43) (constituting a separable complete metric space since G is Polish).

(i) Suppose $(x, y) \in \operatorname{supp} \gamma$ and let $\epsilon > 0$, then

$$\mu(\mathbb{B}(x, \epsilon)) = \gamma(\mathbb{B}(x, \epsilon) \times G) \geq \gamma(\mathbb{B}(x, \epsilon) \times \mathbb{B}(y, \epsilon)) = \gamma(\mathbb{B}((x, y), \epsilon)) > 0.$$

Analogously, we have $\nu(\mathbb{B}(y, \epsilon)) > 0$. So $(x, y) \in \operatorname{supp} \mu \times \operatorname{supp} \nu$.

(ii) Suppose $x \in \operatorname{supp} \mu$, then $\gamma(\mathbb{B}(x, \epsilon) \times G) > 0$ for all $\epsilon > 0$. Since G is Polish, the support of the measure is nonempty whenever the measure is nonzero, and (again, since G is Polish) the support of the measure is closed, there either exists $y \in G$ with $(x, y) \in \operatorname{supp} \gamma$ or there exists a sequence (x_k, y_k) on $\operatorname{supp} \gamma$ with $x_k \to x$ as $k \to \infty$. Hence the assertion follows. □

Lemma A.5 (convergence in product space)**.** *Let G be a Polish space and suppose $(\mu_k), (\nu_k) \subset \mathscr{P}(G)$ are tight sequences. Let $X_k \sim \mu_k$ and $Y_k \sim \nu_k$ and denote by $\gamma_k = \mathcal{L}((X_k, Y_k))$ the joint law of X_k and Y_k. Then (γ_k) is tight.*
If furthermore, $\mu_k \to \mu \in \mathscr{P}(G)$ and $\nu_k \to \nu \in \mathscr{P}(G)$, then cluster points of (γ_k) are in $C(\mu, \nu)$, where the set of couplings $C(\mu, \nu)$ is defined in (44) in Lemma A.4.

Proof. By tightness of (μ_k) and (ν_k), there exists for any $\epsilon > 0$ a compact set $K \subset G$ with $\mu_k(G \setminus K) < \epsilon/2$ and $\nu_k(G \setminus K) < \epsilon/2$ for all $n \in \mathbb{N}$, so also

$$\begin{aligned}\gamma_k(G \times G \setminus K \times K) &\leq \gamma_k((G \setminus K) \times G) + \gamma_k(G \times (G \setminus K)) \\ &= \mu_k(G \setminus K) + \nu_k(G \setminus K) \\ &< \epsilon\end{aligned}$$

for all $k \in \mathbb{N}$, implying tightness of (γ_k). By Prokhorov's Theorem [10], every subsequence of (γ_k) has a convergent subsequence $\gamma_{k_j} \to \gamma$ as $j \to \infty$ where $\gamma \in \mathscr{P}(G \times G)$.

It remains to show that $\gamma \in C(\mu, \nu)$. Indeed, since for every $f \in C_b(G \times G)$ we have $\gamma_{n_k} f \to \gamma f$, we can choose $f(x, y) = g(x)\mathbb{1}_G(y)$ with $g \in C_b(G)$. Also,

$$\mu g \leftarrow \mu_{n_k} g = \gamma_{n_k} f \to \gamma f = \gamma(\cdot \times G)g,$$

which implies the equality $\mu = \gamma(\cdot \times G)$. Similarly $\nu = \gamma(G \times \cdot)$ and hence $\gamma \in C(\mu, \nu)$. □

Lemma A.6 (properties of the Prokhorov-Lèvy distance)**.** *Let (G, d) be a separable complete metric space.*

(i) The Prokhorov-Lèvy distance (Definition 2.6) has the representation

$$d_P(\mu, \nu) = \inf\left\{\epsilon > 0 \,\middle|\, \inf_{\mathcal{L}(X,Y) \in C(\mu,\nu)} \mathbb{P}(d(X, Y) > \epsilon) \leq \epsilon\right\},$$

where the set of couplings $C(\mu, \nu)$ is defined in (44) in Lemma A.4. Furthermore, the inner infimum for fixed $\epsilon > 0$ is attained and the outer infimum is also attained.

(ii) $d_P(\mu, \nu) \in [0, 1]$.

(iii) d_P metrizes convergence in distribution, i.e. for $\mu_k, \mu \in \mathscr{P}(G)$, $k \in \mathbb{N}$ the sequence μ_k converges to μ in distribution if and only if $d_P(\mu_k, \mu) \to 0$ as $k \to \infty$.

(iv) $(\mathscr{P}(G), d_P)$ is a separable complete metric space.

(v) For $\mu_j, \nu_j \in \mathscr{P}(G)$ and $\lambda_j \in [0, 1]$, $j = 1, \ldots, m$ with $\sum_{j=1}^m \lambda_j = 1$ we have

$$d_P(\sum_j \lambda_j \mu_j, \sum_j \lambda_j \nu_j) \leq \max_j d_P(\mu_j, \nu_j).$$

Proof. (i) See [55, Corollary to Theorem 11] for the first assertion. To see that the infimum is attained, let $\gamma_k \in C(\mu, \nu)$ be a minimizing sequence, i.e. for $(X_k, Y_k) \sim \gamma_k$ it holds that $\mathbb{P}(d(X_k, Y_k) > \epsilon) = \gamma_k(U_\epsilon) \to \inf_{(X,Y) \in C(\mu,\nu)} \mathbb{P}(d(X, Y) > \epsilon)$, where $U_\epsilon := \{(x, y) \,|\, d(x, y) > \epsilon\} \subset G \times G$ is open. The sequence (γ_k) is tight and for a cluster point γ we have $\gamma \in C(\mu, \nu)$ by Lemma A.5. From [46, Theorem 36.1] it follows that $\gamma(U_\epsilon) \leq \liminf_j \gamma_{k_j}(U_\epsilon)$.
To see, that the outer infimum is attained, let (ϵ_k) be a minimizing sequence, chosen to be monotonically nonincreasing with limit $\epsilon \geq 0$. One has that $U_\epsilon = \bigcup_k U_{\epsilon_k}$ where $U_{\epsilon_k} \supset U_{\epsilon_{k+1}}$ and hence $\gamma(U_\epsilon) = \lim_k \gamma(U_{\epsilon_k}) \leq \lim_k \epsilon_k = \epsilon$.

(ii) Clear by (i).

(iii) See [10].

(iv) See [49, Lemma 1.4].

(v) If $\epsilon > 0$ is such that $\mu_j(A) \leq \nu_j(\mathbb{B}(A,\epsilon)) + \epsilon$ and $\nu_j(A) \leq \mu_j(\mathbb{B}(A,\epsilon)) + \epsilon$ for all $j = 1, \ldots, m$ and all $A \in \mathcal{B}(G)$, then also $\sum_j \lambda_j \mu_j(A) \leq \sum_j \lambda_j \nu_j(\mathbb{B}(A,\epsilon)) + \epsilon$ as well as $\sum_j \lambda_j \nu_j(A) \leq \sum_j \lambda_j \mu_j(\mathbb{B}(A,\epsilon)) + \epsilon$.

□

Lemma A.7 (properties of the Wasserstein metric)**.** *Recall $\mathscr{P}_p(G)$ and W_p from Definition 2.6.*

(i) The representation of $\mathscr{P}_p(G)$ is independent of x and for $\mu, \nu \in \mathscr{P}_p(G)$ the distance $W_p(\mu,\nu)$ is finite.

(ii) The distance $W_p(\mu,\nu)$ is attained when it is finite.

(iii) The metric space $(\mathscr{P}_p(G), W_p(G))$ is complete and separable.

(iv) If $W_p(\mu_k, \mu) \to 0$ as $k \to \infty$ for the sequence (μ_k) on $\mathscr{P}(G)$, then $\mu_k \to \mu$ as $k \to \infty$.

Proof. (i) See [58, Remark after Definition 6.4].

(ii) From Lemma A.5 we know that a minimizing sequence (γ_k) for $W_p(\mu,\nu)$ is tight and hence there is a cluster point $\gamma \in C(\mu,\nu)$. By continuity of the metric d it follows that d is lsc and bounded from below and from [56, Theorem 9.1.5] it follows that $\gamma d \leq \liminf_j \gamma_{k_j} d = W_p(\mu,\nu)$.

(iii) See [58, Theorem 6.9].

(iv) See [58, Theorem 6.18].

□

Note that the converse to Lemma A.7 (iv) does not hold.

References

[1] D. Ariza-Ruiz, L. Leuştean, and G. López-Acedo. Firmly nonexpansive mappings in classes of geodesic spaces. *Trans. Am. Math. Soc.*, 366(8):4299–4322, 2014.

[2] D. Ariza-Ruiz, G. López-Acedo, and A. Nicolae. The asymptotic behavior of the composition of firmly nonexpansive mappings. *J Optim Theory Appl*, 167:409–429, 2015.

[3] J.-B. Baillon and R. E. Bruck. The rate of asymptotic regularity is $o(\sqrt{n})$. In A. G. Kartsatos, editor, *Theory and applications of nonlinear operators of accretive and monotone type*, volume 178 of *Lecture Notes in Pure and Appl. Math.*, pages 51–81. Marcel Dekker, New York, 1996.

[4] 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.

[5] J.-B. Baillon and G. Haddad. Quelques propriétés des opérateurs angle-bornes et n-cycliquement monotones. *Isr. J. Math.*, 26:137–150, 1977.

[6] C. Baldassi, C. Borgs, J. T. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina. Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes. *Proceedings of the National Academy of Sciences*, 113(48):E7655–E7662, 2016.

[7] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books Math./Ouvrages Math. SMC. Springer, New York, 2011.

[8] R. Bellet. Ergodic properties of Markov processes. *Lecture Notes in Mathematics*, 1881:1–39, 2006.

[9] A. Bërdëllima, F. Lauster, and D. R. Luke. α-firmly nonexpansive operators on metric spaces. *J. Fixed Point Theory Appl.*, 24, 2022.

[10] P. Billingsley. *Convergence of probability measures. 2nd ed.* Chichester: Wiley, 2nd ed. edition, 1999.

[11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundation and Trends in Machine Learning*, 3(1):1–122, 2011.

[12] 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.

[13] F. E. Browder. Convergence theorems for sequences of nonlinear operators in Banach spaces. *Math. Z.*, 100:201–225, 1967.

[14] 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.

[15] 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.

[16] 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.

[17] D. Butnariu, A. N. Iusem, and R. S. Burachik. Iterative methods of solving stochastic convexfeasibility problems and applications. *Computational Optimization and Applications*, 15:269–307, 2000.

[18] D. Butnariu, S. Reich, and A. J. Zaslavski. Asymptotic behavior of inexact orbits for a class of operators in complete metric spaces. *J. Appl. Anal.*, 13(1):1–11, 2007.

[19] V. Cevher, B. C. Vũ, and A. Yurtsever. *Stochastic Forward Douglas-Rachford Splitting Method for Monotone Inclusions*, pages 149–179. Springer International Publishing, Cham, 2018.

[20] W. Cheney and A. A. Goldstein. Proximity maps for convex sets. *Proc. Amer. Math. Soc.*, 10(3):448–450, 1959.

[21] P. L. Combettes and J. Eckstein. Asynchronous block-iterative primal-dual decomposition methods for monotone inclusions. *Math. Program.*, 168(1):645–672, Mar 2018.

[22] P. L. Combettes, S. Salzo, and S. Villa. Consistent learning by composite proximal thresholding. *Math. Program.*, 167(1 (B)):99–127, 2018.

[23] R. Cominetti, J.A. Soto, and J. Vaisman. On the rate of convergence of krasnosel'skiĭ-mann iterations and their connection with sums of bernoullis. *Israel J. Math.*, pages 1–16, 2013.

[24] P. Diaconis and D. Freedman. Iterated Random Functions. *SIAM Review*, 41(1):45–76, 1999.

[25] A. Dieuleveut, A. Durmus, and F. Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *Ann. Stat.*, 48(3):1348–1382, 2020.

[26] Caroline Geiersbach and Georg Ch. Pflug. Projected stochastic gradients for convex constrained problems in Hilbert spaces. *SIAM J. Optim.*, 29(3):2079–2099, 2019.

[27] R. M. Gower and P. Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.

[28] M. Hairer. Convergence of Markov processes. *Lecture notes, University of Warwick*, page 39, 2016.

[29] M. Hardt, T. Ma, and B. Recht. Gradient descent learns linear dynamical systems. *J. Mach. Learn. Res.*, 19:44, 2018. Id/No 29.

[30] N. Hermer. *Random Functions Iterations for Stochastic Feasibility Problems*. PhD thesis, University of Göttingen, SUB Göttingen, April 2019.

[31] N. Hermer, D. R. Luke, and A. Sturm. Random function iterations for consistent stochastic feasibility. *Numer. Funct. Anal. Opt.*, 40(4):386–420, 2019.

[32] O. Hernández-Lerma and J. B. Lasserre. *Markov chains and invariant probabilities*. Basel: Birkhäuser, 2003.

[33] A. N. Iusem, T. Pennanen, and B. F. Svaiter. Inexact variants of the proximal point algorithm without monotonicity. *SIAM J. Optim.*, 13(4):1080–1097, 2003.

[34] O. Kallenberg. *Foundations of Modern Probability*. Probability and Its Applications. Springer, New York, 1997.

[35] L. Kuipers and H. Niederreiter. Uniform distribution of sequences. Pure and Applied Mathematics. New York: Wiley-Interscience. (1974)., 1974.

[36] M. N. Lam, H. N. Phuong, P. Richtárik, K. Scheinberg, M. Takáč, and M. van Dijk. New convergence aspects of stochastic gradient algorithms. *J. Mach. Learn. Res.*, 20:49, 2019. Id/No 176.

[37] A. Lasota and T. Szarek. Lower bound technique in the theory of a stochastic differential equation. *J. Differ. Equations*, 231(2):513–533, 2006.

[38] F. Lauster and D. R. Luke. Convergence of proximal splitting algorithms in CAT(κ) spaces and beyond. *Fixed Point Theory Algorithms Sci Eng*, 2021(13), 2021.

[39] P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.*, 16:964–979, 1979.

[40] D. R. Luke and R. Shefi. A globally linearly convergent method for pointwise quadratically supportable convex-concave saddle point problems. *J. Math. Anal. Appl.*, 457(2):1568–1590, 2018.

[41] 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.

[42] P. Mattila. *Geometry of Sets and Measures in Euclidean Spaces: Fractals and Rectifiability.* Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1995.

[43] S. Meyn and R. L. Tweedie. *Markov chains and stochastic stability.* Cambridge: Cambridge University Press, 2009.

[44] J. J. Moreau. Proximité et dualité dans un espace Hilbertian. *Bull. de la Soc. Math. de France*, 93(3):273–299, 1965.

[45] A. Nedić. Random algorithms for convex minimization problems. *Math. Program.*, 129(2):225–253, Oct 2011.

[46] K. R. Parthasarathy. *Probability measures on metric spaces. Reprint of the 1967 original.* Providence, RI: AMS Chelsea Publishing, reprint of the 1967 original edition, 2005.

[47] Georg Ch. Pflug. Stochastic minimization with constant step-size: Asymptotic laws. *SIAM Journal on Control and Optimization*, 24(4):655–666, 1986.

[48] D. Preiss. Dimension of metrics and differentiation of measures. General topology and its relations to modern analysis and algebra V, Proc. 5th Prague Topol. Symp. 1981, Sigma Ser. Pure Math. 3, 565-568 (1983), 1983.

[49] Yu. V. Prokhorov. Convergence of random processes and limit theorems in probability theory. *Teor. Veroyatn. Primen.*, 1:177–238, 1956.

[50] P. Richtárik and M. Takáč. Distributed coordinate descent method for learning with big data. *J. Mach. Learn. Res.*, 17(75):1–25, 2016.

[51] H. E. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[52] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Control. Optim.*, 14:877–898, 1976.

[53] L. Rosasco, S. Villa, and B. C. Vũ. Convergence of stochastic proximal gradient algorithm. *Appl. Math. Opt.*, Oct 2019.

[54] M. V. Solodov and B. F. Svaiter. A hybrid approximate extragradient – proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Analysis*, 7:323–345, 1999.

[55] V. Strassen. The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, 36(2):423–439, 1965.

[56] D.W. Stroock. *Probability Theory: An Analytic View.* Cambridge University Press, 2010.

[57] T. Szarek. Feller processes on nonlocally compact spaces. *Ann. Probab.*, 34(5):1849–1863, 2006.

[58] C. Villani. *Optimal transport: Old and New.*, volume 338. Berlin: Springer, 2009.

[59] J. von Neumann. *Functional Operators, Vol II. The geometry of orthogonal spaces*, volume 22 of *Ann. Math Stud.* Princeton University Press, 1950. Reprint of mimeographed lecture notes first distributed in 1933.

[60] P. Walters. *An introduction to ergodic theory*, volume 79. New York, NY: Springer, paperback edition, 2000.

[61] D. Worm. *Semigroups on Spaces of Measures.* PhD thesis, Universiteit Leiden, 2010.