Random Function Iterations for Stochastic Feasibility Problems

Dissertation

zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades

"Doctor rerum naturalium"

der Georg-August-Universität Göttingen

im Promotionsprogramm Mathematical Sciences

der Georg-August University School of Science (GAUSS)

vorgelegt von

Neal Hermer

aus Elmshorn

Göttingen, 2019

Betreuungsausschuss:

Prof. Dr. Russell Luke
Institut für Numerische und Angewandte Mathematik
Georg-August-Universität Göttingen

Prof. Dr. Anja Sturm
Institut für Mathematische Stochastik
Georg-August-Universität Göttingen

Mitglieder der Prüfungskommission:

Referent:

Prof. Dr. Russell Luke

Korreferentin:

Prof. Dr. Anja Sturm

Weitere Mitglieder der Prüfungskommission:

Prof. Dr. Gerlind Plonka-Hoch

Prof. Dr. Thorsten Hohage

Prof. Dr. Ingo Witt

Jun.-Prof. Dr. Daniel Rudolf

Tag der mündlichen Prüfung: 24.01.2019

Acknowledgements

Ich bedanke mich bei meinen Betreuern D. R. Luke und A. Sturm für die Möglichkeit in diesem Projekt zu forschen. Es ist schön viel Freiheit in der Forschung zu haben und seine Ideen zu verfolgen, aber auch sehr anstrengend, so dass ich froh bin, dass es nun vorbei ist, aber auch froh, dass ich mich dafür entschieden habe. Besonders mein Erstbetreuer Russell hat viele Ideen und vor allem Motivation zu dem Projekt beigesteuert und Anja konnte uns mit ihrer genauen Arbeitsweise auf Stolpersteine und Korrektheit aufmerksam machen. Dann gibt es da noch die Arbeitsgruppe, deren familiäre Atmosphäre den Arbeitsalltag verschönerte. Anna danke ich für die vielen schönen Gespräche über auch manchmal wissenschaftliche Themen, und die immer ein Ohr für Sorgen und Nöte hatte.

Aus dem nicht-wissenschaftlichen Umfeld, gibt es viele Menschen, die zu einer angenehmen Zeit vor und nach der Arbeit beigetragen haben, und denen ich dankbar bin. Einer dieser ist Anastasia, die mir das schönste Geschenk gemacht hat, das man jemandem machen kann.

Contents

List of Figures

Abstract

The aim of this thesis is to develop a theory that describes errors in fixed point iterations stochastically, treating the iterations as a Markov chain and analyzing them for convergence in distribution. These particular Markov chains are also called iterated random functions. The convergence theory for iterated random averaged operators turns out to be simple in $\mathbb{R}^n$: If an invariant measure for the Markov operator exists, the chain converges to an invariant measure, which may depend on the initial distribution. The stochastic fixed point problem is hence to find invariant measures of the Markov operator. We formulate different error models and study whether the corresponding Markov operator possesses an invariant measure; in some cases also rates of convergence w.r.t. metrics on the space of probability measures can be computed (geometric rates).

There occur two major types of convergence. Weak convergence of the distributions of the iterates (or their average) and almost sure convergence. The stochastic fixed point problem can be seen as either consistent or inconsistent stochastic feasibility problem, where almost sure convergence is observed in the former (see [25]) and weak convergence in the latter. The type of convergence turns out to determine the consistency of the problem. We give conditions for which we can expect convergence in the above terms for general assumptions on the underlying metric space, and nonexpansive, paracontractive or averaged mappings.

Since the focus of this thesis is probabilistic, when applied to algorithms for optimization, convergence is in distribution and the fixed points are measures. This perspective is particularly useful when the underlying problem models systems with measurement errors, or even when the problem is deterministic, but the algorithm for its numerical solution is implemented on conventional computers with finite-precision arithmetic.

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

CHAPTER 1

INTRODUCTION

We consider here only one simple algorithm, that captures many other algorithms in its generality. We are not interested in it for numerical purposes, just to determine its behavior when errors enter in every iteration. This algorithm is a stochastic extension of the simple *fixed point iteration*, that is, for an operator $T : G \to G$, where G is a yet arbitrary set, the sequence (x_k), where $x_{k+1} := Tx_k$, $k \in \mathbb{N}$ and $x_0 \in G$. A description of errors entering this iteration is achieved via i.i.d. random variables $(\xi_k)_{k\in\mathbb{N}_0}$ that map from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ into a measurable space $(I, \mathcal{I})$, where I is called the index set. These errors ξ_k model the random selection of a mapping from a fixed family of mappings $\{T_i : G \to G \mid i \in I\}$. Errors are hence implicitly contained in the choice of the family $(T_i)_{i\in I}$. The *stochastic fixed point iteration*, or as we will refer to it in the following, the *random function iteration* (*RFI*), see also [17], is thus

$$X_{k+1} := T_{\xi_k} X_k, \qquad k \in \mathbb{N}, \quad \text{where } X_0 \sim \mu \in \mathscr{P}(G). \tag{1.1}$$

The iterates X_k form a Markov chain of random variables on the space G, which is not yet specified, but will, in the subsequent analysis, become a separable and complete metric space (we refer to it then as a *Polish space*). Since one is working with random variables, the more general initialization of a random variable is now appropriate, i.e. letting X_0 be any random variable, but with a fixed distribution μ in the *space of probability measures* $\mathscr{P}(G)$ on G. Still the deterministic initialization in a point $x_0 \in G$ is possible by choosing a delta distribution $\mu = \delta_{x_0}$. Also, the deterministic fixed point iteration is representable in this setting, by letting $I = \{1\}$ and $T_1 = T$ in the setting of (1.1).

For a Polish space (G, d) many important classical results of probability theory are still true, see for example [21], this includes in particular the theory of convergence in the weak sense of sequences of probability measures and also the concept of tightness and the equivalence of tightness of a sequence of probability measures and the existence of clusterpoints for any subsequence (Prokhorov's Theorem).

Our aim is to study the behavior of the RFI mainly in the case when convex feasibility problems are considered (in $(G, d) = (\mathbb{R}^n, \|\cdot\|)$) and an error concept for the projection on these sets is introduced. The *convex feasibility problem* consists in finding a point

$x \in \mathbb{R}^n$ in the intersection of the convex and closed sets C_j, $j \in J$, where J is an (mostly finite) index set. Many projection algorithms for solving this problem can be expressed as simple fixed point iteration with nonexpansive, even averaged operators, that fit into the framework studied here.

One way to express the influence of errors of the sets, due to measurement or computational errors, and their projectors is to model them as exact projections onto different, slightly perturbed sets. As an example for the convex feasibility problem with only a single set, we consider an affine subspace $C = \{x \in \mathbb{R}^n \mid \langle a, x\rangle = b\}$ with $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$. In this case, an error model could be given by $C_\xi := \{x \in \mathbb{R}^n \mid \langle a + \xi_1, x\rangle = b + \xi_2\}$ for a random variable $\xi = (\xi_1, \xi_2) \in \mathbb{R}^n \times \mathbb{R}$. This would describe the affine subspace C, but with (in general small) distortions in the normal vector $a \to a + \xi_1$ and the displacement $b \to b + \xi_2$. It is clear that the simple fixed point iteration consisting of just $T = P_C$, where $P_C x = \operatorname{argmin}_{x \in C} \|x - c\|$ is the projector onto C, converges after one step to a point in C, while the RFI for this error model behaves totally different. In general the iteration does not converge to a point in $\mathbb{R}^n$, since the subspaces in every iteration change randomly according to the random variables ξ and ζ. But still, as we will show later on, the distributions of the iterates, also denoted by $\mathcal{L}(X_k)$ (the *law* of X_k) or $\mathbb{P}^{X_k}$ converge in the weak sense to a probability measure on $\mathbb{R}^n$.

Modelling errors of sets in the above sense is useful because, as we will show, convergence of the RFI (more precisely for the distributions in the weak sense) follows for projections algorithms as soon as there exists an invariant measure for the Markov operator. So, a well-posed error model should yield existence of an invariant measure. As some examples indicate, it is often not the specific distribution of the error, but more so the actual error model of the underlying set that has a great influence on the existence of invariant measures.

But the framework of the RFI allows also different interpretations of the random variable ξ. Instead of an error of a set, it could just model a random selection of operators $(T_i)_{i \in I}$ (see also [25]). When $|I|$ is large or infinite, any generic deterministic algorithm to solve the feasibility problem could be too slow or not finish a cycle through all indices after finite time. The stochastic choice of indices can help in this case, if ξ would describe a weighting of the choice of the operators T_i.

If $I = J = \{1, \dots, m\}$ and $T_i = P_{C_i}$ is the projector onto a convex set C_i, then the algorithm resembles the stochastic projection algorithm (stochastic variant of cyclic projections). And instead of possible convergence to a unique limit cycle (in the deterministic case), one would have convergence to an invariant measure for the corresponding RFI.

In contrast to the affine subspace example, there are cases, when not only the distributions converge but also the random variables themselves (almost surely). In these cases we speak of a consistent stochastic feasibility problem, otherwise the problem is called inconsistent. The theory of consistent stochastic feasibility problems is very rich and enables us to analyze this problem in some more depth than the inconsistent problem. Also a great difference is the possible analysis even on Hilbert spaces in contrast to the inconsistent problem, where we need to stay in $\mathbb{R}^n$ to be able to get convergence in distribution.

This thesis consists of content of the article [25] and a not yet submitted article, so in particular the consistent stochastic feasibility problem in this thesis can in parts be found in [25]. A second article is in progress, where the content and examples in the thesis of the inconsistent case are coming from (same authors).

CHAPTER 2

PROBABILITY THEORY

In this section we review the fundamental concepts of probability theory that we need throughout this study. These include conditional expectations for nonintegrable random variables and weak convergence. But first the basics.

2.1. PROBABILITY THEORY: BASICS

Probability theory is a powerful tool to describe natural processes, because it reduces the description from many possibly depending variables to just a relative frequency of events that can be observed. For example, rolling a dice has many free parameters like speed, rotation, height, (refer to these as variables in the *phase space*) that influence its motion on the table after it was rolled. Observation of just 6 relative frequencies, one for each side, enables characterization of its behavior for many turns, but not for a single one. So the introduction of a probability distribution is to give a weight to the set of all the points in the phase space that lead to one possible outcome. This reduces the phase space immensely from $\mathbb{R}^p$, where p is number of free parameters to the set $\{1, 2, \dots, 6\}$, but still captures some properties of the dice with the drawback not to be able to predict an outcome of a single experiment.

The *phase space* is denoted by Ω. A measure on Ω, is defined on a family $\mathcal{F}$ of subsets of Ω. To guarantee richness of operations with the interesting events, that can be observed, this family is assumed to be a σ-algebra, that is, $\Omega \in \mathcal{F}$ and for any $A \in \mathcal{F}$ it holds that the complement $A^c := \Omega \setminus A \in \mathcal{F}$ and for any sequence $(A_n) \subset \mathcal{F}$ the union $\bigcup_n A_n \in \mathcal{F}$. A *measure* μ on $(\Omega, \mathcal{F})$ is a function $\mu : \mathcal{F} \to \mathbb{R}_+ \cup \{\infty\}$ satisfying $\mu(\emptyset) = 0$ and $\mu(A) = \sum_{n=1}^{\infty} \mu(A_n)$ for any pairwise disjoint sequence $(A_n) \subset \mathcal{F}$. A *probability measure* μ satisfies additionally $\mu(\Omega) = 1$. A *probability space* is a triple $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is a set, $\mathcal{F}$ a σ-algebra and $\mathbb{P}$ a probability measure.

Of course, the set of <u>all</u> subsets of Ω (the *power set*) is also a σ-algebra, but measures on this σ-algebra do not satisfy rich properties in general (unless Ω is countable), e.g. there exists no Lebesgue-measure on the power set of $\mathbb{R}$, but it exists on the so called Borel-algebra. It is in general enough to deal with sets in the smallest σ-algebra that includes all open sets of a metric space (G, d), these are called Borel sets and the corresponding σ-algebra $\mathcal{B}(G)$ the *Borel-algebra* of G.

Usually there are no further assumptions on the probability space, except that it is rich enough to guarantee existence of random variables with certain distributions. A *random variable* $X : \Omega \to G$, where G is the *state space*, is a *measurable* function, i.e. $X^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{B}(G)$. The *distribution* μ of the random variable X, denoted by $X \sim \mu$, is a probability measure on the state space G given by $\mu := \mathcal{L}(X) := \mathbb{P}^X := \mathbb{P} \circ X^{-1}$. For example, there exists a uniformly distributed random variable - $X \sim \mathrm{U}(0,1)$ - on $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}([0,1]), \lambda)$, where $\mathbb{P} = \lambda = \mathrm{U}(0,1)$ is the Lebesgue measure; simply take $X = \mathrm{Id}$. The next lemma states that we can find a random variable with given distribution under mild assumptions. (Note that Polish spaces – separable and complete metric spaces – are included in the set of Borel spaces, i.e. a space on which there exists a measurable bijection from it to a Borel-set of $\mathbb{R}$.)

Lemma 2.1.1 (existence of r.v. for given distribution)**.** *Let $(S, \mathcal{S})$ be a Borel space, μ a probability measure on S and $\vartheta \sim \mathrm{U}(0,1)$, then there exists a measurable function $f : [0,1] \to S$ such that $f(\vartheta) \sim \mu$.*

Proof. This is a special case of [28, Theorem 2.22]. □

If we choose the probability space in our example rich enough, i.e. it contains at least 6 elements, then one can define a random variable X that describes the experiment through its probabilities that a certain face is up, when casting a dice. Or, when working with the phase space, let $f : \mathbb{R}^p \to \{1, 2, \ldots, 6\}$ be the solution to the physical model that gives the outcome $i \in \{1, 2, \ldots, 6\}$ depending on the current parameter set (i.e. speed, height, angle and so on). We are interested in determining the probability that $f = i$, but this is only possible if we specify the distribution of each parameter. If we choose deterministic initial distributions, i.e. $\mu = \delta_x$ for $x \in \mathbb{R}^p$, we would get that $\mathbb{P}(f = i) = \mathbb{1}_{\{i\}}(f(x))$ for $i = 1, 2, \ldots, 6$. If we choose the parameters independently and uniformly on $[0,1]$, then $\mathbb{P} := \lambda^p$ is the appropriate probability measure on the phase space or parameter space $\mathbb{R}^p$. We have that $f^{-1}(\{i\})$ are all the parameter constellations that lead in an experiment to the outcome "face i is up" and hence $\mathbb{P}(f = i) = \lambda^p(f^{-1}(\{i\})) = \int_{[0,1]^p} \mathbb{1}_{\{i\}}(f(x))\,\mathrm{d}x$.

2.2. Conditional expectation

Conditional expectations are a useful tool to compute expectations of expressions of two dependent variables, for example $\mathbb{E}[f(X,Y)]$ for an integrable $f : G \times G \to \mathbb{R}$. Note that for another couple $(\tilde{X}, \tilde{Y})$ of random variables with the same *marginals*, that is, $\mathcal{L}(X) = \mathcal{L}(\tilde{X})$ and $\mathcal{L}(Y) = \mathcal{L}(\tilde{Y})$, in general one has $\mathbb{E}[f(X,Y)] \neq \mathbb{E}[f(\tilde{X}, \tilde{Y})]$, unless these variables have the same *joint distribution*, i.e. $\mathcal{L}((X,Y)) = \mathcal{L}((\tilde{X}, \tilde{Y}))$. For the case that X and Y are independent – we also write $X \perp\!\!\!\perp Y$ in this case – we have for every couple $(\tilde{X}, \tilde{Y})$ with the same marginals that $\mathbb{E}[f(X,Y)] = \mathbb{E}[f(\tilde{X}, \tilde{Y})]$.

We will say random variables $(X_i)_{\in \in I}$ for an arbitrary index set I are independent, if for any finite selection $J \subset I$ and any $A_j \in \mathcal{B}(G)$ it holds that $\mathbb{P}(X_j \in A_j,\ \forall j \in J) = \prod_{j \in J} \mathbb{P}(X_j \in A_j)$. One has the following fact.

Theorem 2.2.1 (Existence and Independence, Theorem 2.19 in [28])**.** *With the notation of Lemma 2.1.1, let $\xi_1 = f(\vartheta)$. Let T be another Borel space and η a distribution thereon. Then there exists a measurable function $g : [0,1] \to T$ with $\xi_2 := g(\vartheta) \sim \eta$ such that $\xi_1 \perp\!\!\!\perp \xi_2$.*

This generalizes immediately to sequences by induction, so for any probability measures $\mu_1, \mu_2, \dots$ on a Borel spaces $S_1, S_2, \dots$, there exist independent random variables $\xi_1, \xi_2, \dots$ on the probability space $([0,1], \mathcal{B}([0,1]), \lambda)$ with distributions $\mu_1, \mu_2, \dots$ [28, Theorem 2.19]. One also has that arbitrary transformations of independent variables do not destroy this property. Define for random variable $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (G, \mathcal{B}(G))$ the smallest σ-algebra on Ω that makes X measurable by $\sigma(X)$. Then independence of $(X_i)_{i \in I}$ is equivalent to the independence of $(\sigma(X_i))_{i \in I}$, where the latter is defined as follows. For any finite selection $J \subset I$ and any $B_j \in \sigma(X_j)$ it holds that $\mathbb{P}(B_j\ \forall j \in J) = \prod_{j \in J} \mathbb{P}(B_j)$.

Lemma 2.2.2 (Independence after Transformation)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X \perp\!\!\!\perp Y$ two random variables on some measurable spaces $(S_i, \mathcal{S}_i)$, $i = 1, 2$. Let $f : S_1 \to T_1$ and $g : S_2 \to T_2$ be measurable, where $(T_i, \mathcal{T}_i)$, $i = 1, 2$ are measurable spaces, then $f(X) \perp\!\!\!\perp g(Y)$.*

Proof. One has that $X \perp\!\!\!\perp Y$ iff $\sigma(X) \perp\!\!\!\perp \sigma(Y)$ and since $\sigma(f(X)) \subset \sigma(X)$ and analogous for Y, this assertion follows. □

For any two random variables X, Y one can define a nontrivial third random variable out of these, called conditional expectation. This conditional expectation can be imagined as integrating out all independent parts, i.e. if one would have $X = f(Y, \xi)$, where $\xi \perp\!\!\!\perp Y$, then computing the conditional expectation of X given Y is the random variable

$$\mathbb{E}[X \mid Y] := \int f(Y, u) \mathbb{P}^{\xi}(\mathrm{d}u). \tag{2.1}$$

This decomposition of the random variable X is always possible (for a rich enough probability space), but not almost surely, only in distribution, but still the joint distribution of X and Y is not changed, as the following theorem shows.

Theorem 2.2.3 (Decomposition, Theorem 5.10 in [28])**.** *Let X, Y be random elements on Borel spaces S, T respectively, then there exists a measurable function $f : T \times [0,1] \to S$ such that for any $\xi \sim \mathrm{U}(0,1)$ with $\xi \perp\!\!\!\perp Y$ it holds that $\mathcal{L}(X, Y) = \mathcal{L}(f(Y, \xi), Y)$.*

Here $\mathrm{U}(0,1) = \lambda$ is the uniform distribution on $([0,1], \mathcal{B}([0,1]))$, the probability space needs to be large enough for ξ to exist. One can always enlarge a probability space to guarantee the existence of a $\mathrm{U}(0,1)$ distributed random variable by considering $\Omega \times [0,1]$ as underlying state space with σ-algebra $\mathcal{F} \otimes \mathcal{B}([0,1])$ and probability measure $\mathbb{P} \otimes \lambda$. One can ensure the existence of $\xi \perp\!\!\!\perp Y$ by Theorem 2.2.1, if Y was constructed by Lemma 2.1.1. This theorem means that for any variable $\tilde{X} = f(Y, \tilde{\xi})$ $(\tilde{\xi} \perp\!\!\!\perp Y)$ it holds that $\mathbb{E}[g(X,Y)] = \mathbb{E}[g(\tilde{X}, Y)]$, so X and $\tilde{X}$ are indistinguishable under these integrals for any measurable $g : G \times G \to \mathbb{R}$, whenever the integral exists. Since there always exists such a function f satisfying the above decomposition we could interpret Eq. (2.1) as definition of the conditional expectation (uniqueness, i.e. $\mathbb{E}\left[X \,\middle|\, Y\right] = \mathbb{E}\left[\tilde{X} \,\middle|\, Y\right]$ can also be shown). This enables the interpretation of the conditional expectation $\mathbb{E}\left[X \,|\, Y\right]$ as the random variable that remains after integrating out or taking the expectation of the independent part of X from Y. The more usual definition however is via an a.s. unique density as seen in the next theorem. We will in the following stick to that definition, since it is more common. We will only work with conditional expectations on real-valued random variables.

Theorem 2.2.4 (conditional expectation - basics, see Theorem 5.1 in [28])**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X a real-valued random variable with $\mathbb{E}|X| < \infty$ (X is integrable). Let $\mathcal{F}_0 \subset \mathcal{F}$ a sub-σ-algebra. Then there exists an a.s. unique $\mathcal{F}_0$-mb. random variable $Z := \mathbb{E}\left[X \,|\, \mathcal{F}_0\right]$ with $\mathbb{E}(Z\mathbb{1}_A) = \mathbb{E}(X\mathbb{1}_A)$ for all $A \in \mathcal{F}_0$.*
Let $Y, (X_n)_{n\in\mathbb{N}}$ be integrable random variables. Further properties are:

(i) $\mathbb{E}(\mathbb{E}\left[X \,|\, \mathcal{F}_0\right]) = \mathbb{E}X$;

(ii) X *is* $\mathcal{F}_0$*-mb, then* $\mathbb{E}\left[X \,|\, \mathcal{F}_0\right] = X$ *a.s.;*

(iii) X *independent of* $\mathcal{F}_0$*, then* $\mathbb{E}\left[X \,|\, \mathcal{F}_0\right] = \mathbb{E}X$ *a.s.;*

(iv) $\mathbb{E}\left[aX + bY \,|\, \mathcal{F}_0\right] = a\mathbb{E}\left[X \,|\, \mathcal{F}_0\right] + b\mathbb{E}\left[Y \,|\, \mathcal{F}_0\right]$ *a.s. for all* $a, b \in \mathbb{R}$;

(v) $X \leq Y$*, then* $\mathbb{E}\left[X \,|\, \mathcal{F}_0\right] \leq \mathbb{E}\left[Y \,|\, \mathcal{F}_0\right]$ *a.s.;*

(vi) $0 \leq X_n \nearrow X$ *(monotonically non-decreasing), then* $\mathbb{E}\left[X_n \,|\, \mathcal{F}_0\right] \nearrow \mathbb{E}\left[X \,|\, \mathcal{F}_0\right]$ *a.s.;*

(vii) $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}$ *with* σ*-algebra* $\mathcal{F}_1$*, then* $\mathbb{E}\left[\mathbb{E}\left[X \,|\, \mathcal{F}_1\right] \,|\, \mathcal{F}_0\right] = \mathbb{E}\left[X \,|\, \mathcal{F}_0\right]$;

(viii) Y *is* $\mathcal{F}_0$*-mb. and* $\mathbb{E}[|XY|] < \infty$*, then* $\mathbb{E}\left[XY \,|\, \mathcal{F}_0\right] = Y\mathbb{E}\left[X \,|\, \mathcal{F}_0\right]$.

Note that we set $\mathbb{E}\left[X \,|\, Y\right] := \mathbb{E}\left[X \,|\, \sigma(Y)\right]$ with the definition of the conditional expectation from Theorem 2.2.4. One can generalize the definition of the conditional expectation from integrable random variables to random variables X, where just their negative part $X^- := \max(0, -X)$ is integrable. Therefore, we need to convince ourselves that the positive part $X^+ := \max(0, X)$ is well-behaved, and induces a conditional expectation (existence of a density).

Lemma 2.2.5 (Satz 17.11 in [4])**.** *Let $(\Omega, \mathcal{F})$ be a measurable space and μ be σ-finite ((i.e. there exists $(\Omega_n)_{n\in\mathbb{N}} \subset \mathcal{F}$ with $\mu(\Omega_n) < \infty$ and $\bigcup_n \Omega_n = \Omega$)). Let $f : \Omega \to [0, \infty]$ and set $\nu = f \cdot \mu$ (i.e. $\nu(A) = \int_A f \,\mathrm{d}\mu$ for $A \in \mathcal{F}$). Then f is μ-a.s. unique. Furthermore, ν is σ-finite if and only if f is real-valued μ-a.s.*

Remark 2.2.6: A nonnegative real-valued random variable X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ induces a σ-finite measure $\nu = X \cdot \mathbb{P}$. This is clear by letting $\Omega_n := \{X \leq n\}$.

Theorem 2.2.7 (conditional expectation for nonnegative r.v.)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X \geq 0$ be a real-valued random variable (not necessarily integrable). Let $\mathcal{F}_0 \subset \mathcal{F}$ be a sub-σ-algebra. Then there exists an a.s. unique nonnegative real-valued random variable $Z := \mathbb{E}[X \mid \mathcal{F}_0]$ on $(\Omega, \mathcal{F}_0)$ with $\mathbb{E}(Z\mathbb{1}_A) = \mathbb{E}(X\mathbb{1}_A)$ for all $A \in \mathcal{F}_0$.*
Let additionally $Y, (X_n)$ be nonnegative and real-valued, then all items (i) to (vii) in Theorem 2.2.4 are satisfied for these and (viii) even if $\mathbb{E}[XY] = \infty$.

Proof. From Remark 2.2.6 follows the existence of disjoint sets $\Omega_n \in \mathcal{F}_0$ with $\bigcup_n \Omega_n = \Omega$ and the property that $\int_{\Omega_n} X \,\mathrm{d}\mathbb{P} < \infty$. One has that a.s.

$$\mathbb{1}_{\Omega_n}\mathbb{E}[X \mid \mathcal{F}_0 \cap \Omega_n] = \mathbb{E}[X\mathbb{1}_{\Omega_n} \mid \mathcal{F}_0 \cap \Omega_n] = \mathbb{E}[X \mid \mathcal{F}_0 \cap \Omega_n] = \mathbb{E}[X\mathbb{1}_{\Omega_n} \mid \mathcal{F}_0].$$

Define $Z := \sum_n \mathbb{E}[X \mid \mathcal{F}_0 \cap \Omega_n]$, then $Z = \mathbb{E}[X \mid \mathcal{F}_0]$. The items (i) to (viii) follow now from Theorem 2.2.4 on Ω_n and the Monotone Convergence Theorem, see Theorem A.0.13. □

Now we are ready to formulate the results of Theorem 2.2.4 in a more general form, i.e. for nonintegrable random variables.

Theorem 2.2.8 (conditional expectation for r.v. with integrable negative part)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X be a real-valued random variable with $\mathbb{E}[X^-] < \infty$, where $X^- := \max(0, -X)$. Let $\mathcal{F}_0 \subset \mathcal{F}$ be a sub-σ-algebra. Then there exists an a.s. unique real-valued random variable $Z := \mathbb{E}[X \mid \mathcal{F}_0]$ on $(\Omega, \mathcal{F}_0)$ with $\mathbb{E}(Z\mathbb{1}_A) = \mathbb{E}(X\mathbb{1}_A)$ for all $A \in \mathcal{F}_0$.*
Let additionally $Y, (X_n)$ be real-valued with integrable negative part, then all items (i) to (vii) in Theorem 2.2.4 are satisfied for these and (viii) if $\mathbb{E}[(XY)^-] < \infty$.

Proof. Follows immediately from $X = X^+ - X^-$, where $X^+ := \max(0, X)$ and Theorem 2.2.4 and Theorem 2.2.7. □

2.3. Probability kernel, regular conditional distribution

A major tool when working with conditional expectations is the Disintegration Theorem, see Theorem 2.3.2. This is a more general version of Eq. (2.1) and giving conditions when and how to integrate out independent parts of given random variables. Therefore, we will need two more definitions. A *probability kernel* from $(T, \mathcal{T})$ to $(S, \mathcal{S})$ is a function $p : T \times \mathcal{S} \to [0, 1]$ that is measurable in the first argument, i.e. $p(\cdot, A)$ is measurable for all $A \in \mathcal{S}$ and is a probability measure in the second argument, i.e. $p(x, \cdot)$ is a probability measure for all $x \in T$. A *regular conditional distribution* of $\mathbb{P}(X \in \cdot \mid Y) := \mathbb{E}[\mathbb{1}\{X \in \cdot\} \mid Y] := \mathbb{E}[\mathbb{1}\{X \in \cdot\} \mid \sigma(Y)]$ with X, Y in G, S, respectively is a probability kernel $p : S \times \mathcal{G} \to [0, 1]$

with $p(Y, A) = \mathbb{P}(X \in A \,|\, Y)$ a.s. Note that for $(S, \mathcal{S}) = (\Omega, \mathcal{F}_0)$, where $\mathcal{F}_0 \subset \mathcal{F}$ is a sub σ-algebra and $Y = \mathrm{Id}$, the conditional probability $\mathbb{P}(X \in \cdot \,|\, \mathcal{F}_0) := \mathbb{P}(X \in A \,|\, Y)$ is a regular conditional distribution if there exists a probability kernel $p : \Omega \times \mathcal{B}(G) \to [0, 1]$ with $p(\cdot, A) = \mathbb{P}(X \in A \,|\, Y)$ a.s. One has the following existence theorem.

Theorem 2.3.1 (existence of regular conditional distribution, Theorem 5.3 in [28]). *Let $(S, \mathcal{S})$ be a Borel space and $(T, \mathcal{T})$ a measurable space and let X_1, X_2 be random variables in S, T, respectively. Then there exists a $\mathcal{L}(X_2)$-a.s. unique probability kernel μ from T to S satisfying $\mathbb{P}(X_1 \in \cdot \,|\, X_2) = \mu(X_2, \cdot)$ a.s.*

Theorem 2.3.2 (disintegration). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(S, \mathcal{S}), (T, \mathcal{T})$ be measurable spaces. Let X_1, X_2 be two random variables in S, T, respectively and let $\mathcal{F}_0 \subset \mathcal{F}$ be a sub σ-algebra, such that X_2 is $\mathcal{F}_0$ measurable. Let furthermore $f : G \times S \to \mathbb{R}$ be measurable and $\mathbb{E}[f^-(X_1, X_2)] < \infty$. Suppose μ is a regular version of $\mathbb{P}(X_1 \cdot \,|\, \mathcal{F}_0)$, then*

$$\mathbb{E}[f(X_1, X_2) \,|\, \mathcal{F}_0] = \int f(x_1, X_2)\mu(\cdot, \mathrm{d}x_1) \quad \textit{a.s.},$$

where $\mu(\omega, \cdot) = \mathbb{P}(X_1 \in \cdot \,|\, \mathcal{F}_0)(\omega)$ for $\omega \in \Omega$.

Proof. First we note that by [28, Lemma 1.38 (i)] the rhs. is indeed $\mathcal{F}_0$-measurable. In the proof of [28, Theorem 5.4] is shown that

$$\mathbb{E}[g(X_1, X_2)] = \mathbb{E} \int g(x_1, X_2)\mu(\cdot, \mathrm{d}x_1) \tag{2.2}$$

for all measurable $g \geq 0$. If we now replace X_2 with the $\mathcal{F}_0$-measurable random variable $(X_2, \mathbb{1}_A) \in G \times \{0, 1\}$ with $A \in \mathcal{F}_0$, and let $g(X_1, (X_2, \mathbb{1}_A)) := f(X_1, X_2)\mathbb{1}_A$ the statement follows for $f \geq 0$. By uniqueness of $\mathbb{E}[f(X_1, X_2) \,|\, \mathcal{F}_0]$ (Theorem 2.2.7) linearity it also holds for measurable functions with $\mathbb{E}[f^-(X_1, X_2)] < \infty$. □

Remark 2.3.3 (disintegration for independent variables)**:** If $\mathcal{F}_0 = \sigma(X_2)$ and $X_1 \perp\!\!\!\perp X_2$, then $\mathbb{P}(X_1 \in \cdot \,|\, X_2) = \mathcal{L}(X_1)$ a.s. and

$$\mathbb{E}[f(X_1, X_2) \,|\, X_2] = \int f(x_1, X_2)\mathbb{P}^{X_1}(\mathrm{d}x_1) \qquad \text{a.s.}$$

2.4. Support of a measure

Theorem 2.4.1 (support of a measure). *Let (G, d) be a Polish space and $\mathcal{B}(G)$ its Borel σ-algebra. Let π be a measure on $(G, \mathcal{B}(G))$ and define its support via*

$$\operatorname{supp} \pi = \{x \in G \,|\, \pi(\mathbb{B}(x, \epsilon)) > 0 \; \forall \epsilon > 0\}.$$

Then the following hold

(i) $\operatorname{supp} \pi \neq \emptyset$, *if* $\pi \neq 0$.

(ii) $\operatorname{supp} \pi$ *is closed.*

(iii) $\pi(A) = \pi(A \cap \operatorname{supp} \pi)$ *for all* $A \in \mathcal{B}(G)$*, i.e.* $\pi((\operatorname{supp} \pi)^c) = 0$.

(iv) For closed $S \subset G$ *with* $\pi(A \cap S) = \pi(A)$ *for all* $A \in \mathcal{B}(G)$ *it holds that* $\operatorname{supp} \pi \subset S$.

(v) Let $\pi(G) < \infty$*. For closed* $S \subset G$ *with* $\pi(S) = \pi(G)$ *it holds that* $\operatorname{supp} \pi \subset S$.

Proof. (i) If $\pi(G) > 0$, then due to separability one could find for any $\epsilon_1 > 0$ a countable cover of G with balls with radius ϵ_1, where at least one needs to have nonzero measure, because $0 < \pi(G) \leq \sum_n \pi(\mathbb{B}(x_n, \epsilon))$. Now just consider $B_1 := \mathbb{B}(x_N, \epsilon_1)$ such that $\pi(B_1) > 0$ and apply the above procedure of countable covers with $\epsilon_2 < \epsilon_1$ iteratively, then there is a sequence $\epsilon_n \to 0$ and $\mathbb{B}(x_{n+1}, \epsilon_{n+1}) \subset \mathbb{B}(x_n, \epsilon_n)$, such that $x_n \to x$, i.e. $x \in \operatorname{supp} \pi$.

(ii) Let $(x_n)_{n\in\mathbb{N}} \subset \operatorname{supp} \pi$ with $x_n \to x$ as $n \to \infty$. Let $\epsilon > 0$ and $N > 0$ such that $d(x_n, x) < \epsilon$ for all $n \geq N$. Then $x_n \in \mathbb{B}(x, \epsilon)$ and $\exists \tilde{\epsilon} > 0$ with $\mathbb{B}(x_n, \tilde{\epsilon}) \subset \mathbb{B}(x, \epsilon)$, so we get

$$\pi(\mathbb{B}(x, \epsilon)) \geq \pi(\mathbb{B}(x_n, \tilde{\epsilon})) > 0,$$

i.e. $x \in \operatorname{supp} \pi$.

(iii) Write $S = (\operatorname{supp} \pi)^c$. Choose $\{x_n\}_{n\in\mathbb{N}} \subset S$ dense. By openness of S there exists $\epsilon_n > 0$ with $\mathbb{B}(x_n, \epsilon_n) \subset S$, hence $S = \bigcup_{n\in\mathbb{N}} \mathbb{B}(x_n, \epsilon_n)$ and

$$\pi(S) \leq \sum_{i\in\mathbb{N}} \pi(\mathbb{B}(x_n, \epsilon_n)) = 0.$$

(It holds $\pi(\mathbb{B}(x_n, \epsilon_n)) = 0$, because otherwise, one could find for any small enough $\epsilon > 0$ a countable cover of $\mathbb{B}(x_n, \epsilon_n)$ with balls with radius ϵ, where at least one needs to have nonzero measure. Since this holds for all ϵ, there is a contradiction to $\mathbb{B}(x_n, \epsilon_n) \subset S$.)

(iv) Let $x \in \operatorname{supp} \pi$. So $\pi(\mathbb{B}(x, \epsilon) \cap S) > 0$ for all $\epsilon > 0$, i.e. $\mathbb{B}(x, \epsilon) \cap S \neq \emptyset$ for all $\epsilon > 0$. Let x_n be such that $x_n \in \mathbb{B}(x, \epsilon_n) \cap S$, where $\epsilon_n \to 0$ as $n \to \infty$. Then by closedness of S, $x_n \to x \in S$.

(v) We have that $S = G \setminus N$ with $N \subset G$ and $\pi(N) = 0$. For any $A \in \mathcal{B}(G)$ it holds that $\pi(A \cap S) = \pi(A) - \pi(A \cap N) = \pi(A)$. The assertion follows from (iv).

□

From Theorem 2.4.1 (v) it follows that the support of a probability measure μ on G can equivalently be defined as the smallest closed set $S \subset G$, for which $\mu(S) = 1$. The next Lemma shows the connection between a random variable and the support of its law.

Lemma 2.4.2 (support of random variable)**.** *Let* $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (G, \mathcal{B}(G))$ *be a random variable, and* G *a Polish space. Then*

$$\operatorname{supp} \mathcal{L}(X) = \bigcap_{\mathbb{P}(N)=0} \overline{X(\Omega \setminus N)},$$

where $\mathcal{L}(X) = \mathbb{P}(X \in \cdot)$ *is the distribution (law) of* X*. In particular, if* $X(\Omega \setminus N) \subset \operatorname{supp} \mathcal{L}(X)$ *for a nullset* $N \subset \Omega$*, then* $\operatorname{supp} \mathcal{L}(X) = \overline{X(\Omega \setminus N)}$.

Proof. First let $x \in \overline{X(\Omega \setminus N)}$ for all $\mathbb{P}$-nullsets $N \subset \Omega$, i.e. there exists a sequence $(\omega_n^N)_{n\in\mathbb{N}} \subset \Omega \setminus N$ with $X(\omega_n^N) \to x$ as $n \to \infty$. Would hold $\mathbb{P}(N_\epsilon) = 0$, where $N_\epsilon := X^{-1}\mathbb{B}(x,\epsilon)$ for some $\epsilon > 0$, then $\overline{X(\Omega \setminus N_\epsilon)} \subset G \setminus \mathbb{B}(x,\epsilon)$, i.e. a contradiction to the existence of a convergent sequence in $\overline{X(\Omega \setminus N_\epsilon)}$ to x. So $\mathbb{P}(N_\epsilon) > 0$ for all $\epsilon > 0$, i.e. $x \in \operatorname{supp} \mathcal{L}(X)$.
Let now $x \in \operatorname{supp} \mathcal{L}(X)$, then $\mathbb{P}^X(\mathbb{B}(x,\epsilon)) > 0$ for all $\epsilon > 0$, i.e. for any $\mathbb{P}$-nullset $N \subset \Omega$ holds $\{\omega \in \Omega \setminus N \,|\, X(\omega) \in \mathbb{B}(x,\epsilon)\} \neq \emptyset$. So one can find a sequence $(\omega_n^N)_{n\in\mathbb{N}} \subset \Omega \setminus N$ with $X(\omega_n^N) \to x$ as $n \to \infty$, so $x \in \overline{X(\Omega \setminus N)}$ for all nullsets N. □

2.5. Weak convergence, its metrization and tightness

A nice source and consistent summary on weak convergence on metric spaces are the lecture notes in [21]. These are based on the books [50, Chapter 9], [43, Chapter II], [8, Chapter 1] that give detailed and further results on this and other topics. Let (G,d) be a Polish space with induced Borel-σ-algebra $\mathcal{B}(G)$. A sequence (μ_n) of probability measures on G is said to converge to $\mu \in \mathscr{P}(G)$ (in the weak sense) if for any $f \in C_b(G)$ (i.e. continuous and bounded function $f : G \to \mathbb{R}$) it holds that

$$\mu_n f = \int f(y)\,\mathrm{d}\mu_n \to \int f(y)\,\mathrm{d}\mu = \mu f \qquad \text{as } n \to \infty.$$

One has the following useful characterizations of weak convergence. Recall, that $f : G \to \mathbb{R}$ is lower semi-continuous (l.s.c.) if $\liminf_{x\to x_0} f(x) \geq f(x_0)$ for all $x_0 \in G$ and upper semi-continuous (u.s.c.) if $-f$ is l.s.c. Recall also, a sequence (ν_n) of probability measures is called *tight*, if for any $\epsilon > 0$ there exists a compact $K \subset G$ with $\nu_n(K) > 1 - \epsilon$ for all $n \in \mathbb{N}$. At last recall, $\operatorname{cl} A := \overline{A}$ is the closure of A, i.e. the set of all clusterpoints of any sequence in A and $\operatorname{int} A$ the interior, i.e. the set of points in A, such that there exists a ball centered around it which is contained in A.

Theorem 2.5.1 (Portmanteau). *Let $(\mu_n) \subset \mathscr{P}(G)$ and $\mu \in \mathscr{P}(G)$. The following are equivalent*

(i) $\mu_n \to \mu$ *as* $n \to \infty$ *in the weak sense.*

(ii) $\mu_n f \to \mu f$ *for all* $f \in C_b(G)$.

(iii) $\mu_n f \to \mu f$ *for all bounded and uniformly continuous* $f : G \to \mathbb{R}$.

(iv) $\mu_n f \to \mu f$ *for all bounded and Lipschitz continuous* $f : G \to \mathbb{R}$.

(v) $\limsup_n \mu_n f \leq \mu f$ *for all u.s.c.* $f : G \to \mathbb{R}$ *that are bounded from above.*

(vi) $\liminf_n \mu_n f \geq \mu f$ *for all l.s.c.* $f : G \to \mathbb{R}$ *that are bounded from below.*

(vii) $\limsup_n \mu_n(B) \leq \mu(B)$ *for all closed* $B \in \mathcal{B}(G)$.

(viii) $\liminf_n \mu_n(U) \geq \mu(U)$ *for all open* $U \in \mathcal{B}(G)$.

(ix) $\mu_n(A) \to \mu(A)$ *for all* $A \in \mathcal{B}(G)$ *with* $\mu(\operatorname{cl} A \setminus \operatorname{int} A) = 0$.

(x) $\mu_n f \to \mu f$ *for all bdd. and mb.* $f : G \to \mathbb{R}$ *with* $\mu(\{x \mid f \text{ is continuous at } x\}) = 1$.

(xi) (μ_n) *is tight and every convergent subsequence has the same limit* μ.

(xii) $d_P(\mu_n, \mu) \to 0$, *where* d_P *is defined in Theorem 2.5.4.*

(xiii) $d_0(\mu_n, \mu) \to 0$, *where* d_0 *is defined in Theorem 2.5.5.*

Furthermore, the weak limit μ *is unique.*

Proof. The last three items are proved below in separate theorems, see Theorems 2.5.3, 2.5.4 and 2.5.5. All of the other points can be found in [43, Theorem 6.1] and [50, Theorem 9.1.5], except item (iv). Since any bounded Lipschitz function is contained in $C_b(G)$, to finish the proof, we just need to show that (iv) implies (viii). Given an open set $U \in \mathcal{B}(G)$, we define a sequence of bounded Lipschitz continuous functions $f_m = \min(1, m\, d(x, U^c))$, $m \in \mathbb{N}$ and note that $0 \le f_m \uparrow \mathbb{1}_U$, since U is open, and hence

$$\liminf_n \mu_n(U) \ge \liminf_n \mu_n f_m = \mu f_m \uparrow \mu(U)$$

by the Monotone Convergence Theorem. This proves (viii). For uniqueness of the weak limit, note that if two limits μ, ν would exist, then we get that

$$\mu(U) \uparrow \mu(f_m) = \nu(f_m) \uparrow \nu(U)$$

with the Monotone Convergence Theorem. This holds for all open $U \in \mathcal{B}(G)$ and hence equality $\nu = \mu$ follows from Theorem A.0.18. (In particular we also get that two probability measure are equal, if $\mu(f) = \nu(f)$ for all $f \in C_b(G)$ that are Lipschitz continuous). $\square$

Remark 2.5.2: Weak convergence of probability measures is in functional analysis also referred to as weak-$*$ convergence of corresponding functionals on $C_b(G)$. To see that, consider the space of probability measures as a subset of linear functionals on the Banach space $(C_b(G), \|\cdot\|_\infty)$ of continuous and bounded functions $f : G \to \mathbb{R}$ with the supremum norm. Every probability measure ν induces a functional Φ_ν on $C_b(G)$ through $\Phi_\nu(f) := \langle \nu, f \rangle := \int_G f(x)\nu(\mathrm{d}x)$. Weak convergence of the probability measures $\nu_n \to \nu$ can then be understood as weak-$*$ convergence of Φ_{ν_n} to Φ_ν, i.e. $\langle \nu_n, f \rangle \to \langle \nu, f \rangle$ as $n \to \infty$ for all $f \in C_b(G)$.

We turn our attention to the last three items of Theorem 2.5.1. For item (xi) we need the following concept of compactness in the space of probability measures.

Theorem 2.5.3 (Prokhorov's Theorem)**.** *Let* (G, d) *be a Polish space and* $(\nu_n) \subset \mathscr{P}(G)$. *Then* (ν_n) *is tight, if and only if* (ν_n) *is weakly compact in* $\mathscr{P}(G)$, *i.e. any subsequence of* (ν_n) *has a convergent subsequence in the weak sense.*

Proof. See [8, Theorem 5.1, Theorem 5.2]. $\square$

Note that with help of Theorem A.0.16 we get immediately the assertion (xi) in Theorem 2.5.1. There are further characterizations of weak convergence. The following characterizations are based on viewing the space of probability measures equipped with certain metrics as metric space, where convergence with respect to the metric is equivalent to weak convergence of the measures.

Theorem 2.5.4 (properties of the Prokhorov-Levi distance)**.** *Let G be a Polish space. Define for $\mu, \nu \in \mathscr{P}(G)$ the Prokhorov-Levi distance*

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

(i) It holds the representation

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

where $C(\mu,\nu) := \{\gamma \in \mathscr{P}(G\times G) \,|\, \gamma(\cdot \times G) = \mu, \quad \gamma(G\times \cdot) = \nu\}$ is called the set of couplings *for μ and ν. 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 weak convergence, i.e. for $\mu_n, \mu \in \mathscr{P}(G)$, $n \in \mathbb{N}$ holds $\mu_n \to \mu$ if and only if $d_P(\mu_n,\mu) \to 0$ as $n \to \infty$.

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

(v) For $\mu_i, \nu_i \in \mathscr{P}(G)$ and $\lambda_i \in [0,1]$, $i = 1,\dots,m$ with $\sum_{i=1}^m \lambda_i = 1$ holds

$$d_P(\sum_i \lambda_i\mu_i, \sum_i \lambda_i\nu_i) \le \max_i d_P(\mu_i,\nu_i).$$

Proof. (i) See [49, Corollary] for the first assertion. To see that the infimum is attained, let $\gamma_n \in C(\mu,\nu)$ be a minimizing sequence, i.e. for $(X_n, Y_n) \sim \gamma_n$ holds $\mathbb{P}(d(X_n,Y_n) > \epsilon) = \gamma_n(U_\epsilon) \to \inf_{(X,Y)\in C(\mu,\nu)} \mathbb{P}(d(X,Y) > \epsilon)$, where $U_\epsilon := \{(x,y) \,|\, d(x,y) > \epsilon\} \subset G\times G$ is open. The sequence (γ_n) is tight and for a clusterpoint γ holds $\gamma \in C(\mu,\nu)$ by Lemma 2.6.3. From Theorem 2.5.1 (viii) it follows that $\gamma(U_\epsilon) \le \liminf_k \gamma_{n_k}(U_\epsilon)$. To see, that the outer infimum is attained, let (ϵ_n) be a minimizing sequence, chosen to be monotonically nonincreasing with limit $\epsilon \ge 0$. One has that $U_\epsilon = \bigcup_n U_{\epsilon_n}$ where $U_{\epsilon_n} \supset U_{\epsilon_{n+1}}$ and hence $\gamma(U_\epsilon) = \lim_n \gamma(U_{\epsilon_n}) \le \lim_n \epsilon_n = \epsilon$.

(ii) Clear by (i).

(iii) See [50, Theorem 9.1.11].

(iv) See [50, Theorem 9.1.11].

(v) If $\epsilon > 0$ is such that $\mu_i(A) \le \nu_i(\mathbb{B}(A,\epsilon)) + \epsilon$ and $\nu_i(A) \le \mu_i(\mathbb{B}(A,\epsilon)) + \epsilon$ for all $i = 1,\dots,m$ and all $A \in \mathcal{B}(G)$, then also $\sum_i \lambda_i\mu_i(A) \le \sum_i \lambda_i\nu_i(\mathbb{B}(A,\epsilon)) + \epsilon$ as well as $\sum_i \lambda_i\nu_i(A) \le \sum_i \lambda_i\mu_i(\mathbb{B}(A,\epsilon)) + \epsilon$.

□

Another metric that metrizes weak convergence is the Kantorovich-Rubinshtein or Fortet-Mourier metric.

Theorem 2.5.5 (Kantorovich-Rubinshtein or Fortet-Mourier metric)**.** *Let G be a Polish space. Define for $\mu, \nu \in \mathscr{P}(G)$ the Kantorovich-Rubinshtein or Fortet-Mourier metric*

$$d_0(\mu, \nu) = \sup \{\mu f - \nu f \,|\, f \in \mathrm{Lip}_1(G),\ \|f\|_\infty \leq 1\},$$

where $\mathrm{Lip}_1(G) := \{f : G \to \mathbb{R} \,|\, |f(x) - f(y)| \leq d(x, y)\, \forall x, y \in G\}$. Then d_0 metrizes weak convergence, i.e. for $\mu_n, \mu \in \mathscr{P}(G)$, $n \in \mathbb{N}$ it holds that $\mu_n \to \mu$ if and only if $d_0(\mu_n, \mu) \to 0$ as $n \to \infty$. Furthermore, $(\mathscr{P}(G), d_0)$ is a Polish space.

Proof. See [10, Section 8.3]. □

2.6. Measures on the product space, couplings

The product space is needed in the description of metrics on the space of probability measures. We will give properties of couplings. For a metric space (G, d) we can define a product space $(G \times G, d_\times)$, which is also a metric space, via any metric $d_\times : G^2 \times G^2 \to \mathbb{R}_+$ that satisfies

$$d_\times \left(\left(\begin{smallmatrix} x_n \\ y_n \end{smallmatrix}\right), \left(\begin{smallmatrix} x \\ y \end{smallmatrix}\right) \right) \to 0 \qquad \Leftrightarrow \qquad d(x_n, x) \to 0 \quad \text{and} \quad d(y_n, y) \to 0. \tag{2.3}$$

Examples would be

$$d_\times \left(\left(\begin{smallmatrix} x_1 \\ y_1 \end{smallmatrix}\right), \left(\begin{smallmatrix} x_2 \\ y_2 \end{smallmatrix}\right) \right) = \max(d(x_1, x_2), d(y_1, y_2)) \tag{2.4}$$

$$d_\times \left(\left(\begin{smallmatrix} x_1 \\ y_1 \end{smallmatrix}\right), \left(\begin{smallmatrix} x_2 \\ y_2 \end{smallmatrix}\right) \right) = \left(d^p(x_1, x_2) + d^p(y_1, y_2)\right)^{\frac{1}{p}}, \qquad p \geq 1. \tag{2.5}$$

This product space satisfies a desirable property as the next lemma shows.

Lemma 2.6.1. *Let (G, d) be a Polish space and let the metric $d_\times$ on $G \times G$ satisfy Eq. (2.3), 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 Eq. (2.3). 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 that the rhs. is always contained in the lhs. For the other direction, note that any metric $d_\times$ with the property (2.3) 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 (2.3), then for $(a_n, b_n) \in A$ with $(a_n, b_n) \to (a, b) \in G \times G$ w.r.t. $\tilde{d}_\times$ it holds that $d(a_n, a) \to 0$ and $d(b_n, b) \to 0$ and hence $d_\times((a_n, b_n), (a, b)) \to 0$, 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 Eq. (2.3). Furthermore separability of $G \times G$ yields that any open set is the countable union of balls: by Theorem A.0.20 there exists $(u_n)_{n \in \mathbb{N}} \subset U$ dense for $U \subset G \times G$ open. We can find $\epsilon_n > 0$ with $\bigcup_n \mathbb{B}(u_n, \epsilon_n) \subset U$. 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 put $\epsilon_m = \epsilon/2$ and 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 Eq. (2.4). Then for any open $U \subset G \times G$ there exist $(u_n) \subset U$ and $\epsilon_n > 0$ with $U = \bigcup_n \mathbb{B}(u_n, \epsilon_n)$ and since

$$\mathbb{B}(u_n, \epsilon_n) = \mathbb{B}(u_{n,1}, \epsilon_n) \times \mathbb{B}(u_{n,2}, \epsilon_n) \in \mathcal{B}(G) \otimes \mathcal{B}(G),$$

for $u_n = (u_{n,1}, u_{n,2}) \in G \times G$ we also get that the lhs. is contained in the rhs, so equality of the σ-algebras follows. □

As we have seen in the proof above the advantage is that we can equip $G \times G$ with the metric in Eq. (2.4), so that balls have a simple structure, that will be helpful as well in the next lemma.

We call a pair of random variables (X, Y) with $X \sim \mu$ and $Y \sim \nu$ a *coupling* of μ and ν. We define for given probability measures μ, ν on G

$$C(\mu, \nu) := \{\gamma \in \mathscr{P}(G \times G) \,|\, \gamma(\cdot \times G) = \mu, \quad \gamma(G \times \cdot) = \nu\},$$

by abuse of language, we also call this the set of *couplings* for μ and ν. We have the following properties of couplings.

Lemma 2.6.2 (couplings)**.** *Let (G, d) be a Polish space and let $\mu, \nu \in \mathscr{P}(G)$. Let $\gamma \in C(\mu, \nu)$, 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 Eq. (2.4).

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

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

Analogous follows $\nu(\mathbb{B}(y, \epsilon)) > 0$. 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 $\epsilon > 0$. By Theorem 2.4.1 there either exists $y \in G$ with $(x, y) \in \operatorname{supp} \gamma$ or there exists a sequence $(x_n, y_n) \in \operatorname{supp} \gamma$ with $x_n \to x$. Hence the assertion follows.

□

As a last point, we want to give a result on tightness of couplings and their clusterpoints.

Lemma 2.6.3 (weak convergence in product space)**.** *Let (G, d) be a Polish space and suppose $(\mu_n), (\nu_n) \subset \mathscr{P}(G)$ are tight sequences. Let $X_n \sim \mu_n$ and $Y_n \sim \nu_n$ and denote by $\gamma_n = \mathcal{L}((X_n, Y_n))$ the joint law of X_n and Y_n. Then (γ_n) is tight.*
If furthermore, $\mu_n \to \mu \in \mathscr{P}(G)$ and $\nu_n \to \nu \in \mathscr{P}(G)$ in the weak sense, then clusterpoints of (γ_n) are in $C(\mu, \nu)$.

Proof. For the proof idea see [47, Theorem 3.12]. By tightness of (μ_n) and (ν_n), there exists for any $\epsilon > 0$ a compact set $K \subset G$ with $\mu_n(G \setminus K) < \epsilon/2$ and $\nu_n(G \setminus K) < \epsilon/2$ for all $n \in \mathbb{N}$, so also

$$\begin{aligned}\gamma_n(G \times G \setminus K \times K) &\leq \gamma_n((G \setminus K) \times G) + \gamma_n(G \times (G \setminus K)) \\ &= \mu_n(G \setminus K) + \nu_n(G \setminus K) \\ &< \epsilon\end{aligned}$$

for all $n \in \mathbb{N}$, implying tightness of (γ_n). By Prokhorov's Theorem, every subsequence of (γ_n) has a convergent subsequence and so there exists $(n_k) \subset \mathbb{N}$ and $\gamma \in \mathscr{P}(G \times G)$ with $\gamma_{n_k} \to \gamma$. One has $\gamma \in C(\mu, \nu)$: Since for every $f \in C_b(G \times G)$ holds $\gamma_{n_k} f \to \gamma f$, we may choose $f(x, y) = g(x) \mathbb{1}_G(y)$ with $g \in C_b(G)$. One has $\mu_{n_k} g \to \mu g$ and

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

as $k \to \infty$. Since $\mu_{n_k} g = \gamma_{n_k} f$ for all k one has the equality $\mu = \gamma(\cdot \times G)$. So similarly $\nu = \gamma(G \times \cdot)$ and hence $\gamma \in C(\mu, \nu)$. □

2.7. Markov chains, Random Function Iterations, Markov operator

Recall the definition of a time-homogeneous Markov chain with transition kernel p.

Definition 2.7.1. A sequence of random variables $(X_k)_{k \geq 0}$, $X_k : (\Omega, \mathcal{F}, \mathbb{P}) \to (G, \mathcal{B}(G))$ is called Markov chain with transition kernel p if for all $k \in \mathbb{N}_0$ and $A \in \mathcal{B}(G)$ $\mathbb{P}$-a.s. the following hold:

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

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

From Theorem 2.3.1 follows the existence of regular versions for $\mathbb{P}(X_{k+1} \in \cdot \mid X_k)$. One has the following fact.

Proposition 2.7.2 (existence of update function, Markov chain property, Proposition 7.6 in [28])**.** *The sequence $(X_k)_{k \geq 0}$ of random variables on a Borel space G is a time-homogeneous Markov chain if and only if there exist a measurable function $\Phi : G \times [0,1] \to G$, called* update function*, and a $\mathrm{U}(0,1)$ i.i.d. sequence $(\xi_k) \perp\!\!\!\perp X_0$ with $X_{k+1} = \Phi(X_k, \xi_k)$ for $k \in \mathbb{N}_0$.*

That ξ_k is $\mathrm{U}(0,1)$ distributed is not necessary for (X_k) to have the Markov property, only the independence of ξ_k and $X_0, \ldots, X_k$ is important, as the next proposition shows.

Proposition 2.7.3 (Markov chain via update function)**.** *Let $(I, \mathcal{I})$ be a measurable space. If an i.i.d. sequence (ξ_k) of I-valued random variables is given and a measurable function $\Phi : G \times I \to G$, then $X_{k+1} = \Phi(X_k, \xi_k)$ defines a Markov chain, if $\xi_k \perp\!\!\!\perp X_0$ for all $k \in \mathbb{N}$.*

Proof. The assertion is an easy consequence of Theorem 2.3.2. Let $f : G \times I \to \mathbb{R}$, $f(x, y) := \mathbb{1}_A(\Phi(x, y))$, where $A \in \mathcal{B}(G)$, $\mathcal{F}_0 = \sigma(X_0, X_1, \dots, X_k)$, $Y = \xi_k$ and $X = X_k$. Then by construction $\mathcal{F}_0$ and Y are independent, X is $\mathcal{F}_0$ measurable, and hence $\mathbb{P}(Y \in \cdot \,|\, \mathcal{F}_0) = \mathcal{L}(Y)$ has a regular version. Let $g(x) := \int f(x, y)\mathbb{P}^Y(\mathrm{d}y)$. Then, since $\mathbb{E}|f(X, Y)| \leq 1 < \infty$, Theorem 2.3.2 yields

$$\mathbb{E}\left[f(X, Y) \,|\, \mathcal{F}_0\right] = g(X),$$

and hence

$$\mathbb{P}\left(X_{k+1} \in A \,|\, X_0, \dots, X_k\right) = g(X_k).$$

By the same argument with $\mathcal{F}_0 = \sigma(X_k)$ one has that

$$\mathbb{P}\left(X_{k+1} \in A \,|\, X_k\right) = g(X_k),$$

hence g is the transition kernel of the Markov chain $(X_k)_{k \in \mathbb{N}_0}$. Note that $g(y) = \mathbb{E}(f(X, y)) = \mathbb{P}(\Phi(X, y) \in A)$, hence $p(\cdot, A) := g(\cdot)$ for A fixed is the transition kernel, as claimed. □

Let us write also $T_y x := \Phi(x, y)$, so we implicitly introduce a family of mappings

$$\{T_y : G \to G \,|\, y \in [0, 1]\},$$

then the transition kernel p of a Markov chain with update function Φ is given by

$$(x \in G)(A \in \mathcal{B}(G)) \qquad p(x, A) := \mathbb{P}(\Phi(x, \xi) \in A) = \mathbb{P}(T_\xi x \in A) \tag{2.6}$$

as follows from Proposition 2.7.3. So we can write the Markov chain formally as the following algorithm.

Algorithm 1 Random Function Iteration (RFI)

Initialization: $X_0 \sim \mu$, (ξ_k) i.i.d., $X_0 \perp\!\!\!\perp (\xi_k)$
 for $k = 0, 1, 2, \dots$ **do**
  $X_{k+1} = T_{\xi_k} X_k$
 return $\{X_k\}_{k \in \mathbb{N}}$

Thus the RFI appears naturally in the study of convergence of Markov chains. We will use the notation $X_k^{X_0} := T_{\xi_{k-1}} \dots T_{\xi_0} X_0$ to denote the RFI sequence with update function $T_y x = \Phi(x, y)$ initialized with $X_0 \sim \mu$. This is particularly helpful when characterizing RFI sequences initialized with the delta distribution of a point $x \in G$, where X_k^x denotes this sequence initialized with $X_0 \sim \delta_x$. We throughout assume that the variables $X_0, (\xi_k)$ are constructed by Theorem 2.2.1, so that we can always construct more independent or dependent variables in the later analysis.

The *Markov operator* $\mathcal{P}$ acting on a measure $\mu \in \mathscr{P}(G)$ is defined via

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

With this notation, one is able to write for the distribution of the k-th iterate of the Markov chain with Markov operator $\mathcal{P}$ that $\mathcal{L}(X_k) = \mathbb{P}^{X_k} = \mu\mathcal{P}^k$ (see [28, Proposition 7.2]).

One defines the operation of the Markov operator acting on a measurable function $f : G \to \mathbb{R}$ via

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

Note that

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

The symbol $\mathcal{P}$ is used ambiguously, sometimes authors use the symbol $\mathcal{P}^*$ to mean the operator on $\mathscr{P}(G)$ with the property that $\mathcal{P}^*\mu = \mu\mathcal{P}$, but it will always be clear from the context, which operator is meant in a formula.

An important regularity property for the Markov operator is the *Feller property*, i.e. $\mathcal{P}f \in C_b(G)$ whenever $f \in C_b(G)$, it is needed in Theorem 2.8.2 to ensure we can construct invariant measures for $\mathcal{P}$ with help of the average of the distributions of the Markov chain iterates.

Theorem 2.7.4 (Feller property for continuous mappings)**.** *If T_i is continuous for every $i \in I$, then the Markov operator $\mathcal{P}$ is Feller.*

Proof. See also [23, Theorem 4.22]. By continuity of T_i, $i \in I$, the update function Φ is continuous in the first argument. It follows for $f \in C_b(G)$ and $x_n \to x$ as $n \to \infty$ by Lebesgue's Dominated Convergence Theorem

$$\mathcal{P}f(x_n) = \int_I f(\Phi(x_n, u))\mathbb{P}^\xi(\mathrm{d}u) \to \int_I f(\Phi(x, u))\mathbb{P}^\xi(\mathrm{d}u) = \mathcal{P}f(x).$$

Note that $\mathcal{P}f$ is bounded, whenever f is a bounded function. $\square$

2.8. INVARIANT MEASURE

A fixed point of the Markov operator $\mathcal{P}$ is called an *invariant distribution*, i.e. $\pi \in \mathscr{P}(G)$ is invariant if and only if $\pi\mathcal{P} = \pi$. We denote the set of all invariant distributions by $\operatorname{inv}\mathcal{P}$. In terms of corresponding random variables, we have the following lemma.

Lemma 2.8.1. *Let π be an invariant probability measure for $\mathcal{P}$ defined in Eq. (2.6) and let $Y \sim \pi$. Suppose ξ is independent of Y, then $T_\xi Y \sim \pi$.*

Proof. For $A \in \mathcal{B}(G)$

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

where we used Theorem 2.3.2 and Fubini's Theorem. □

More generally, if we choose Y independent of (ξ_k), also denoted $Y \perp\!\!\!\perp (\xi_k)$, we get that $X_k^Y \sim \pi$ for all $k \in \mathbb{N}$ with the notation from above. A fundamental result and central in our analysis of the inconsistent feasibility problem is the next theorem.

Theorem 2.8.2 (construction of an invariant measure)**.** *Let $\mu \in \mathscr{P}(G)$ and $\mathcal{P}$ be the Markov operator for a given transition kernel p, which is assumed to be Feller. Let $(\mu\mathcal{P}^n)_{n\in\mathbb{N}}$ be a tight family of probability measures on a Polish space (G,d), i.e. for any $\epsilon > 0$ there exists $K_\epsilon \subset G$ compact with $(\mu\mathcal{P}^n)(G \setminus K_\epsilon) < \epsilon$ for all $n \in \mathbb{N}$. Then any clusterpoint of (ν_n) where $\nu_n = \frac{1}{n}\sum_{i=1}^n \mu\mathcal{P}^i$ is an invariant measure for $\mathcal{P}$.*

Proof. This is basically [24, Theorem 1.10]. The tightness of $(\mu\mathcal{P}^n)$ implies tightness of (ν_n) and therefore there exists a weakly converging subsequence (ν_{n_k}) with limit $\pi \in \mathscr{P}(G)$ by Prokhorov's Theorem. 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

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

□

We have the following existence result.

Proposition 2.8.3 (existence of invariant measure)**.** *Let (G,d) be a Polish space and $\mathcal{P}$ be the Markov operator corresponding to the transition kernel in (2.6). Suppose $\mathcal{P}$ is Feller and that there exists a compact set $K \subset G$ and $\mu \in \mathscr{P}(G)$ with*

$$\limsup_{n\to\infty} \nu_n^\mu(K) := \limsup_{n\to\infty} \tfrac{1}{n}\sum_{i=1}^n \mu\mathcal{P}^i(K) > 0,$$

then there exists an invariant measure for $\mathcal{P}$.

Proof. See [35, Proposition 3.1]. □

Remark 2.8.4: It readily can be shown that there exists an invariant probability measure π that has compact support, if and only if $\limsup_{n\to\infty} \nu_n^\mu(K) = 1$ for some compact K and some $\mu \in \mathscr{P}(G)$.

A Markov operator need not possess a unique invariant probability measure or any invariant measure at all, as the next example illustrates.

Example 2.8.5 (no invariant measure)**.** Consider the case of two subspaces C_1 and C_2 in $\mathbb{R}^n$ that have noncompact intersection (i.e. both at least 2 dimensional) and which are distorted by an affine noise model as follows. Let ξ_1, ξ_2 be independent affine perturbations of the projectors onto C_1 and C_2 respectively, i.e. we let

$$(x \in \mathbb{R}^n) \qquad P_{\xi_1} x = P_{C_1} x + \xi_1, \qquad P_{\xi_2} x = P_{C_2} x + \xi_2 \tag{2.7}$$

where P_{C_i} denotes the projector onto C_i $(i = 1, 2)$. We consider the fixed point mapping T_ξ in Algorithm 1 corresponding to the composition of the projectors: $T_\xi x := P_{\xi_1} P_{\xi_2} x$ where $\xi = (\xi_1, \xi_2)$. The noise ξ satisfies the property $\mathbb{P}(\langle h, \xi_i\rangle > 0) = \alpha > 0$ $(i = 1, 2)$ for all $h \in C_1 \cap C_2$ with $\|h\| = 1$. (In particular, this holds for isotropic noise.) For this noise model, one can show that there does not exist an invariant measure.

To see this, note that we can find an $h \in \mathbb{R}^n$ with $\|h\| = 1$ such that $th \in C_1 \cap C_2$ for all $t \in \mathbb{R}$. For any $c = c(t) = th$, let

$$H_{>c} := \{x \in \mathbb{R}^n \mid \langle h, x - c\rangle > 0\}$$

be the open half-space with normal h that contains $th+c$ for $t > 0$. Note that for $x \in H_{>c}$, the probability to end up in $H_{>c}$ after one projection is $\geq \alpha$, since

$$\langle h, P_{\xi_i} x - c\rangle = \langle h, P_{C_i} x - c\rangle + \langle h, \xi_i\rangle, \qquad i = 1, 2$$

Here $\langle h, P_{C_i} x - c\rangle > 0$ since C_i is a subspace and $\mathbb{P}(\langle h, \xi_i\rangle > 0) = \alpha$, by the assumption on the noise. So one has for $x \in H_{>c}$ that

$$\begin{aligned}\mathbb{P}(T_\xi x \in H_{>c}) &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\{P_{\xi_1} P_{\xi_2} x \in H_{>c}\} \mid \xi_2\right]\right] \\ &\geq \alpha \mathbb{E}\left[\mathbb{1}\{P_{\xi_2} x \in H_{>c}\}\right] \geq \alpha^2.\end{aligned}$$

This is in contradiction to the existence of an invariant measure π. Indeed,

$$\pi(H_{>c}) \geq \int_{H_{>c}} p(x, H_{>c}) \pi(\mathrm{d}x) \geq \alpha^2$$

for any $c \in C_1 \cap C_2$. But π is tight, so there is a compact set K for which $\pi(K) > 1 - \alpha^2$. If c is chosen such that $K \cap H_{>c} = \emptyset$, the contradiction follows.

Consider the Polish space $(\mathbb{R}^n, \|\cdot\|)$ for the case that $T_i = P_i$, $i \in I$ is a projector onto a nonempty closed and convex set. 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 [13, Theorem 2]). We try to translate this into our setting. A sufficient condition for the existence of an invariant measure for $\mathcal{P}$ is then the existence of a compact set $K \subset \mathbb{R}^n$ with $p(x, K) \geq \epsilon$ for all $x \in \mathbb{R}^n$ and some $\epsilon > 0$. This would be given in the case that I consists of only finitely many sets, where one of them, say $C_i = K$, is compact and $\mathbb{P}(\xi = i) = \epsilon$, since $p(x, K) = \mathbb{P}(P_\xi x \in K) \geq \mathbb{P}(P_i x \in K, \xi = i) = \mathbb{P}(\xi = i) = \epsilon$ for all $x \in \mathbb{R}^n$. More generally, we have the following result

Corollary 2.8.6 (existence of invariant measures for finite collections of continuous mappings)**.** *Let (G, d) be a Polish space and let $T_i : G \to G$ be continuous for $i \in I$, where I is a finite index. 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

$$\mathbb{P}(X_{k+1} \in K) = \mathbb{E}[\mathbb{P}\left(T_{\xi_k} X_k \in K \,|\, X_k\right)] \geq \mathbb{P}(\xi = i) \qquad \forall k \in \mathbb{N}_0.$$

The assertion follows now immediately from Proposition 2.8.3, since $\mathbb{P}(\xi = i) > 0$ and $\mathcal{P}$ is Feller by continuity of T_j for all $j \in I$. $\square$

Next we mention an existence version for invariant measures of the RFI Markov operator, if the corresponding RFI sequence (X_k) has uniformly bounded expectation.

Lemma 2.8.7 (existence in $\mathbb{R}^n$, RFI)**.** *Let $T_i : \mathbb{R}^n \to \mathbb{R}^n$ $(i \in I)$. Let (X_k) be an 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 Eq. (2.6).*

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 Proposition 2.8.3 since closed balls in $\mathbb{R}^n$ 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 $\mathcal{P}$. $\square$

2.9. Wasserstein metric

A useful metric later on, to describe convergence of a sequence of probability measures, in many examples is the so called Wasserstein metric on the space of probability measures $\mathscr{P}(G)$. Especially to describe geometric convergence this metric will become very helpful. We give in the following the definition and some properties we make use of.

Lemma 2.9.1 (properties of the Wasserstein metric). *Let (G, d) be a Polish space. Denote the Wasserstein metric $W_p : \mathscr{P}(G) \times \mathscr{P}(G) \to [0, \infty]$ for $p \in [0, \infty)$ and $\mu, \nu \in \mathscr{P}(G)$ through*

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in C(\mu,\nu)} \int d^p(x, y)\gamma(\mathrm{d}x\,, \mathrm{d}y) \right)^{\frac{1}{p}}.$$

Define the subsets $\mathscr{P}_p(G) \subset \mathscr{P}(G)$ via

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

(i) The representation of $\mathscr{P}_p(G)$ is independent of x and for $\mu, \nu \in \mathscr{P}_p(G)$ we have $W_p(\mu, \nu) < \infty$.

(ii) If $W_p(\mu, \nu) < \infty$, then the infimum is attained.

(iii) $(\mathscr{P}_p(G), W_p(G))$ is a Polish space.

(iv) If $W_p(\mu_n, \mu) \to 0$ for $(\mu_n) \subset \mathscr{P}(G)$, then $\mu_n \to \mu$ (but not vice versa).

Proof. (i) See [53, Remark after Definition 6.4].

(ii) From Lemma 2.6.3 we know that the sequence (γ_n) which is assumed to be a minimizing sequence for $W_p(\mu, \nu)$ is tight and hence there is a clusterpoint $\gamma \in C(\mu, \nu)$. By continuity of the metric d it follows that d is lower semi-continuous and bounded from below and from [50, Theorem 9.1.5] follows then $\gamma d \le \liminf_k \gamma_{n_k} d = W_p(\mu, \nu)$.

(iii) See [53, Theorem 6.9].

(iv) See [53, Theorem 6.18].

□

We want to emphasize here that the Wasserstein metric may take the value infinity, unless G is compact. For example in $\mathbb{R}^n$ the Wasserstein metric takes the value infinity if $\mu = \mathcal{L}(X)$ with $\mathbb{E}[\|X\|] = \infty$ and $\nu \in \mathscr{P}(G)$ has compact support. But if we would choose $\nu = \mu(\cdot - a)$ for some $a \in \mathbb{R}^n$, then $W(\mu, \nu) \le \|a\| < \infty$. Also (iv) of Lemma 2.9.1 is an important implication. If we are able to show convergence of $(\mathcal{L}(X_n))$ to a probability measure μ in the Wasserstein metric, then we immediately get weak convergence $\mathcal{L}(X_n) \to \mu$. Since convergence in the Wasserstein metric is stronger than weak convergence [53, Definition 6.8] the implication is not true in general, unless G is compact [53, Corollary 6.13].

In Section 8.2 we give an important application of the Wasserstein metric to describe geometric convergence of the RFI Markov chain for the case the mappings T_i $(i \in I)$ are contractions in expectation, as for example is the case for T_i being the projection onto an affine subspace (Example 8.2.6).

Theorem 2.9.2 (Geometric convergence for contractions in expectation)**.** *Let (G, d) be a Polish space and let $\{T_i \,|\, i \in I\}$ be* contractions in expectation*, i.e. there exists $r \in (0, 1)$ such that $T_i : G \to G$ $(i \in I)$ are mappings with $\mathbb{E}[d(T_\xi x, T_\xi y)] \le r d(x, y)$ for all $x, y \in G$, see also [48]. Suppose furthermore that there exists $y \in G$ with $\mathbb{E}[d(T_\xi y, y)] < \infty$, then there exists a unique invariant measure $\pi \in \mathscr{P}_1(G)$ for $\mathcal{P}$ with*

$$W(\mu\mathcal{P}^n, \pi) \le r^n W(\mu, \pi)$$

for all $\mu \in \mathscr{P}_1(G)$.

Proof. Note that for any pair of distributions $\mu, \nu \in \mathscr{P}_1(G)$ and a pair random variables (X, Y) that is an optimal coupling for $W(\mu, \nu)$ (possible by Lemma 2.9.1) satisfies

$$W(\mu\mathcal{P}, \nu\mathcal{P}) \le \mathbb{E}[d(T_\xi X, T_\xi Y)] \le r\mathbb{E}[d(X, Y)] = rW(\mu, \nu),$$

if (X, Y) is chosen independent of ξ (possible due to Theorem 2.2.1) and by application of Remark 2.3.3. We have that $\mathcal{P} : \mathscr{P}_1(G) \to \mathscr{P}_1(G)$: let $X \sim \mu \in \mathscr{P}_1(G)$ with $X \perp\!\!\!\perp \xi$ then

$$\mathbb{E}[d(T_\xi X, y)] \le \mathbb{E}[d(T_\xi X, T_\xi y)] + \mathbb{E}[d(T_\xi y, y)] < \infty,$$

that means $\mu\mathcal{P} \in \mathscr{P}_1(G)$. So in total we have shown that $\mathcal{P}$ is a contraction on the Polish space $(\mathscr{P}_1(G), W)$ and hence Banach's Fixed Point Theorem yields the assertion of the theorem. □

We want to note here that there is no further assumption like continuity or nonexpansiveness on the mappings T_i $(i \in I)$ needed, also the underlying metric space is a general Polish space, which emphasizes the generality of this assertion.

As a consequence of Example 2.8.5, from Theorem 2.9.2 we can conclude that the mapping T_ξ cannot be contractive in expectation, indeed $\|T_\xi\| = 1$.

2.10. TV-norm

Another metric on the space of probability measures is the TV-norm. We will make use of it in some examples later on to describe geometric convergence of the sequence $(\mathcal{L}(X_n))$ to an invariant probability measure for the RFI Markov chain for projectors onto compact intervals (Example 8.2.8). In the following we give its definition on some properties we make use of.

Lemma 2.10.1 (properties of the TV-norm)**.** *Let (G, d) be a metric space. Define the TV-norm of two probability measures as*

$$\|\mu - \nu\|_{\mathrm{TV}} := \sup_{A \in \mathcal{B}(G)} |\mu(A) - \nu(A)|.$$

Then

(i) $\|\mu-\nu\|_{\mathrm{TV}} = \frac{1}{2}\sup_{f\colon G\to[-1,1]\ mb.} |\mu f - \nu f| = \inf_{\mathcal{L}(X,Y)\in C(\mu,\nu)} \mathbb{P}(X\neq Y)$.

(ii) $(\mathscr{P}(G), \|\cdot\|_{\mathrm{TV}})$ *is complete.*

(iii) $\sup_{x,y\in D}\|p(x,\cdot)-p(y,\cdot)\|_{\mathrm{TV}} = \sup_{\mu,\nu\in\mathscr{P}(D),\mu\neq\nu} \frac{\|\mu\mathcal{P}-\nu\mathcal{P}\|_{\mathrm{TV}}}{\|\mu-\nu\|_{\mathrm{TV}}}$ *for* $D\subset G$.

Proof. (i) See [46, Proposition 3].

(ii) See [23, Theorem 4.28].

(iii) See [19, Corollary 3.14].

□

Next we show that in view of Remark 2.5.2 the TV-norm is equivalent to the operator norm in the dual space of $(C_b(G), \|\cdot\|_\infty)$, so that convergence in either norm implies convergence in the other.

Lemma 2.10.2 (TV-norm equivalent to operator norm)**.** *Let* (G,d) *be a metric space and let* $\mu,\nu\in\mathscr{P}(G)$*. Then*

$$\|\mu-\nu\|_{\mathrm{TV}} \le \|\mu-\nu\|_* \le 2\|\mu-\nu\|_{\mathrm{TV}},$$

where for $\phi\in C_b^*(G)$

$$\|\phi\|_* = \sup_{\substack{f\in C_b(G)\\ \|f\|_\infty=1}} \langle\phi, f\rangle\,.$$

Proof. From (i) in Lemma 2.10.1 we get that

$$\|\mu-\nu\|_* \le 2\|\mu-\nu\|_{\mathrm{TV}}.$$

For the other direction, fix $U\subset G$ open and let (f_n) be a sequence in $C_b(G)$ with $\|f_n\|_\infty = 1$ and $0\le f_n\le \mathbb{1}_U$ (construction in the proof of Theorem 2.5.1). Then

$$\|\mu-\nu\|_* = \sup_{\substack{f\in C_b(G)\\ \|f\|_\infty=1}} |(\mu-\nu)f| \ge \lim_{n\to\infty} |(\mu-\nu)f_n| = |\mu(U)-\nu(U)|.$$

This inequality follows immediately also for all closed sets A, since $G\setminus A$ is open.
For any real-valued measure η (σ-additive and $\eta(\emptyset)=0$) on a measure space $(S,\mathcal{S})$ there exist disjoint $S^+,S^-\in\mathcal{S}$ with $\eta^+(A) := \eta(S^+\cap A)\ge 0$ and $\eta^-(A) := -\eta(S^-\cap A)\ge 0$ for all $A\in\mathcal{S}$ (*Hahn-Jordan decomposition* [9, Theorem 3.1.1]). In particular, $\eta=\eta^+-\eta^-$.
Letting $\eta=\mu-\nu$, we have that

$$(\mu-\nu) = (\mu-\nu)^+ - (\mu-\nu)^-,$$

and the existence of $G^+\in\mathcal{B}(G)$ with $(\mu-\nu)(G^+) = (\mu-\nu)^+(G)$.
Since $(\mu-\nu)^+$ is a finite measure, by Theorem A.0.18 there exists a sequence (A_n) of

closed sets with $A_n \subset G^+$ and $(\mu-\nu)^+(A_n) \to (\mu-\nu)^+(G^+)$. Hence we have (by Theorem A.0.18) that

$$\limsup_n (\mu-\nu)^-(A_n) = \limsup_n (\mu-\nu)(A_n) - (\mu-\nu)^+(A_n) \le (\mu-\nu)(G^+) - (\mu-\nu)^+(G^+) = 0,$$

which implies (for n large enough)

$$\|\mu-\nu\|_* \ge \lim_n |\mu(A_n) - \nu(A_n)| = \lim_n (\mu-\nu)^+(A_n) - (\mu-\nu)^-(A_n) = (\mu-\nu)^+(G^+).$$

Now, since

$$\begin{aligned}\|\mu-\nu\|_{\mathrm{TV}} &= \sup_{A\in\mathcal{B}(G)} |\mu(A)-\nu(A)| = \sup_{A\in\mathcal{B}(G)} \mu(A)-\nu(A)\\ &= \sup_{A\in\mathcal{B}(G)} (\mu-\nu)^+(A) - (\mu-\nu)^-(A) = (\mu-\nu)^+(G^+),\end{aligned}$$

it follows that $\|\mu-\nu\|_* \ge \|\mu-\nu\|_{\mathrm{TV}}$. □

We want to note that convergence of $(\mathcal{L}(X_k))$ in the TV-norm is a strong type of convergence, that implies convergence in the Wasserstein and Levi-Prokhorov metric. In contrast to convergence in the weak sense, convergence in the TV-norm implies that $\mu_n f \to \mu f$ for all measurable and bounded $f : G \to \mathbb{R}$.

Now we give a helpful result on geometric convergence of a Markov chain in the TV-norm, under the assumption that the *global minorization* condition for the transition kernel is satisfied.

Theorem 2.10.3 (Doeblin, Theorem 4.29 in [23])**.** *Let (G,d) be a metric space. Assume p satisfies the global minorization property, i.e. there exists a probability measure ν and $\kappa > 0$ such that $p(x,A) \ge \kappa\nu(A)$ for all $A \in \mathcal{B}(G)$ and $x \in G$. Then there exists a unique invariant measure π for $\mathcal{P}$ and, for all probability measures μ, it holds that $\|\mu\mathcal{P}^n - \pi\|_{TV} \le (1-\kappa)^n$.*

Note that the global minorization property can be interpreted as a (strong) regularity property of the transition kernel of the Markov chain. It is an interesting question, if regularity properties of the kernel can be formulated similar to metric regularity of mappings to describe linear (geometric) convergence of fixed point iterations also in cases where $\operatorname{inv}\mathcal{P}$ is not a singleton.

CHAPTER 3

The Stochastic Fixed Point Problem

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a (rich enough) probability space and (G, d) be a Polish space. Let $\{T_i \,|\, i \in I\}$ be a family of mappings from $G \to G$, where I is an arbitrary index set. Let ξ be a random variable into the measurable space $(I, \mathcal{I})$, where $\mathcal{I}$ is a σ-algebra on I. We can define an update function $\Phi : G \times I \to G$ that generates an RFI on G via $\Phi(x, i) = T_i x$ in Algorithm 1, so that, if an i.i.d. sequence (ξ_k) and $X_0 \sim \mu$ with $X_0 \perp\!\!\!\perp (\xi_k)$ are given, where $\mu \in \mathscr{P}(G)$ is the initial distribution, we have

$$X_{k+1} = T_{\xi_k} X_k, \qquad k \in \mathbb{N}_0.$$

With the notation from Section 2.7 the induced transition kernel (RFI kernel) on G is

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

Then the the stochastic fixed point problem is to

$$\text{Find } \pi \in \operatorname{inv} \mathcal{P}, \tag{3.1}$$

where $\mathcal{P}$ is the Markov operator defined through the RFI kernel (see Section 2.7) and where

$$\operatorname{inv} \mathcal{P} := \{\pi \in \mathscr{P}(G) \,|\, \pi \mathcal{P} = \pi\},$$

see Section 2.8. For the stochastic fixed point problem to be well-defined, we will henceforth assume without exception as a *standing assumption* that Φ is measurable.

There are two further specializations which are fundamentally different. The *consistent* and *inconsistent stochastic feasibility* problem. While the former is characterized by observation of almost sure convergence of the RFI sequence of random variables for the latter at most convergence of the distributions in the weak sense can be expected.

3.1. Consistent Stochastic Feasibility Problem

The *stochastic feasibility* problem is to find a point

$$x^* \in C := \{x \in G \,|\, \mathbb{P}(x \in \operatorname{Fix} T_\xi) = 1\}\,, \tag{3.2}$$

where the fixed point set of the operator T_i is denoted as

$$\operatorname{Fix} T_i = \{x \in G \,|\, x = T_i x\}\,.$$

The problem is called *consistent*, if $C \neq \emptyset$.

Letting $T_i = P_i$ be a projector onto a convex, closed and nonempty set C_i in a Hilbert space ($i \in I$), this specializes immediately to the stochastic feasibility problem formulated by Butnariu and Flåm [15] where $\operatorname{Fix} T_\xi = C_\xi$. In order to make sense of this formulation of stochastic feasibility, we need the event $\{x \in \operatorname{Fix} T_\xi\}$ to be an element of $\mathcal{F}$ for any $x \in G$.

Remark 3.1.1: Since $\{x\} \in \mathcal{B}(G)$ and the function $\Phi_\xi : G \times \Omega \to G\,,\ (x, \omega) \mapsto \Phi_\xi(x, \omega) := (\Phi \circ (\operatorname{Id}, \xi))(x, \omega) = T_{\xi(\omega)} x$ is measurable as composition of two measurable functions, we find

$$\begin{aligned}
\{x \in \operatorname{Fix} T_\xi\} &= \left\{\omega \in \Omega \,\middle|\, x \in \operatorname{Fix} T_{\xi(\omega)}\right\} \\
&= \left\{\omega \in \Omega \,\middle|\, T_{\xi(\omega)} x = x\right\} \\
&= \left\{\omega \in \Omega \,\middle|\, (x, \omega) \in \Phi_\xi^{-1}\{x\}\right\} \\
&\in \mathcal{F},
\end{aligned}$$

since slices of sets in the product σ-field are measurable with respect to the single σ-fields (see Lemma A.0.19).

Furthermore, the definition of C does not depend on the variable ξ itself only on its distribution as can be seen by the following computation:

$$\mathbb{P}(x \in \operatorname{Fix} T_\xi) = \int_\Omega \mathbb{1}\{T_\xi x = x\}\, \mathrm{d}\mathbb{P} = \int_I \mathbb{1}\{T_i x = x\} \mathbb{P}^\xi(\mathrm{d}i).$$

3.2. Consistent Stochastic Feasibility for Continuous Mappings

We will specialize from here on to continuous functions. This will have too major advantages, for one, the consistent feasibility problem then becomes a deterministic feasibility problem which will be used throughout, and, for the other, consistent feasibility characterizes a very strong kind of convergence for Markov chains, namely, almost sure convergence. We tend to the latter now.

Lemma 3.2.1 ($C \neq \emptyset$ is necessary for a.s. convergence)**.** *Let (G,d) be a Polish space and T_i be continuous $(i \in I)$. Let there exist $\pi \in \operatorname{inv}\mathcal{P}$ and let $X \sim \pi$ with $X \perp\!\!\!\perp (\xi_k)$. If the RFI sequence (X_k^X) converges almost surely as $k \to \infty$, then* $\operatorname{supp}\pi \subset C$*.*

Proof. We assume from now on that $C = \emptyset$, i.e. $\mathbb{P}(T_\xi x = x) < 1$ for all $x \in G$ and lead that to a contradiction. Fix $x \in \operatorname{supp}\mathcal{L}(X) = \operatorname{supp}\pi$ and let for $\epsilon > 0$

$$A^\epsilon := \{i \in I \,|\, d(T_i x, x) \geq \epsilon\}\,.$$

Since $A^{1/n} \uparrow A^{0+} := \{i \in I \,|\, d(T_i x, x) > 0\}$ as $n \to \infty$ and $\mathbb{P}^\xi(A^{0+}) = \mathbb{P}(T_\xi x \neq x) > 0$, there exists $\epsilon_0 > 0$ with $\mathbb{P}^\xi(A^{\epsilon_0}) > 0$. Define now in view of continuity of T_i $(i \in I)$ the sets

$$A_n^\epsilon := \left\{ i \in I \,\middle|\, d(T_i y, T_i x) \leq \epsilon \quad \forall y \in \mathbb{B}(x, \tfrac{1}{n}) \right\}$$

for $n \in \mathbb{N}$. The set A_n^ϵ is measurable, since

$$\begin{aligned} g = \inf_{y \in D_n} g_y, \qquad &\text{where} \\ i \in I \mapsto g_y(i) = \mathbb{1}\{d(T_i y&, T_i x) \leq \epsilon\}, \end{aligned}$$

is measurable, where $D_n \subset \mathbb{B}(x, \frac{1}{n})$ is countable and dense (exists by Theorem A.0.20), and application of Lemma A.0.19. Since $A_n^\epsilon \uparrow I$ as $n \to \infty$ for any $\epsilon > 0$, there exists $M > 0$ with $1/M < \epsilon_0/2$ and $\mathbb{P}^\xi(A) > 0$, where $A := A_M^{\epsilon_0/2} \cap A^{\epsilon_0}$. For each $i \in A$ we have that

$$d(T_i y, x) \geq d(T_i x, x) - d(T_i y, T_i x) \geq \frac{\epsilon_0}{2} > \frac{1}{M} \qquad \forall y \in \mathbb{B}(x, \tfrac{1}{M}).$$

We then have, denoting $B := \mathbb{B}(x, \frac{1}{M})$, that

$$\mathbb{P}(X_k^X \in B, X_{k+1}^X \notin B) \geq \mathbb{P}(X_k^X \in B, \xi_k \in A) = \pi(B)\mathbb{P}^\xi(A) > 0,$$

for all $k \in \mathbb{N}_0$, where we used independence of ξ_k and X_k^X and that $\mathbb{P}(X_k^X \in B) = \pi(B) > 0$ for all $k \in \mathbb{N}_0$, since by Lemma 2.8.1 it holds that $X_k^X \sim \pi$ for all $k \in \mathbb{N}$. Since by assumption $X_k^X \to Y$ a.s. for some random variable Y, it also follows that

$$\mathbb{P}(X_k^X \in B, X_{k+1}^X \notin B) \to \mathbb{P}(Y \in B, Y \notin B) = 0,$$

which is a contradiction and the assumption $\mathbb{P}(T_\xi x = x) < 1$ needs to be false, so $x \in C$. □

Next we convince ourselves that the formulation in Eq. (3.2) is indeed a fixed point problem in the classical deterministic sense. Therefore denote in the following for $A \subset \Omega$,

$$C(A) := \bigcap_{\omega \in A} \operatorname{Fix} T_{\xi(\omega)}.$$

Lemma 3.2.2 (equivalence of stochastic and deterministic feasibility problems)**.** *Let (G, d) be a Polish space. Let T_i be continuous $(i \in I)$. If $C \neq \emptyset$, there exists a $\mathbb{P}$-nullset $N \subset \Omega$, such that*

$$C = C(\Omega \setminus N) = \bigcap_{\omega \in \Omega \setminus N} \operatorname{Fix} T_{\xi(\omega)}.$$

Furthermore, $C \subset G$ is closed.

Proof. For the direction "$\supset$", let $x \in C(\Omega \setminus N)$ for a $\mathbb{P}$-nullset $N \subset \Omega$, then $\mathbb{P}(x \in \operatorname{Fix} T_\xi) = \mathbb{P}(\Omega \setminus N) = 1$, i.e. $x \in C$.
Consider now the direction "$\subset$". Let Q be a dense and countable subset of C (exists by Theorem A.0.20). Since for each $q \in Q$, $\mathbb{P}(q \in \operatorname{Fix} T_\xi) = 1$, there is $N_q \subset \Omega$ with $\mathbb{P}(N_q) = 0$ and $q \in C(\Omega \setminus N_q)$. Set $N = \bigcup_{q \in Q} N_q$, then $\mathbb{P}(N) = 0$ and $q \in C(\Omega \setminus N)$ for all $q \in Q$.
Now let $c \in C$, so $\exists (q_n)_{n \in \mathbb{N}} \subset Q$ with $q_n \to c$ as $n \to \infty$. Since, for all $i \in I$, $\operatorname{Fix} T_i$ is closed by continuity of T_i, we get $c = \lim_{n \to \infty} q_n \in C(\Omega \setminus N)$.
The set $C(\Omega \setminus N)$ is defined as intersection over closed sets and hence closed itself. $\square$

Remark 3.2.3 (interpretation)**:** Lemma 3.2.2 shows that the feasible set C in the separable case can be written as intersection of a selection of sets $\operatorname{Fix} T_{\xi(\omega)}$ as in the deterministic formulation of the fixed point problem, but where $\omega \in \Omega \setminus N$ for a nullset $N \subset \Omega$. In fact $C(\Omega)$ is in general a proper subset of $C = C(\Omega \setminus N)$ or can even be empty. But note that, even though the construction of C in Lemma 3.2.2 appears to depend on the random variable ξ, in fact C only depends on the *distribution* $\mathbb{P}^\xi$ as pointed out earlier.
Furthermore, in the context of more general Markov chains, we have,

$$(c \in C) \qquad p(c, \{c\}) = \mathbb{P}(T_\xi c \in \{c\}) = \mathbb{P}(\Omega \setminus N) = 1.$$

Hence

$$(A \in \mathcal{B}(G)) \qquad \delta_c \mathcal{P}(A) = p(c, A) = \mathbb{1}_A(c) = \delta_c(A).$$

In other words, the delta function δ_c for $c \in C$ is an invariant measure for $\mathcal{P}$.

Corollary 3.2.4 ($\mathbb{P}^\xi$ nullset, separable space)**.** *Under the assumptions of Lemma 3.2.2 there exists a $\mathbb{P}$-nullset N with $C = C(\Omega \setminus N)$, such that $\xi(N) := \{\xi(\omega) \,|\, \omega \in N\}$ is a $\mathbb{P}^\xi$-nullset, where we denote $\mathbb{P}^\xi = \mathbb{P}(\xi \in \cdot)$, and it satisfies*

$$C = \bigcap_{i \in \xi(\Omega) \setminus \xi(N)} \operatorname{Fix} T_i.$$

Proof. We will construct a $\mathbb{P}$-nullset N for which $\xi(\Omega \setminus N) = \xi(\Omega) \setminus \xi(N)$, where $\xi(N)$ is a $\mathbb{P}^\xi$-nullset, in that case immediately follows that

$$\bigcap_{\omega \in \Omega \setminus N} \operatorname{Fix} T_{\xi(\omega)} = \bigcap_{i \in \xi(\Omega) \setminus \xi(N)} \operatorname{Fix} T_i.$$

Let $A_x := \{i \in I \,|\, T_i x = x\}$ for $x \in G$, then analogously to Remark 3.1.1

$$A_x = \left\{ i \in I \,\middle|\, (x,i) \in \Phi^{-1}\{x\} \right\} \in \mathcal{I}$$

and so is $A := \bigcap_{c \in C} A_c = \bigcap_{q \in Q} A_q$ as countable intersection of measurable sets ($Q \subset C$ dense and countable, see proof of Lemma 3.2.2). Let $\tilde{N}$ be the $\mathbb{P}$-nullset from Lemma 3.2.2, i.e. $C = C(\Omega \setminus \tilde{N})$, note that due to

$$C = \bigcap_{\omega \in \Omega \setminus \tilde{N}} \operatorname{Fix} T_{\xi(\omega)} = \bigcap_{i \in \xi(\Omega \setminus \tilde{N})} \operatorname{Fix} T_i \neq \emptyset$$

it holds $\xi(\Omega \setminus \tilde{N}) \subset A_c \neq \emptyset$, for all $c \in C$. Set $N := \Omega \setminus \xi^{-1}A$, then from $\Omega \setminus \tilde{N} \subset \xi^{-1}A$ follows $N \subset \tilde{N}$ is a $\mathbb{P}$-nullset and

$$\mathbb{P}^{\xi}(A) = \mathbb{P}(\xi^{-1}A) \geq \mathbb{P}(\Omega \setminus \tilde{N}) = 1,$$

i.e. $\mathbb{P}^{\xi}(\xi(N)) = 1 - \mathbb{P}^{\xi}(A) = 0$. By definition of A we have for $\omega \in \xi^{-1}A$, that any $c \in C$ satisfies $c \in \operatorname{Fix} T_{\xi(\omega)}$, so it follows $C \subset C(\xi^{-1}A)$. Due to $C(\Omega \setminus N) \subset C(\Omega \setminus \tilde{N})$ holds

$$C = \bigcap_{\omega \in \xi^{-1}A} \operatorname{Fix} T_{\xi(\omega)} = \bigcap_{i \in \xi(\xi^{-1}A)} \operatorname{Fix} T_i = \bigcap_{i \in \xi(\Omega) \setminus \xi(N)} \operatorname{Fix} T_i.$$

Note that from $N = \Omega \setminus \xi^{-1}A$ follows $\xi(N) = \xi(\Omega) \setminus \xi(\xi^{-1}A)$. □

If ξ is not surjective, then $\xi(\Omega) \neq I$. In that case, there is a $\mathbb{P}^{\xi}$-nullset $\xi(N)$ of indices in I, that are not needed to characterize the fixed point set, and these indices can be removed from the index set I. Note also that, in general, the $\mathbb{P}$-nullsets occurring in Lemma 3.2.2 and Corollary 3.2.4 are different. If there is $N \subset \Omega$ with $C = C(\Omega \setminus N)$, then it need not be the case that $C = \bigcap_{i \in \xi(\Omega) \setminus \xi(N)} \operatorname{Fix} T_i$.

In the context of the iterates X_k of the RFI in many of the results below we construct the set N in Lemma 3.2.2 as follows:

$$N = \bigcup_k N_k \quad \text{where } N_k := \Omega \setminus \{\omega \in \Omega \mid T_{\xi_k(\omega)} c = c \ \forall c \in C\}. \tag{3.3}$$

From Lemma 3.2.2 we have that N_k is a set of measure zero, hence so is N.

3.3. Inconsistent Stochastic Feasibility

If $C = \emptyset$ we call this the *inconsistent stochastic feasibility* problem.

Example 3.3.1 (inconsistent stochastic feasibility)**.** Consider the (trivially convex, non-empty and closed) sets $C_{-1} := \{-1\}$ and $C_1 := \{1\}$ together with a random variable ξ such that $\mathbb{P}(\xi = 1) = \mathbb{P}(\xi = -1) = 1/2$. The mappings $T_i x = P_{C_i} x = i$ for $x \in \mathbb{R}$ and $i \in I = \{-1, 1\}$ are the projections onto the sets C_{-1} and C_1. The RFI iteration then amounts to just random flipping between the values -1 and 1. So it holds that

$\mathbb{P}(T_\xi x = i) = 1/2$ for all $x \in \mathbb{R}$ and hence there is clearly no feasible fixed point to this iteration, that is, $C = \emptyset$. Nevertheless, by Theorem 2.3.2 we have

$$\mathbb{P}(X_{k+1} = i) = \mathbb{E}[\mathbb{P}\left(T_{\xi_k} X_k = i \,|\, X_k\right)] = \frac{1}{2}$$

for all $k \in \mathbb{N}_0$. That means the unique invariant distribution to which the distributions of the iterates of the RFI (i.e. $(\mathbb{P}(X_k \in \cdot))_k$) converges is $\pi = \frac{1}{2}(\delta_{-1} + \delta_1)$, and this is attained after one iteration.

As the above example demonstrates, inconsistent stochastic fixed point problems are neither exotic nor devoid of meaningful notions of convergence. As we have seen in Lemma 3.2.1, one can not expect the RFI to converge to a point in the case $C = \emptyset$, but still one can ask for convergence of the distributions $(\mathcal{L}(X_k))$ of the iterates (X_k) of the RFI to some invariant measure π for $\mathcal{P}$.

3.4. Notions of Convergence for Inconsistent Feasibility

In the following, we will describe two modes of convergence for the sequence $(\mathcal{L}(X_k))$ on $\mathscr{P}(G)$.

1. Weak convergence of the Cesáro average of the distributions of the variables (X_k) to a probability measure $\pi \in \mathscr{P}(G)$, i.e. for continuous and bounded $f \in C_b(G)$

$$\nu_n f := \frac{1}{n}\sum_{k=1}^{n} \mathcal{L}(X_k) f = \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} f(X_k)\right] \to \pi f, \qquad \text{as } n \to \infty.$$

2. Weak convergence of the probability distributions of the variables (X_k) to a probability measure $\pi \in \mathscr{P}(G)$, i.e. for continuous and bounded $f \in C_b(G)$

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

Clearly, the second mode implies the first, but the latter will not occur as natural as seen in Example 8.1.11.

If we again assume continuity of the family of mappings T_i on the Polish space (G, d) we get that, indeed, the limiting measure is an invariant measure for $\mathcal{P}$.

Lemma 3.4.1. *Let (G, d) be a Polish space. Let T_i be continuous ($i \in I$). Suppose $\nu_n^\mu \to \pi$ or $\mu\mathcal{P}^k \to \pi$ in the weak sense, then $\pi \in \operatorname{inv} \mathcal{P}$.*

Proof. An elementary fact from the theory of Markov chains (Theorem 2.8.2) is that, if π is a cluster point of (ν_n) in the weak sense, then π is an invariant probability measure for the Markov operator $\mathcal{P}$. From Theorem 2.5.3 we have that (ν_n) is tight and hence relatively compact in $\mathscr{P}(G)$. In particular, this means any subsequence has a convergent subsequence (ν_{n_k}) with

$$(\forall f \in C_b(G)) \qquad \nu_{n_k} f = \mathbb{E}\left[\frac{1}{n_k}\sum_{j=1}^{n_k} f(X_j)\right] \to \pi f, \qquad \text{as } k \to \infty.$$

Convergence of the whole sequence, i.e. $\nu_n \to \pi$, amounts then to showing that π is the unique cluster point of (ν_k) (see Theorem A.0.16), which is clear in our case.
For the other case, one has

$$\pi f \leftarrow \mu \mathcal{P}^{k+1} f = \mu \mathcal{P}^k(\mathcal{P} f) \to \pi(\mathcal{P} f)$$

as $k \to \infty$ for any $f \in C_b(G)$. So it needs to hold that $\pi = \pi\mathcal{P}$. $\square$

The notion of convergence we considered for the consistent stochastic feasibility problem was much stronger. Clearly, almost sure convergence of the sequence implies the more general notion above. This is common in the studies of stochastic algorithms in optimization, though this does not require the full power of the theory of general Markov processes.

CHAPTER 4

Convergence Analysis - Consistent Feasibility

We achieve convergence of iterated random functions for consistent stochastic feasibility in several different settings under different assumptions on the metric spaces and the mappings T_i $(i \in I)$. The main properties of the mappings we consider are:

- quasi-nonexpansive mappings, i.e.

$$(\forall x \notin \operatorname{Fix} T_i)(\forall y \in \operatorname{Fix} T_i) \qquad d(T_i x, y) \leq d(x, y). \tag{4.1}$$

- paracontractions, i.e. T_i is continuous and

$$(\forall x \notin \operatorname{Fix} T_i)(\forall y \in \operatorname{Fix} T_i) \qquad d(T_i x, y) < d(x, y); \tag{4.2}$$

- nonexpansive mappings, i.e.

$$(\forall x, y \in G) \qquad d(T_i x, T_i y) \leq d(x, y); \tag{4.3}$$

- averaged mappings on a normed linear space $\mathcal{H}$, i.e. mappings $T : \mathcal{H} \to \mathcal{H}$ for which there exists an $\alpha \in (0, 1)$ such that

$$(\forall x, y \in \mathcal{H}) \qquad \|Tx - Ty\|^2 + \frac{1-\alpha}{\alpha}\|(x - Tx) - (y - Ty)\|^2 \leq \|x - y\|^2. \tag{4.4}$$

Note that for a quasi-nonexpansive mapping $T : G \to G$ the condition $x \in \operatorname{Fix} T$ implies that $d(Tx, y) = d(x, y)$ for all $y \in G$. The set of quasi-nonexpansive mappings contains the paracontractions and the nonexpansive mappings. The set of projectors onto convex sets or more generally the set of averaged mappings on a Hilbert space $\mathcal{H}$ is contained in both the set of nonexpansive mappings and the set of paracontractions [7, Remark 4.24 and 4.26]. For an example of a paracontraction that is not averaged see Example 4.0.2 and Appendix B. Averaged mappings were first used in the work of Mann, Krasnoselski, Edelstein, Gurin, Polyak and Raik who wrote seminal papers in the analysis of (firmly) nonexpansive and averaged mappings [20, 22, 31, 38] although the terminology “averaged” wasn’t coined until sometime later [3].

Example 4.0.2. Let $f : \mathbb{R} \to \mathbb{R}_+$ be continuous. Let $f(0) = 0$ and $|f(x)| < |x|$ for all $x \in \mathbb{R} \setminus \{0\}$, then f is paracontractive. This includes also convex functions, e.g. Huber functions, which are not averaged in general (see Appendix B). For other examples on $\mathbb{R}^n$ also see Appendix B.

4.1. RFI on a compact metric space

In this section we establish convergence of the RFI on a compact metric space. The next example illustrates why nonexpansivity alone does not suffice to guarantee convergence to the intersection set C.

Example 4.1.1 (nonexpansive mappings, negative result)**.** For non-expansive mappings in general, one cannot expect that the support of every invariant measure is contained in the feasible set C. Consider a rotation in positive direction in $\mathbb{R}^2$

$$A = \begin{pmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{pmatrix}, \qquad \varphi \in (0, 2\pi),$$

and set $\xi = 1$ and $I = \{1\}$, $T_1 = A$. Then $C = \{0\}$ and, since $\|A\| = 1$, A is nonexpansive, but $\|Ax\| = \|x\|$ for all $x \in \mathbb{R}^2$. So the (deterministic) iteration $X_{k+1} = AX_k$ will not converge to 0, whenever $X_0 \sim \delta_x$, $x \neq 0$.

A sufficient requirement on the mappings T_i to ensure convergence of the RFI is paracontractiveness. The next Lemma is the main ingredient for proving a.s. convergence of (X_k) to a random point in C. The support of a probability measure $\nu \in \mathscr{P}(G)$ is the smallest closed set $S \subset G$, for which $\nu(G \setminus S) = 0$ (see also Theorem 2.4.1 for equivalent representations); we then write $S = \operatorname{supp} \nu$.

Lemma 4.1.2 (invariant measures for paracontractions)**.** *Under the standing assumptions and if T_i $(i \in I)$ is paracontracting on a compact metric space, then the set of invariant measures for $\mathcal{P}$ is $\{\pi \in \mathscr{P}(G) \,|\, \operatorname{supp} \pi \subset C\}$.*

Proof. It is clear that $\pi \in \mathscr{P}(G)$ with $\operatorname{supp} \pi \subset C$ is invariant, since $p(x, \{x\}) = \mathbb{P}(T_\xi x \in \{x\}) = \mathbb{P}(x \in \operatorname{Fix} T_\xi) = 1$ for all $x \in C$ and hence $\pi\mathcal{P}(A) = \int_C p(x, A)\pi(\mathrm{d}x) = \pi(A)$ for all $A \in \mathcal{B}(G)$.

The other implication is not so immediate. Suppose $\operatorname{supp} \pi \setminus C \neq \emptyset$ for some invariant measure π of $\mathcal{P}$. Then due to compactness of $\operatorname{supp} \pi$ (as it is closed in G) we can find $s \in \operatorname{supp} \pi$ maximizing the continuous function $\operatorname{dist}(\cdot, C)$ on G. So $d_{\max} = \operatorname{dist}(s, C) > 0$. We show that the probability mass around s will be attracted to the feasible set C, implying that the invariant measure loses mass around s in every step, which yields a contradiction.

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

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

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

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

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

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

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

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

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

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

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

Since $p(x, K(\epsilon)) = 0$ for $x \in G$ with $\operatorname{dist}(x, C) \leq d_{\max} - \epsilon$ due to paracontractiveness, it holds by invariance of π that

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

It follows, then, that

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

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

Theorem 4.1.3 (almost sure convergence for a compact metric space)**.** *Under the standing assumptions, let T_i be paracontractive, $i \in I$, and let (G, d) be a compact metric space. Then the RFI sequence (X_k) of random variables converges almost surely to a random variable $X_\mu \in C$ depending on the initial distribution μ.*

Proof. Since $\mathcal{P}$ is Feller and G compact, Theorem 2.8.2 implies that any subsequence of (ν_n), where $\nu_n = \frac{1}{n}\sum_{i=1}^n \mathcal{L}(X_i)$, has a convergent subsequence and clusterpoints are invariant measures for $\mathcal{P}$. Let (ν_{n_k}) be a convergent subsequence with limit π. So for the bounded and continuous function $\operatorname{dist}(\cdot, C)$ it holds that $\nu_{n_k} \operatorname{dist}(\cdot, C) \to \pi \operatorname{dist}(\cdot, C) = 0$ as $k \to \infty$ by weak convergence of the probability measures and the fact that, by Lemma 4.1.2, $\operatorname{supp} \pi \subset C$.
Due to quasi-nonexpansiveness and Lemma 3.2.2 (a compact metric space is separable), we have a.s. (for all $\omega \notin N$ with N given by (3.3)) that $d(X_{k+1}, c) \leq d(X_k, c)$ for all $c \in C$ and $k \in \mathbb{N}$, which implies $\operatorname{dist}(X_{k+1}, C) \leq \operatorname{dist}(X_k, C)$ for all $k \in \mathbb{N}$ a.s. It therefore follows that

$$\mathbb{E}[\operatorname{dist}(X_{n_k}, C)] \leq \frac{1}{n_k}\sum_{i=1}^{n_k} \mathbb{E}[\operatorname{dist}(X_i, C)] = \nu_{n_k} \operatorname{dist}(\cdot, C) \to 0$$

by monotonicity of $(\mathbb{E}[\operatorname{dist}(X_k, C)])_k$. This yields $\mathbb{E}[\operatorname{dist}(X_k, C)] \to 0$ as $k \to \infty$. Now since $(\operatorname{dist}(X_k, C))_k$ is nonincreasing, it must be that $\operatorname{dist}(X_k, C) \to 0$ a.s. Hence for any cluster point x_ω of $(X_k(\omega))_k$ we have $x_\omega \in C$. This together with a.s. monotonicity of $(d(X_k, c))_k$ for all $c \in C$ implies that $d(X_k(\omega), x_\omega) \to 0$ for any cluster-point x_ω of $(X_k(\omega))_k$, which implies the uniqueness of x_ω. In other words, (X_k) converges almost surely to a random variable X_μ, with $X_\mu(\omega) = x_\omega \in C$, $\omega \notin N$, as claimed. $\square$

4.2. Finite dimensional normed vector space

The results for compact metric spaces can be applied, with minor adjustments, to finite dimensional vector spaces. In the following let $(G, d) = (V, \|\cdot\|)$ be a finite dimensional normed vector space over $\mathbb{R}$. This means in particular, that V is also complete and every closed and bounded set is compact (Heine-Borel property) and all norms on V are equivalent. So actually, since all n-dimensional vector spaces are isomorphic, it is enough to study convergence in $\mathbb{R}^n$ equipped with the euclidean norm $\|\cdot\|$.

The following result for $\mathbb{R}^n$ is a straight forward application of Theorem 4.1.3.

Theorem 4.2.1 (almost sure convergence in $\mathbb{R}^n$)**.** *Under the standing assumptions, let $T_i : \mathbb{R}^n \to \mathbb{R}^n$ be paracontractive, $i \in I$. Then the RFI sequence (X_k) of random variables converges almost surely to a random variable $X_\mu \in C$ depending on the initial distribution μ.*

Proof. First, suppose $\mu = \delta_x$ for $x \in \mathbb{R}^n$. Let N be given by (3.3). The quasi-nonexpansiveness property gives us $\|X_{k+1} - c\| \leq \|X_k - c\|$ for all $c \in C$ a.s. (i.e. if $\omega \notin N$). Letting $c \in C$ with $\operatorname{dist}(x, C) = \|x - c\|$, this implies $X_k \in \overline{\mathbb{B}}(c, \|x - c\|)$, where $\overline{\mathbb{B}}(s, \epsilon) \subset \mathbb{R}^n$ is the closed ball around $s \in \mathbb{R}^n$ with radius ϵ. The assertion $X_k \to X_{\delta_x}$ a.s. then follows from Theorem 4.1.3. Denote the corresponding invariant measure as $\pi_x := \mathcal{L}(X_{\delta_x})$.

Suppose now that $\mu \in \mathscr{P}(\mathbb{R}^n)$ is arbitrary. For $f \in C_b(\mathbb{R}^n)$ one has $p^k(x, f) \leq \|f\|_\infty$ for all $k \in \mathbb{N}$ and $x \in \mathbb{R}^n$. Note that $p^k(x, f) = \delta_x \mathcal{P}^k f$, and from the above argument

$\delta_x \mathcal{P}^k \to \pi_x$ in the weak sense as $k \to \infty$. Hence by Lebesgue's Dominated Convergence Theorem, we get

$$\mu \mathcal{P}^k f = \int_{\mathbb{R}^n} p^k(x, f)\mu(\mathrm{d}x) \to \int_{\mathbb{R}^n} \pi_x f \mu(\mathrm{d}x) =: \mu\pi_x f =: \pi_\mu f, \qquad \text{as } k \to \infty.$$

We conclude that $\mathcal{L}(X_k) = \mu \mathcal{P}^k \to \pi_\mu$ weakly. The measure $\pi_\mu = \mu\pi_x$ is an invariant probability measure for $\mathcal{P}$, since π_x is a invariant probability measure for $\mathcal{P}$.

Choosing $f = \min\{\operatorname{dist}(\cdot, C), M\} \in C_b(\mathbb{R}^n)$ with $M > 0$ yields $\mathcal{L}(X_k)f \to \pi_\mu f = 0$. Since $f(X_{k+1}) \leq f(X_k)$ a.s. and $\mathcal{L}(X_k)f = \mathbb{E}[f(X_k)] \to 0$, it holds that $f(X_k) \to 0$ a.s. In particular, $\operatorname{dist}(X_k, C) \to 0$ a.s. So for a converging subsequence $(X_{n_k}(\omega))_k$ with limit x_ω it holds that $x_\omega \in C$. Moreover, since $(\|X_k - x_\omega\|)_k$ is monotone, actually $X_k(\omega) \to X_\mu(\omega) := \lim_k X_k(\omega) = x_\omega \in C$, $\omega \notin N$. □

4.3. Weak convergence in Hilbert spaces

In this section (G, d) is a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$. Under the standing assumptions the following extended-valued function

$$R(x) := \mathbb{E}\left[\|x - T_\xi x\|^2\right] = \int_\Omega \left\|x - T_{\xi(\omega)}x\right\|^2 \mathbb{P}(\mathrm{d}\omega) = \int_I \|x - T_u x\|^2 \mathbb{P}^\xi(\mathrm{d}u) \tag{4.5}$$

is measurable from $\mathcal{H}$ to $[0, \infty]$. Following [41] we use this function to characterize convergence of the consistent fixed point problem under the weaker assumption that the mappings T_ξ are *averaged* (see Eq. (4.4)).

Lemma 4.3.1 (properties of R and C for quasi-nonexpansive mappings)**.** *In addition to the standing assumptions, suppose that T_i $(i \in I)$ is quasi-nonexpansive and continuous. Then*

(i) $C = R^{-1}(0)$;

(ii) R is finite everywhere;

(iii) R is continuous;

(iv) C is convex and closed.

Proof. (i) We have $x \in C \Leftrightarrow x \in \operatorname{Fix} T_\xi$ a.s. $\Leftrightarrow x = T_\xi x$ a.s. $\Leftrightarrow R(x) = 0$.

(ii) Fix $x \in C$, then $x = T_\xi x$ a.s. Using quasi-nonexpansivity we get a.s., that

$$\|y - T_\xi y\| \leq \|y - x\| + \|x - T_\xi y\| \leq 2\|x - y\| \qquad \forall y \in \mathcal{H}, \tag{4.6}$$

$$\Longleftrightarrow$$

$$\|y - T_\xi y\|^2 \leq 4\|y - x\|^2 \qquad \forall y \in \mathcal{H}. \tag{4.7}$$

From (4.7) it follows that $R(y) \leq 4\|y - x\|^2 < \infty$ for all $y \in \mathcal{H}$.

(iii) Let $x, x_n \in \mathcal{H}$, $n \in \mathbb{N}$, with $x_n \to x$ as $n \to \infty$. Define the functions $f_n(\omega) = \left\|x_n - T_{\xi(\omega)}x_n\right\|^2$ on Ω $(n \in \mathbb{N})$. Then, by continuity of $T_{\xi(\omega)}$ for fixed $\omega \in \Omega$, one has $f_n \to f := \|x - T_\xi x\|^2$ for all $\omega \in \Omega$. Define the constant function $g(\omega) = 8\epsilon^2 + 8\|x - c\|^2$ for some $c \in C$ and some $\epsilon > 0$. By (4.6) we have that $\|y - T_\xi y\| \leq 2\|y - c\|$ for all $y \in \mathcal{H}$. For $y \in \mathbb{B}(x, \epsilon)$ this yields $\|y - T_\xi y\| \leq 2\epsilon + 2\|x - c\|$. We conclude that g is $\mathbb{P}$-integrable and $f_n \leq g$ for all $n \in \mathbb{N}$ with $x_n \in \mathbb{B}(x, \epsilon)$. Finally, application of Lebesgue's Dominated Convergence Theorem yields $R(x_n) = \mathbb{E}f_n \to \mathbb{E}f = R(x)$ as $n \to \infty$.

(iv) This follows from [7, Proposition 4.13, Proposition 4.14]. Note that for any $\alpha \in \mathbb{R}$, $a, b \in \mathcal{H}$ we have [7, Corollary 2.14]

$$\|\alpha a + (1-\alpha)b\|^2 = \alpha\|a\|^2 + (1-\alpha)\|b\|^2 - \alpha(1-\alpha)\|a - b\|^2.$$

Let $z = \lambda x + (1-\lambda)y$ with $x, y \in R^{-1}(0) = C$, $\lambda \in [0, 1]$. One has with $T_\xi x = x$ and $T_\xi y = y$ a.s. that a.s. holds

$$\begin{aligned}
\|T_\xi z - z\|^2 &= \|\lambda(T_\xi z - x) + (1-\lambda)(T_\xi z - y)\|^2 \\
&= \lambda\|T_\xi z - x\|^2 + (1-\lambda)\|T_\xi z - y\|^2 - \lambda(1-\lambda)\|x - y\|^2 \\
&\leq \lambda\|z - x\|^2 + (1-\lambda)\|z - y\|^2 - \lambda(1-\lambda)\|x - y\|^2 \\
&= \|\lambda(z - x) + (1-\lambda)(z - y)\|^2 \\
&= 0.
\end{aligned}$$

So $R(z) = 0$, i.e. $z \in R^{-1}(0)$. Closedness of $R^{-1}(0)$ follows by continuity of R.

□

In the next theorem we need to compute conditional expectations of nonnegative real-valued random variables, which are non-integrable in general (for example, if the random variable X_0 with distribution μ does not have a finite expectation, $\mathbb{E}[\|X_0\|] = +\infty$). But for these random variables the classical results on integrable random variables are still applicable (see Theorem 2.2.7), also the disintegration theorem is still valid (see Theorem 2.3.2).

The stage is now set to show convergence for the corresponding Markov chain. The next several results concern weak convergence of sequences of random variables with respect to the Hilbert space, namely, $x_n \xrightarrow{\text{w}} x$ if $\langle x_n, y\rangle \to \langle x, y\rangle$ for all $y \in \mathcal{H}$.

Theorem 4.3.2 (weak cluster points belong to feasible set for averaged mappings)**.** *Under the standing assumptions, let T_i be α_i-averaged with $\alpha_i \leq \alpha < 1$ for all $i \in I$. Then weak cluster points (in the sense of Hilbert spaces) of the RFI sequence $(X_k)_{k\in\mathbb{N}_0}$ of random variables in $\mathcal{H}$ are a.s. contained in C.*

Proof. Fix $c \in C$. Since T_ξ is averaged we have for all $k \in \mathbb{N}$ that

$$\|X_{k+1} - c\|^2 \leq \|X_k - c\|^2 - \frac{1-\alpha}{\alpha}\|X_{k+1} - X_k\|^2 \tag{4.8}$$

everywhere but on a $\mathbb{P}$-nullset N_c, which may depend on c. Let $\mathcal{F}_k = \sigma(X_0, \xi_0, \ldots, \xi_{k-1})$ be the σ-algebra of all iterations of the algorithm up to the k-th and apply Lemma A.0.22. We get that $\sum_{k\in\mathbb{N}_0} R(X_k) < \infty$ a.s., where from Theorem 2.3.2 follows that $\mathbb{E}\left[\|X_{k+1} - X_k\|^2 \,\middle|\, \mathcal{F}_k\right] = R(X_k)$. Hence there is $\tilde{N} \subset \Omega$ with $\mathbb{P}(\tilde{N}) = 0$ and $R(X_k(\omega)) \to 0$ as $k \to \infty$ for $\omega \in \Omega \setminus (N_c \cup \tilde{N})$.
By nonexpansiveness of T_ξ for all we find for any $x, x_n \in \mathcal{H}$

$$\begin{aligned}
\|x - T_\xi x\|^2 &= \|x_n - T_\xi x\|^2 + \|x - x_n\|^2 + 2\langle x_n - T_\xi x, x - x_n\rangle \\
&= \|x_n - T_\xi x\|^2 - \|x - x_n\|^2 + 2\langle x - T_\xi x, x - x_n\rangle \\
&= \|x_n - T_\xi x_n\|^2 + \|T_\xi x - T_\xi x_n\|^2 + 2\langle x_n - T_\xi x_n, T_\xi x_n - T_\xi x\rangle \\
&\quad - \|x - x_n\|^2 + 2\langle x - T_\xi x, x - x_n\rangle \\
&\le \|x_n - T_\xi x_n\|^2 + 2\langle x_n - T_\xi x_n, T_\xi x_n - T_\xi x\rangle + 2\langle x - T_\xi x, x - x_n\rangle \\
&\le \|x_n - T_\xi x_n\|^2 + 2\|x_n - T_\xi x_n\|\|x_n - x\| + 2\langle x - T_\xi x, x - x_n\rangle .
\end{aligned}$$

Taking expectation and using Jensen's inequality yields

$$R(x) \le R(x_n) + 2\sqrt{R(x_n)}\|x_n - x\| + 2\mathbb{E}[\langle x - T_\xi x, x - x_n\rangle]. \tag{4.9}$$

Now assume that the sequence (x_n) is weakly converging to $x \in \mathcal{H}$, i.e. $x_n \xrightarrow{\text{w}} x$. Then the functions $f_n = \langle x - T_\xi x, x - x_n\rangle$, $n \in \mathbb{N}$, on Ω satisfy $f_n \to 0$ a.s. Defining the $\mathbb{P}$-integrable function $g(\omega) := \left\|x - T_{\xi(\omega)} x\right\| \sup_n \|x - x_n\|$ gives us $|f_n| \le g$ for all $n \in \mathbb{N}$ and hence by Lebesgue's Dominated Convergence Theorem $\mathbb{E}[\langle x - T_\xi x, x - x_n\rangle] \to 0$ as $n \to \infty$.
So for $\omega \in \Omega \setminus (N_c \cup \tilde{N})$ there is a weakly convergent subsequence of the bounded sequence $(X_k(\omega))_{k\in\mathbb{N}}$, denoted $x_n := X_{k_n}(\omega) \xrightarrow{\text{w}} x_\omega =: x$ as $n \to \infty$. As shown above this subsequence satisfies $R(x_n) \to 0$ as $n \to \infty$. We conclude with (4.9) that $R(x) = 0$, i.e. $x \in C$ and hence any weak cluster point of the sequence $(X_k(\omega))_k$ is contained in C. □

In the case of separable Hilbert spaces, we are able to show Fejér monotonicity (a sequence (x_k) in a Hilbert space $\mathcal{H}$ is Fejér monotone w.r.t. $S \subset \mathcal{H}$, if $\|x_{k+1} - s\| \le \|x_k - s\|$ for all $s \in S$ and $k \in \mathbb{N}$) of the sequence (X_k) a.s., so the classical theory of convergence analysis from [7] can be applied in this case. An analogous statement for nonseparable Hilbert spaces remains open since we do not have the representation Lemma 3.2.2 at hand.

Theorem 4.3.3 (almost sure weak convergence under separability)**.** *Under the same assumptions as in Theorem 4.3.2 assume additionally that $\mathcal{H}$ is a separable Hilbert space. Then the sequence (X_k) is a.s. weakly convergent (in the sense of Hilbert spaces) to a random variable $X_\mu \in C$, depending on the initial distribution μ. Furthermore $P_C X_k \to X_\mu$ strongly a.s. as $k \to \infty$.*

Proof. Instead of a nullset N_c, which may depend on $c \in C$, as in the proof of Theorem 4.3.2, separability gives with help of Lemma 3.2.2 that there is a nullset N, such that on $\Omega \setminus N$ Eq. (4.8) is satisfied for all $c \in C$. This implies a.s. Fejér monotonicity of (X_k). Since from Theorem 4.3.2 follows that weak clusterpoints of (X_k) are contained in

C a.s., we can now apply Theory in [7] developed for Fejér monotone sequences, we get: From [7, Theorem 5.5] (a Fejér monotone sequence w.r.t. C that has all weak clusterpoints in C is weakly convergent to a point in C) follows that $X_k \xrightarrow{w} X_\mu \in C$ a.s.

For strong convergence of $(P_C X_k)$ a.s. we apply [7, Proposition 5.7]. From [7, Corollary 5.8] we get from $X_k \xrightarrow{w} X_\mu$ a.s., that $P_C X_k \to X_\mu$ a.s. strongly as $k \to \infty$. □

Example 4.3.4 (convergence to projection for affine subspaces)**.** Let $\mathcal{H}$ be separable and C_i be an affine subspace, $i \in I$, where I is an arbitrary index set. Let $T_i = P_i$ be the projector onto C_i. Under the standing assumptions holds that $\lim_k X_k = X_\mu = P_C X_0$ for $X_0 \sim \mu$ and any $\mu \in \mathscr{P}(\mathcal{H})$.

We show, that $P_C X_{k+1} = P_C X_k$ for any $k \in \mathbb{N}_0$. This allows us to conclude that $P_C X_k = P_C X_0$ for any $k \in \mathbb{N}_0$, and thus $P_C X_0$ is the only possible weak cluster point of (X_k) by Theorem 4.3.3. Using the characterization [16, Theorem 4.1] (if $K \subset \mathcal{H}$ is nonempty, closed and convex and $u \in K$ then $\langle x - u, k - u \rangle \leq 0$ for all $k \in K$ iff $u = P_K x$) of a projection, we find with help of [16, Theorem 4.9] (for a subspace S holds that $\langle x - P_S x, s \rangle = 0$ for all $s \in S$), that for $c \in C$ holds that

$$\langle X_{k+1} - P_C X_k, c - P_C X_k \rangle = \underbrace{\langle P_{\xi_k} X_k - X_k, c - P_C X_k \rangle}_{=0} + \underbrace{\langle X_k - P_C X_k, c - P_C X_k \rangle}_{\leq 0} \leq 0.$$

Hence by [16, Theorem 4.1] we have that $P_C X_{k+1} = P_C X_k$.

Chapter 5

Geometric Convergence - Consistent Feasibility

We will assume in this section that $\mathcal{H}$ is a separable Hilbert space and T_i is α_i-averaged, $i \in I$. We will furthermore assume, that $\alpha_i \leq \alpha$ for some $\alpha < 1$. As with the deterministic case, geometric convergence of the algorithm can be analyzed by introducing a condition on the set of fixed points. In the context of set feasibility with finitely many sets, the condition is equivalent to *linear regularity* of the sets [42, Assumption 2]: There exists $\kappa > 0$ such that

$$\operatorname{dist}^2(x, C) \leq \kappa R(x) \qquad \forall x \in \mathcal{H}, \tag{5.1}$$

where R is defined by (4.5). In the more general context of fixed point mappings, this property is more appropriately called *global metric subregularity* of R at all points in C for 0 [32]; in particular there exists a $\kappa > 0$ such that

$$\operatorname{dist}^2(x, R^{-1}(0)) \leq \kappa R(x) \qquad \forall x \in \mathcal{H}.$$

Here $C = R^{-1}(0)$, so the above is just another way of writing (5.1). The smallest constant satisfying this inequality will be called the regularity constant, it is given by

$$\sup_{x \in \mathcal{H} \setminus C} \frac{\operatorname{dist}^2(x, C)}{R(x)}.$$

Theorem 5.0.5. *In addition to the standing assumptions, suppose the regularity condition in Eq. (5.1) is satisfied and T_i is α_i-averaged, $i \in I$ with $\alpha_i \leq \alpha$ for some $\alpha < 1$. Then the RFI converges geometrically in expectation to the fixed point set, i.e. for any initial distribution*

$$\mathbb{E}[\operatorname{dist}(X_{k+1}, C)] \leq \sqrt{1 - \kappa^{-1}\frac{1-\alpha}{\alpha}}\,\mathbb{E}[\operatorname{dist}(X_k, C)] \qquad \forall k \in \mathbb{N}_0. \tag{5.2}$$

Proof. Revisiting (4.8) in the proof of Theorem 4.3.2 gives us for $\omega \in \Omega \setminus N$ (N given by (3.3)) and $x = P_C X_k(\omega)$

$$\operatorname{dist}^2(X_{k+1}(\omega), C) \leq \|X_{k+1}(\omega) - x\|^2 \leq \operatorname{dist}^2(X_k(\omega), C) - \frac{1-\alpha}{\alpha}\|X_{k+1}(\omega) - X_k(\omega)\|^2.$$

With help of Jensen's inequality and concavity of $x \mapsto \sqrt{x}$ on $[0, \infty)$, we get that

$$\begin{aligned}\mathbb{E}\left[\operatorname{dist}(X_{k+1}, C) \,|\, \mathcal{F}_k\right] &\leq \mathbb{E}\left[\sqrt{\operatorname{dist}^2(X_k, C) - \frac{1-\alpha}{\alpha}\|T_{\xi_k}X_k - X_k\|^2} \,\middle|\, \mathcal{F}_k\right] \\ &\leq \sqrt{\operatorname{dist}^2(X_k, C) - \frac{1-\alpha}{\alpha}\mathbb{E}\left[\|T_{\xi_k}X_k - X_k\|^2 \,\middle|\, \mathcal{F}_k\right]} \\ &= \sqrt{\operatorname{dist}^2(X_k, C) - \frac{1-\alpha}{\alpha}R(X_k)} \\ &\leq \sqrt{1 - \kappa^{-1}\frac{1-\alpha}{\alpha}}\operatorname{dist}(X_k, C).\end{aligned}$$

□

Note that it could be $\mathbb{E}[\operatorname{dist}(X_k, C)] = \infty$ for all $k \in \mathbb{N}$, depending on the initial distribution μ.

The next theorem concerns the *Wasserstein distance* of two probability measures. For two measures $\nu_1, \nu_2 \in \mathscr{P}(G)$ this is given by

$$W(\nu_1, \nu_2) = \inf_{\substack{Y_1 \sim \nu_1 \\ Y_2 \sim \nu_2}} \mathbb{E}[\|Y_1 - Y_2\|].$$

Theorem 5.0.6 (strong convergence and geometric convergence of measures)**.** *Under the standing assumptions, suppose the regularity condition in Eq. (5.1) is satisfied and T_i is α_i-averaged, $i \in I$ with $\alpha_i \leq \alpha$ for some $\alpha < 1$. Then $X_k \to X$ strongly a.s. as $k \to \infty$ and the Wasserstein distances $W(\mathcal{L}(X_k), \mathcal{L}(X))$ also converge geometricly, there is $r \in (0,1)$ such that*

$$W(\mathcal{L}(X_k), \mathcal{L}(X)) \leq 2r^k W(\mathcal{L}(X_0), \mathcal{L}(X)).$$

Proof. See also [7, Theorem 5.12]. One has a.s. that

$$\|X_k - X_{k+m}\| \leq \|X_k - P_C X_k\| + \|P_C X_k - X_{k+m}\| \leq 2\operatorname{dist}(X_k, C) \leq 2\sqrt{\kappa R(X_k)}.$$

We used here, that T_ξ is nonexpansive and it satisfies $T_\xi c = c$ for any $c \in C$ a.s., hence $\|P_C X_k - X_{k+m}\| = \left\|T_{\xi_{k+m-1}} \cdots T_{\xi_k} P_C X_k - X_{k+m}\right\| \leq \operatorname{dist}(X_k, C)$. This gives us that (X_k) is a Cauchy sequence a.s., since $R(X_k) \to 0$ as seen in the proof of Theorem 4.3.2. Its limit X is contained in C, since its weak limit needs to coincide with the strong limit. Letting $m \to \infty$ one arrives at $\|X_k - X\| \leq 2\operatorname{dist}(X_k, C)$. Taking the expectation yields $\mathbb{E}[\|X_k - X\|] \leq 2\mathbb{E}[\operatorname{dist}(X_k, C)]$. Hence, using Theorem 5.0.5 gives us $\mathbb{E}[\|X_k - X\|] \leq 2r^k\mathbb{E}[\operatorname{dist}(X_0, C)]$ with $r = \sqrt{1 - \kappa^{-1}\frac{1-\alpha}{\alpha}}$ and using the fact that $\mathbb{E}[\operatorname{dist}(X_0, C)] \leq W(\mathcal{L}(X_0), \mathcal{L}(X))$, we have, by the definition of the Wasserstein distance,

$$W(\mathcal{L}(X_k), \mathcal{L}(X)) \leq 2r^k W(\mathcal{L}(X_0), \mathcal{L}(X)).$$

□

Note that it could be $W(\mathcal{L}(X_0), \mathcal{L}(X)) = \infty$, depending on the initial distribution μ.

Remark 5.0.7 (ϵ-fixed point)**:** In order to assure that, with probability greater than $1-\beta$, the k-th iterate is in an ϵ neighborhood of the feasible set C, it is sufficient that $k \geq \ln\left(\frac{\beta\epsilon}{\sqrt{\kappa R(x)}}\right) / \ln(c)$, where $c = \sqrt{1 - \frac{1-\alpha}{\alpha}\kappa^{-1}}$ and $X_0 \sim \delta_x$. To see this, note that, by Markov's inequality,

$$\begin{aligned}
\mathbb{P}(X_k \in C + \epsilon\mathbb{B}(0,1)) &= \mathbb{P}(\operatorname{dist}(X_k, C) < \epsilon) \\
&= 1 - \mathbb{P}(\operatorname{dist}(X_k, C) \geq \epsilon) \\
&\geq 1 - \frac{\mathbb{E}[\operatorname{dist}(X_k, C)]}{\epsilon} \\
&\geq 1 - r^k \frac{\operatorname{dist}(x, C)}{\epsilon} \\
&\geq 1 - r^k \frac{\sqrt{\kappa R(x)}}{\epsilon}.
\end{aligned}$$

Remark 5.0.8: As seen in Example 4.3.4 the probability $\mathbb{P}(X_k \in C)$ can increase to 1 as $k \to \infty$, but this is not necessarily the case, as we will see in Examples 8.1.7 and 8.1.9. There, one finds that $\mathbb{P}(X_k \in C) = \mathbb{P}(X_0 \in C)$ for $k \in \mathbb{N}$. In Example 8.1.8 it holds that $\mathbb{P}(X_k \in C) = \mathbb{P}(X_1 \in C)$ for all $k \in \mathbb{N}$.

Theorem 5.0.9 (necessary and sufficient conditions for geometric convergence)**.** *Under the standing assumptions, let T_i be α_i-averaged, $i \in I$ with $\alpha_i \leq \alpha$ for some $\alpha < 1$. The regularity condition in Eq. (5.1) is satisfied if and only if there exists $r \in [0,1)$ such that*

$$\mathbb{E}[\operatorname{dist}(T_\xi x, C)] \leq r \operatorname{dist}(x, C) \qquad \forall x \in \mathcal{H}. \tag{5.3}$$

Furthermore, condition Eq. (5.1) is necessary and sufficient for geometric convergence in expectation of Algorithm 1 to the fixed point set C as in Eq. (5.2) with a uniform constant for all initial probability measures.

Proof. Eq. (5.1) implies Eq. (5.2), which in turn implies Eq. (5.3) (with $X_0 \sim \delta_x$) by Theorem 5.0.5 with $r = \sqrt{1 - \kappa^{-1}\frac{1-\alpha}{\alpha}}$. The other implication follows the same proof pattern as [36, Theorem 3.11]. We note that, by Theorem 2.3.2, if $X_0 \sim \delta_x$ for $x \in \mathcal{H}$, then

$$\mathbb{E}\left[\|X_1 - X_0\| \,|\, \xi_0\right] = \|T_{\xi_0} x - x\|,$$

hence by Hölder's inequality

$$\mathbb{E}[\|X_1 - X_0\|] \leq \sqrt{R(x)}.$$

Furthermore we can estimate

$$\|X_1 - X_0\| = \|X_1 - P_C X_1 + P_C X_1 - X_0\| \geq \operatorname{dist}(X_0, C) - \operatorname{dist}(X_1, C).$$

Taking the expectation above, the assumption that $\mathbb{E}[\operatorname{dist}(X_1, C)] \leq r\mathbb{E}[\operatorname{dist}(X_0, C)]$ yields

$$(\forall x \in \mathcal{H}) \qquad R(x) \geq (1-r)^2 \operatorname{dist}^2(x, C),$$

i.e. the constant κ in Eq. (5.1) is finite with $\kappa \leq (1-r)^{-2} < \infty$. So Eq. (5.3) implies Eq. (5.1).

For the last implication of the theorem, note that, in case Eq. (5.2) is satisfied with the same constant $r \in (0,1)$ for all Dirac measures δ_x with $x \in \mathcal{H}$, then Eq. (5.3) also holds (letting $X_0 \sim \delta_x$) and hence by the above equivalence Eq. (5.1) is satisfied. This completes the proof. □

Remark 5.0.10: Conventional analytical strategies invoke strong convexity in order to achieve geometric convergence. Our analysis makes no such assumption on the sets C_i. Theorem 5.0.9 shows that geometric convergence is a by-product, mainly, of the regularity of the set of fixed points. The results of [36] indicate that one could formulate a necessary regularity condition for *sublinear* convergence, which also might be useful for stochastic algorithms.

CHAPTER 6

Convergence Analysis - Inconsistent Feasibility

We now analyze convergence of the inconsistent feasibility problem. More exactly, we analyze both the consistent and inconsistent feasibility problem at once, by analyzing the stochastic fixed point problem as formulated in Eq. (3.1). The consistent and inconsistent feasibility problem are then just specializations of this formulation. The next example illustrates how characterization of convergence of Markov chains is more subtle (see the notions of convergence in section 3.4) depending on the contraction properties of the class of mappings defining the RFI.

Example 6.0.11 (nonexpansive mappings, negative result)**.** For non-expansive mappings in general, one cannot expect that the sequence $(\mathcal{L}(X_k))$ converges to an invariant probability measure. Consider a rotation by 180° in $\mathbb{R}^2$, i.e. $A = -\operatorname{Id}$. We have in the RFI setup $\xi = 1$ and $I = \{1\}$, $T_1 = A$. Then A is nonexpansive with $\|Ax\| = \|x\|$ for all $x \in \mathbb{R}^2$. Furthermore, if $X_0 \sim \delta_x$ for $x \neq 0$, then $X_{2k} = x$ and $X_{2k+1} = -x$ for all $k \in \mathbb{N}$. This implies 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 $B \in \mathcal{B}(\mathbb{R}^2)$. Nevertheless the Cesáro average $\nu_n := \frac{1}{n}\sum_{i=1}^n \mathbb{P}^{X_i}$ converges to π_x in the weak sense.

The analysis of Markov chains often separates into the ergodic analysis, i.e. when starting in an invariant (and ergodic) measure, and general convergence theory, when also initial measures which are not invariant are considered.

6.1. Ergodic theory

We understand here under ergodic theory more generally analysis of the properties of the RFI Markov chain when starting it in the support of any ergodic measure for the Markov operator $\mathcal{P}$. 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.

As Example 6.0.11 shows, meaningful notions of ergodic convergence are possible, even when convergence in distribution can not be expected, in our case, convergence of the average. We develop ergodic facts for the Markov chain under consideration for general classes of mappings and at the end specialize to continuous mappings T_i ($i \in I$). We begin with fundamental facts of ergodic chains.

An invariant probability measure π of $\mathcal{P}$ 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. Two measures π_1, π_2 are called *mutually singular* when there is $A \in \mathcal{B}(G)$ with $\pi_1(A^c) = \pi_2(A) = 0$.

Remark 6.1.1 (A different notion of Ergodicity)**:** In many standard literature works ergodicity is introduced in a different manner, which is not quite suitable for what we need. Usually ergodicity is introduced on the product space $G^\infty := \bigtimes_{i=0}^\infty G$. A map $T : G^\infty \to G^\infty$ is called *measure preserving* or μ-preserving, if $\mu \in \mathscr{P}(G^\infty)$ and $\mu(T^{-1}A) = \mu(A)$ for all $A \in \mathcal{B}(G^\infty)$. A measure preserving map T is called *ergodic*, if $\mu(I) \in \{0, 1\}$ for all *invariant sets* $I \in \mathcal{B}(G^\infty)$, i.e. those sets I, which satisfy $T^{-1}I = I$. To see the connection of this ergodicity notation and the one we use, let $\theta : G^\infty \to G^\infty$ be the left-shift, i.e.

$$x = (x_i)_{i\in\mathbb{N}_0} \mapsto \theta(x) = (x_{i+1})_{i\in\mathbb{N}_0}.$$

Now some authors argue that θ is ergodic if and only if $\mu = \mathbb{P}_\pi$, where $\mathbb{P}_\pi$ is the distribution of the sequence $(X_k)_{k\in\mathbb{N}_0}$ with $X_0 \sim \pi$, where π is ergodic and invariant. The problem with that is that θ need not be measure preserving, because $\theta^{-1}(A) = G \times A$ and in general $\mathbb{P}_\pi(G \times A) \neq \mathbb{P}_\pi(A)$, so one has to work with the doubly infinite sequence space $G_{-\infty}^\infty = \bigtimes_{i=-\infty}^\infty G$ instead, to overcome that problem.

In contrast to processes on $\mathbb{N}$, a stochastic process indexed by $\mathbb{Z}$ does not require an initial distribution. That is the case for example, when the process is *stationary*, i.e. when $\theta(X_k)_{k\in\mathbb{Z}} \overset{\mathrm{d}}{=} (X_k)_{k\in\mathbb{Z}}$. One has by [28, Lemma 9.2] that for every stationary process $X_0, X_1, \ldots$ there exist random variables $X_{-1}, X_{-2}, \ldots$ such that $\ldots, X_{-1}, X_0, X_1, \ldots$ is stationary. In particular stationarity implies that $\pi = \mathcal{L}(X_k) = \mathcal{L}(X_{k+1})$ for all $k \in \mathbb{Z}$ is an invariant measure. The left shift is measure preserving for the stationary time-homogeneous Markov process $(X_k)_{k\in\mathbb{Z}}$ with transition kernel p: for the cylindrical sets $A_{n_1,\ldots,n_l} \subset G_{-\infty}^\infty$ with indices $n_i \in \mathbb{Z}$, $i \in \{1, \ldots, l\}$, $l \in \mathbb{N}$ – i.e. there exist $A_1, \ldots, A_l \subset G$ with $a \in A_{n_1,\ldots,n_l}$ exactly when $a_{n_k} \in A_k$ – we have

$$\mathbb{P}_\pi(A) = \pi \otimes p^{n_2-n_1} \otimes \cdots \otimes p^{n_l-n_{l-1}}(A_1 \times \cdots \times A_l),$$

where for two kernels p_i on $G\times\mathcal{F}_i$, $i = 1, 2$ denote $p_1\otimes p_2(s, B) := \int p_1(s, \mathrm{d}t) \int p_2(t, \mathrm{d}u) \mathbb{1}_B(t, u)$ for $B \in \mathcal{F}_1 \otimes \mathcal{F}_2$. One then has

$$\mathbb{P}_\pi(\theta^{-1}A) = \mathbb{P}_\pi(A_{n_1+1,\ldots,n_l+1}) = \mathbb{P}_\pi(A_{n_1,\ldots,n_l}) = \mathbb{P}_\pi(A)$$

that means the shift operator θ is actually invariant for this measure, since the cylindrical sets generate $\bigotimes_{k\in\mathbb{Z}} \mathcal{B}(G)$. Furthermore $\mathbb{P}^{(X_k)_{k\geq 0}} = \mathbb{P}_\pi(G_{-\infty} \times \cdot)$, where $G_{-\infty} = \bigtimes_{i=-\infty}^{-1} G$. One then has the desired connection by [23, Corollary 5.11] that if $\pi \in \mathscr{P}(G)$ is an invariant probability measure for the Markov operator $\mathcal{P}$, then θ is ergodic with respect to $\mathbb{P}_\pi$, if and only if π is ergodic (where P_π is now defined on $G_{-\infty}^\infty$ instead of G^∞).

The following decomposition theorem is key to our development.

Theorem 6.1.2 (Theorem 1.7 in [24])**.** *Given a Markov kernel p, denote by $\mathcal{I}$ the set of all invariant probability measures for $\mathcal{P}$ and by $\mathcal{E} \subset \mathcal{I}$ the set of all those that are ergodic. Then, $\mathcal{I}$ is convex and $\mathcal{E}$ 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 $\mathcal{E}$ such that $\pi(A) = \int_{\mathcal{E}} \nu(A) Q_\pi(\mathrm{d}\nu)$. In other words, every invariant measure is a convex combination of ergodic invariant measures. Finally, any two distinct elements of $\mathcal{E}$ are mutually singular.*

Remark 6.1.3: If there exists only one invariant probability measure of $\mathcal{P}$, we know by Theorem 6.1.2 that it is ergodic. If there exist more invariant probability measures, then there exist uncountably many invariant and at least two ergodic ones.

Theorem 6.1.4 (Birkhoff's ergodic theorem, Theorem 9.6 in [28])**.** *Let π be an ergodic invariant probability measure for $\mathcal{P}$, and let $(G, \mathcal{G})$ be a measure space, $f : G \to \mathbb{R}$ be such that $\pi|f|^p < \infty$ and $p \in [1, \infty]$, then*

$$\frac{1}{n} \sum_{i=1}^{n} f(X_i) \to \pi f, \qquad \text{a.s. and in } L_p \text{ as } n \to \infty,$$

where the sequence (X_n) is generated by Algorithm 1 with $X_0 \sim \pi$.

Corollary 6.1.5. *Same assumptions as in Theorem 6.1.4. Let $f : G \to \mathbb{R}$ be measurable and bounded, i.e. $\|f\|_\infty := \sup_{x \in G} |f| < \infty$, then*

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

where $p^i(x, f) := \delta_x \mathcal{P}^i f$.

Proof. Let $Z, Z_1, Z_2, \ldots$ be real-valued random variables with $Z_n \to Z$ a.s. as $n \to \infty$ and $|Z_n| \le Y$ where $\mathbb{E}|Y| < \infty$. Then $\mathbb{E}[Z_n \,|\, \mathcal{F}_0] \to \mathbb{E}[Z \,|\, \mathcal{F}_0]$ a.s. as $n \to \infty$ for any sub-σ-algebra $\mathcal{F}_0 \subset \mathcal{F}$ (c.f. [29, Satz 8.14 (viii)]). For a bounded function f (so $\pi|f| \le \|f\|_\infty$), the statement follows by application of this fact with $Y = \|f\|_\infty$, $Z_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$, $Z = \pi f$ and $\mathcal{F}_0 = \sigma(X_0)$ from Theorem 6.1.4. □

To get weak convergence of the Cesáro average ν_n^x to π from Birkhoff's Theorem, the fact in Corollary 6.1.5 needs to be strengthened such that the limit is well-defined for all $f \in C_b(G)$ except on a π-nullset. This is done in the following for compact metric spaces and the Euclidean space $\mathbb{R}^n$.

Corollary 6.1.6 (Weak convergence of (ν_n^x))**.** *Let (G, d) be a compact metric space and let π be an ergodic invariant probability measure for $\mathcal{P}$. Then for π-a.e. $x \in G$ the sequence $\nu_n^x \to \pi$ as $n \to \infty$, where $\nu_n^x = \frac{1}{n}\sum_{i=1}^{n} p^i(x, \cdot)$.*

Proof. Our proof follows the pattern of [33, Theorem 1.1, Chapter 3]. By Corollary 6.1.5, we have for $f \in C_b(G)$ a.s. that

$$\nu_n^{X_0} f = \frac{1}{n} \sum_{i=1}^{n} p^i(X_0, f) \to \pi f, \qquad n \to \infty, \tag{6.1}$$

where $X_0 \sim \pi$. Since $(C_b(G), \|\cdot\|_\infty) = (C(G), \|\cdot\|_\infty)$ is separable by compactness of G, there exists a countable dense subset $(g_k)_{k\in\mathbb{N}} \subset C_b(G)$. Let $N_k \subset \Omega$ be the $\mathbb{P}$-nullset, where (6.1) is not satisfied for g_k. Define the $\mathbb{P}$-nullset $N = \bigcup_k N_k$, it holds for $\omega \in \Omega \setminus N$ that

$$\nu_n^{X_0(\omega)} g_k \to \pi g_k, \qquad n \to \infty, \quad \forall k \in \mathbb{N}.$$

Let $f \in C_b(G)$, then we want to show that also $\nu_n^x f \to \pi f$ for $x \in X_0(\Omega \setminus N)$. Let $\epsilon > 0$. By denseness of $(g_k) \subset C_b(G)$ there is $m \in \mathbb{N}$ with $\|f - g_m\|_\infty < \epsilon$. One has that

$$\begin{aligned} |\nu_n^x f - \pi f| &\le |\nu_n^x f - \nu_n^x g_m| + |\pi f - \pi g_m| + |\nu_n^x g_m - \pi g_m| \\ &\le 2\|f - g_m\|_\infty + |\nu_n^x g_m - \pi g_m| \\ &< 3\epsilon \end{aligned}$$

for n large enough. □

Corollary 6.1.7 (Weak convergence of (ν_n^x) in $\mathbb{R}^n$)**.** *Let $(G, d) = (\mathbb{R}^n, \|\cdot\|)$ and let π be an ergodic invariant probability measure for $\mathcal{P}$. Then for π-a.e. $x \in \mathbb{R}^n$ the sequence $\nu_n^x \to \pi$ as $n \to \infty$, where $\nu_n^x = \frac{1}{n}\sum_{i=1}^n p^i(x, \cdot)$.*

Proof. We proceed similar to the proof above. We know that the metric space $(C_c(\mathbb{R}^n), \|\cdot\|_\infty)$ of compactly supported continuous functions with the supremum-norm is separable. So for $f \in C_c(\mathbb{R}^n)$ holds a.s. that

$$\nu_n^{X_0} f \to \pi f, \qquad n \to \infty,$$

where $X_0 \sim \pi$. Defining the $\mathbb{P}$-nullset N as above and using the 3ϵ-argument then gives for π-a.e. $x \in \mathbb{R}^n$ that

$$\nu_n^x f \to \pi f, \qquad n \to \infty, \quad \forall f \in C_c(\mathbb{R}^n).$$

Choosing $f \in C_c(\mathbb{R}^n)$ to be a smoothed indicator function of a ball, the tightness of (ν_n^x) can be shown, implying that

$$\nu_n^x f \to \pi f, \qquad n \to \infty, \quad \forall f \in C_b(\mathbb{R}^n).$$

Let $\epsilon > 0$ and $M > 0$ such that $\pi(\overline{\mathbb{B}}(0, M)) > 1 - \epsilon$. Let ϕ_δ be 1 on $\overline{\mathbb{B}}(0, M)$ and 0 outside of $\overline{\mathbb{B}}(0, M + \delta)$ and else continuous and nonnegative. Since $\phi_\delta \in C_c(\mathbb{R}^n)$, one has that $\nu_n^x \phi_\delta \to \pi \phi_\delta \ge \pi(\overline{\mathbb{B}}(0, M)) > 1 - \epsilon$, so there is N such that for all $n \ge N$ holds that $\nu_n(\overline{\mathbb{B}}(0, M + \delta)) > 1 - \epsilon$. After possibly making δ larger (but still finite) this is true for all n. Tightness implies the existence of weakly convergent subsequences of (ν_n^x), but all possible clusterpoints coincide with π on $C_c(\mathbb{R}^n)$ and hence on all compact boxes of $\mathbb{R}^n$ and then by regularity the clusterpoint is unique. □

An open question for us is if continuity of the mappings T_i $(i \in I)$ is already enough to get from π-a.e. x to the statement that ν_n^x converges to π for all $x \in \operatorname{supp}\pi$. If it was true, this could simplify Theorem 6.2.3 for the case that (G,d) is the Euclidean space $\mathbb{R}^n$. If it was not true, then a counter example would be nice to see, which could not be found yet.

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 $\mathcal{P}$. In the results that follow, we assume continuity of T_i. Lemma 2.8.1 could one make think that the support of any invariant measure is invariant under T_ξ, i.e. $T_\xi \operatorname{supp}\pi \subset \operatorname{supp}\pi$, but this need not be the case for Markov operators generated from discontinuous mappings T_i. Indeed, 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. $T = \mathrm{Id}$, we have that $\mu\mathcal{P}^k = \mu$ for all $k \in \mathbb{N}$. Consequently, $\pi = \mu$ is invariant and $S_\pi = [0,1]$, but $T([0,1]) = \{-1\} \cup [0,1] \cap (\mathbb{R}\setminus\mathbb{Q})$, which is not contained in $[0,1]$.

Lemma 6.1.8 (invariance of the support of invariant measures)**.** *Let (G,d) be a Polish space and let $T_i : G \to G$ be continuous for all $i \in I$. Let $S_\pi := \operatorname{supp}\pi$ 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.*

Proof. We can write with Fubini for any $A \in \mathcal{B}(G)$ that

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

Since $1 = \pi(S_\pi) = \mathbb{E}\pi(T_\xi^{-1}S_\pi)$ and $\pi(\cdot) \le 1$, we find that $\pi(T_\xi^{-1}S_\pi) = 1$ a.s. We have that $T_i^{-1}S_\pi$ is closed for all $i \in I$ due to continuity of T_i and closedness of S_π, hence it must hold that $S_\pi \subset T_\xi^{-1}S_\pi$ a.s. by Theorem 2.4.1. (v), i.e. $T_\xi S_\pi \subset S_\pi$ a.s. □

This Lemma 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. E.g. 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 $T_2 S_\pi = \{1\}$.

Corollary 6.1.9 (characterization of support)**.** *Under the assumptions of Lemma 6.1.8, we have that*

$$S_\pi = \overline{\bigcup_{x \in S_\pi} \operatorname{supp}\mathcal{L}(T_\xi x)}.$$

Proof. From Lemma 6.1.8 it is clear that there exists a $\mathbb{P}$-nullset $N \subset \Omega$ such that $T_{\xi(\Omega\setminus N)}x \subset S_\pi$ for all $x \in S_\pi$. Hence by Lemma 2.4.2 $\operatorname{supp}\mathcal{L}(T_\xi x) \subset S_\pi$ for all $x \in S_\pi$. Would there on the other hand exist $x \in S_\pi \setminus \overline{\bigcup_{y\in S_\pi} \operatorname{supp}\mathcal{L}(T_\xi y)}$, then one can find $\epsilon_0 > 0$ such that for all $\epsilon \in (0,\epsilon_0)$ $\mathbb{P}(T_\xi y \in \mathbb{B}(x,\epsilon)) = 0$ for all $y \in S_\pi$, which is a contradiction to invariance $\pi(\mathbb{B}(x,\epsilon)) = \pi\mathcal{P}(\mathbb{B}(x,\epsilon))$. □

In the following we will denote $S_\pi := \operatorname{supp}\pi$ and

$$S := \bigcup_{\pi\in\mathcal{E}} S_\pi, \tag{6.2}$$

where $\mathcal{E}$ is the set of ergodic measures, see Theorem 6.1.2.

6.2. Ergodic theory for nonexpansive mappings

Our subsequent developments make heavy use of the following important fact in [51] about tightness of the sequence of the iterated kernel $(p^k(s,\cdot))$, where $s \in S$.

Theorem 6.2.1 (tightness of $(\delta_s\mathcal{P}^k)$)**.** *Let (G,d) be a Polish space. Let $T_i : G \to G$ be nonexpansive, $i \in I$. Suppose there exists an invariant measure for $\mathcal{P}$. Then $(\delta_s\mathcal{P}^k)$ is tight for all $s \in S$ defined by (6.2).*

Proof. We will apply [51, Proposition 2.1]. Therefore, we need to ensure, that we have equicontinuity of the sequence of functions $(x \mapsto \delta_x P^k f)$ for any Lipschitz continuous $f : G \to \mathbb{R}$: Let $\epsilon > 0$ and $x, y \in G$ with $d(x,y) < \epsilon/\|f\|_{\mathrm{Lip}}$, then, using nonexpansivity, we get

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

for all $k \in \mathbb{N}$.
Furthermore, letting $f = \mathbb{1}_{\mathbb{B}(s,\epsilon)}$ for some $s \in S_\pi$, where $\pi \in \mathcal{E}$ and $\epsilon > 0$ in Theorem 6.1.4 shows that

$$\limsup_{n\to\infty} \nu_n^x(\mathbb{B}(s,\epsilon)) = \lim_n \nu_n^x(\mathbb{B}(s,\epsilon)) = \pi(\mathbb{B}(s,\epsilon)) > 0 \qquad \text{for } \pi\text{-a.e. } x \in G,$$

where $\nu_n^x = \frac{1}{n}\sum_{i=1}^n \delta_x\mathcal{P}^i$. So, the assumptions in [51, Proposition 2.1] for $(\delta_s\mathcal{P}^k)$ to be tight are met. □

Remark 6.2.2 (tightness of (ν_n^s))**:** Note that (ν_n^s) is tight for $s \in S$, since by Theorem 6.2.1, 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 $\nu_n^s(K) > 1-\epsilon$ for all $n \in \mathbb{N}$.

Theorem 6.2.3 (weak convergence of (ν_n^s) on Polish spaces)**.** *Let (G,d) be a Polish space and let $T_i : G \to G$, $i \in I$ be nonexpansive. Let π be an ergodic invariant probability measure for $\mathcal{P}$. Then for all $s \in S_\pi := \operatorname{supp}\pi$ the sequence $\nu_n^s \to \pi$ as $n \to \infty$, where $\nu_n^s = \frac{1}{n}\sum_{i=1}^n p^i(s,\cdot)$.*

Proof. Let $f \in C_b(G)$. Then by Corollary 6.1.5 there is a nullset $N = N(f)$ (depending on f) such that for all $x \in X_0(\Omega \setminus N)$ holds that $\nu_n^x f \to \pi f$. Since $X_0 \sim \pi$ one has that $\overline{X_0(\Omega \setminus N)} = S_\pi$ and hence for any $\epsilon > 0$ and any $s \in S_\pi$ there exists $\tilde{s} = \tilde{s}(f, s, \epsilon) \in S_\pi$ with $d(s, \tilde{s}) < \epsilon$ and $\nu_n^{\tilde{s}} f \to \pi f$ as $n \to \infty$. Let ν be a cluster point of (ν_n^s), i.e. $\nu_{n_l}^s \to \nu$. Then for the Kantorovich-Rubinstein or also called Fortet-Mourier metric (see also [10, Section 8.3]) holds that

$$\begin{aligned}
d_0(\nu, \pi) &= \sup\left\{\nu f - \pi f \mid f \in \mathrm{Lip}_1(G),\, \|f\|_\infty \le 1\right\} \\
&= \sup\left\{\lim_l \nu_{n_l}^s f - \nu_{n_l}^{\tilde{s}(f,s,\epsilon)} f \,\middle|\, f \in \mathrm{Lip}_1(G),\, \|f\|_\infty \le 1\right\} \\
&\le \sup\left\{\lim_l \frac{1}{n_l}\sum_{i=1}^{n_l}\int \left|f(X_i^s) - f(X_i^{\tilde{s}(f,s,\epsilon)})\right| \mathrm{d}\mathbb{P} \,\middle|\, f \in \mathrm{Lip}_1(G),\, \|f\|_\infty \le 1\right\} \\
&\le \sup\left\{d(s, \tilde{s}(f,s,\epsilon)) \mid f \in \mathrm{Lip}_1(G),\, \|f\|_\infty \le 1\right\} \\
&\le \epsilon
\end{aligned}$$

for all $\epsilon > 0$, where $\mathrm{Lip}_1(G) := \{f : G \to \mathbb{R} \mid |f(x) - f(y)| \le d(x,y)\, \forall x, y \in G\}$, which means $d_0(\nu, \pi) = 0$, i.e. $\nu = \pi$. □

Remark 6.2.4: In Theorem 6.1.2 the decomposition of any invariant measure into a convex combination of ergodic invariant measures was stated and in particular two ergodic measures π_1, π_2 were mutually singular. Note that still it could be that $\operatorname{supp}\pi_1 \cap \operatorname{supp}\pi_2 \neq \emptyset$. But Theorem 6.2.3 establishes that for nonexpansive mappings T_i, $i \in I$ this is not possible, so the singularity of ergodic measures extends to their support. In particular for two ergodic measures $\pi, \tilde{\pi}$ holds $S_\pi \cap S_{\tilde{\pi}} = \emptyset$ if and only if $\pi \neq \tilde{\pi}$.
This seemingly simple property that two ergodic measures are mutually singular with respect to their support has tremendous implications for the ergodic behavior of the Markov chain generated by the RFI algorithm. In particular it gives us uniqueness of invariant measures on the metric space (S_π, d).

Corollary 6.2.5. *Under the assumptions of Theorem 6.2.3 for two ergodic measures $\pi, \tilde{\pi}$ it holds that $S_\pi \cap S_{\tilde{\pi}} = \emptyset$ if and only if $\pi \neq \tilde{\pi}$.*

Theorem 6.2.6 (convergence on S)**.** *Let (G, d) be a Polish space and $T_i : G \to G$ be nonexpansive, $i \in I$. Let $\mu \in \mathscr{P}(S)$ with $S \neq \emptyset$, i.e. we assume there exists an invariant probability measure π for $\mathcal{P}$. Then $\nu_n := \frac{1}{n}\sum_{i=1}^n \mu\mathcal{P}^i$ converges to an invariant probability measure for $\mathcal{P}$.*

Proof. The case $\mu = \delta_x$ with $x \in S$ follows easily from Theorem 6.2.3 and Remark 6.2.4 which ensure the existence of a unique ergodic invariant measure π_x with $x \in S_{\pi_x} = \operatorname{supp}\pi_x$ and $\nu_n^x \to \pi_x$.

For the general case, let $\mu \in \mathscr{P}(S)$ be arbitrary. One finds for $f \in C_b(G)$ that

$$\nu_n f = \tfrac{1}{n}\sum_{i=1}^n \mu\mathcal{P}^i f = \int \tfrac{1}{n}\sum_{i=1}^n \delta_x \mathcal{P}^i f \mu(\mathrm{d}x) = \int \nu_n^x f \mu(\mathrm{d}x).$$

From the special case established above we have $\nu_n^x \to \pi_x$ as $n \to \infty$, so that application of Lebesgue's Dominated Convergence Theorem yields

$$\int \nu_n^x f \mu(\mathrm{d}x) \to \int \pi_x f \mu(\mathrm{d}x), \qquad n \to \infty.$$

Using a loose notation the measure $\mu\pi_x(\cdot) := \int \pi_x(\cdot)\mu(\mathrm{d}x)$ is invariant by invariance of π_x and $\nu_n f \to \mu\pi_x f$ as $n \to \infty$. □

Remark 6.2.7: Theorem 6.2.6 only establishes convergence of the Markov chain when it is initialized with a measure in the support of an ergodic invariant measure; moreover, it is only the Cesáro average of the distributions of the iterates that converges.

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 6.2.8 (positive transition probability for ergodic measures)**.** *Let (G, d) 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 $s, \tilde{s} \in S_\pi$ it holds that*

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

Proof. Given $\tilde{s} \in S_\pi$ 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 $s \in S_\pi$ let $X_0 \sim \delta_s$ and (X_k) generated by Algorithm 1. By Theorem 6.2.6 (ν_n) converges to π, where $\nu_n := \frac{1}{n}\sum_{i=1}^n p^i(s, \cdot)$. So in particular $\nu_n f \to \pi f \geq \pi(\mathbb{B}(\tilde{s}, \frac{\epsilon}{2})) > 0$ as $n \to \infty$. Hence, for n large enough there is $\delta > 0$ with

$$\nu_n f = \tfrac{1}{n}\sum_{i=1}^{n} p^i(s, f) \geq \delta.$$

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

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

□

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. Denote by $C(\mu, \nu)$ the set of all couplings for μ and ν, i.e. probability measures on $G \times G$ with marginals μ and ν.

Lemma 6.2.9 (distance of supports is attained)**.** *Let (G, d) be a Polish space and $T_i : G \to G$ be nonexpansive, $i \in I$. Suppose $\pi, \tilde{\pi}$ are ergodic probability measures for $\mathcal{P}$. 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 6.1.8 it holds a.s. that

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

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 6.2.8 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

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

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_n^m on $G \times G$ via

$$\gamma_n^m f := \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n f(X_i^s, X_i^{\tilde s_m})\right]$$

for measurable $f : G \times G \to \mathbb{R}$. Then $\gamma_n^m \in C(\nu_n^s, \nu_n^{\tilde s_m})$ and by Lemma 2.6.3 and Theorem 6.2.3 $(\gamma_n^m)_n$ is tight for fixed $m \in \mathbb{N}$ and there exists a clusterpoint $\gamma^m \in C(\pi, \tilde\pi)$. The sequence $(\gamma^m) \subset C(\pi, \tilde\pi)$ is again tight by Lemma 2.6.3 and hence for any clusterpoint $\gamma \in C(\pi, \tilde\pi)$ holds for the bounded and continuous function $(x, y) \mapsto f^M(x, y) = \min(M, d(x, y))$ that

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

for all $M \ge d(s, \tilde s_m)$, $m \in \mathbb{N}$. Since by the Monotone Convergence Theorem $\gamma^m f^M \nearrow \gamma^m d$ as $m \to \infty$, it follows $\gamma^m f^M = \gamma^m d$ for all $M \ge d(s, \tilde s_1)$. By the same argument holds for $M \ge d(s, \tilde s_1)$ and a subsequence (γ^{m_k}) with limit γ that $\gamma d = \gamma f^M$. Hence,

$$\gamma d = \gamma f^M = \lim_k \gamma^{m_k} f^M = \lim_k \gamma^{m_k} d \le \lim_k d(s, \tilde s_{m_k}) = \operatorname{dist}(s, S_{\tilde\pi}).$$

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 $d(x, y) \ge \operatorname{dist}(S_\pi, S_{\tilde\pi})$ 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 2.6.2. □

As a consequence of Lemma 6.2.9 one can show that S is closed.

Lemma 6.2.10 (S is closed)**.** *Let (G, d) be a Polish space and let $T_i : G \to G$ be nonexpansive, $i \in I$. Then S is closed, and $S = \bigcup_{\pi \in \operatorname{inv} \mathcal{P}} S_\pi$.*

Proof. Let $(s_m) \subset G$ be a sequence with $s_m \in \pi_m \in \mathcal{E}$ and $s_m \to s$. We have to show that $s \in S$. The representation of $S = \bigcup_{\pi \in \operatorname{inv} \mathcal{P}} S_\pi$ follows from closedness of S and Theorem 6.1.2.

First we show tightness of $(\delta_s\mathcal{P}^k)$. From Theorem 6.2.1 we have that $(\delta_{s_m}\mathcal{P}^k)_{k\in\mathbb{N}}$ is tight for any $m \in \mathbb{N}$. Fix $\epsilon > 0$ and $m \in \mathbb{N}$ with $d(s, s_m) \leq \epsilon$. According to [34, Chp. 3, Theorem 2.2] there exists a compact set $K \subset G$ such that

$$\delta_{s_m}\mathcal{P}^k(\mathbb{B}(K,\epsilon)) \geq 1-\epsilon \qquad \forall k \in \mathbb{N}.$$

Due to nonexpansiveness of T_i, $i \in I$ we have with the notation $X_k^x := T_{\xi_{k-1}} \cdots T_{\xi_0}x$ for $x \in G$ that

$$\operatorname{dist}(X_k^s, \mathbb{B}(K,\epsilon)) \leq d(X_k^s, X_k^{s_m}) + \operatorname{dist}(X_k^{s_m}, \mathbb{B}(K,\epsilon)) \leq 2\epsilon$$

for all $k \in \mathbb{N}$. Hence,

$$\delta_s\mathcal{P}^k(\mathbb{B}(K,2\epsilon)) \geq \delta_{s_m}\mathcal{P}^k(\mathbb{B}(K,\epsilon)) \geq 1-\epsilon \qquad \forall k \in \mathbb{N}.$$

By [34, Chp. 3, Theorem 2.2] that implies tightness of $(\delta_s\mathcal{P}^k)$ and hence of the Cesáro average (ν_n^s).
Let ν be a clusterpoint of (ν_n^s) such that $\nu_{n_k}^s \to \nu$. Then

$$d_0(\nu, \pi^{s_m}) = \lim_k d_0(\nu_{n_k}^s, \nu_{n_k}^{s_m}) \leq d(s, s_m).$$

Hence, $\pi^{s_m} \to \nu$ as $m \to \infty$, which implies that ν is the unique limit of (ν_n^s) and also by the Feller property of $\mathcal{P}$ that ν is invariant.
We show now ergodicity of ν. By Theorem 6.1.2 we have that $\nu(S) = 1$. Since by Theorem 2.4.1 (iii) it holds that $1 = \nu(S) = \nu(S \cap S_\nu)$, there exists $\tilde{s} \in S_\nu \cap S$. W.l.o.g. let $\tilde{s} \in S_{\tilde{\pi}}$ for some $\tilde{\pi} \in \mathcal{E}$. We assume from now on that $s \notin S$ and lead that to a contradiction, yielding the desired result. Let $\epsilon < \operatorname{dist}(s, S_{\tilde{\pi}})/2$, where the latter expression is positive, since $S_{\tilde{\pi}}$ is closed and $s \notin S$. Fix $m \in \mathbb{N}$ such that $d(s, s_m) < \epsilon$. From Lemma 6.1.8 we can find $\tilde{s}_m \in S_{\tilde{\pi}}$ such that $\operatorname{dist}(S_{\pi^{s_m}}, S_{\tilde{\pi}}) = d(s_m, \tilde{s}_m)$. Then,

$$\begin{aligned}
\operatorname{dist}(X_k^s, S_{\tilde{\pi}}) &\geq \operatorname{dist}(X_k^{s_m}, S_{\tilde{\pi}}) - d(X_k^s, X_k^{s_m}) \\
&\geq \operatorname{dist}(S_{\pi^{s_m}}, S_{\tilde{\pi}}) - d(s, s_m) \\
&= d(s_m, \tilde{s}_m) - d(s, s_m) \\
&\geq d(s, \tilde{s}_m) - 2d(s, s_m) \\
&\geq \operatorname{dist}(s, S_{\tilde{\pi}}) - 2\epsilon > 0.
\end{aligned}$$

This is a contradiction to the fact that $\tilde{s} \in S_\nu \cap S_{\tilde{\pi}}$. So, we have that $s \in S$, and hence S is closed. □

6.3. General convergence theory for nonexpansive mappings

When the Markov chain is initialized with a point not supported in S, i.e. we allow $\mu \in \mathscr{P}(G)$, the convergence results on general Polish spaces are much weaker than for the ergodic case in the previous section. One major problem is that the sequences (ν_n^x) for $x \in G \setminus S$ need not be tight anymore. The right-shift operator θ on l^2, for example, with the initial distribution δ_{e_1}, generates the sequence $\theta^k e_1 = e_k$. Spaces, on which we can guarantee tightness are of course compact metric spaces, since then $(\mathscr{P}(G), d_P)$ is compact, i.e. the space of probability measures equipped with the Prokhorov-Levi metric (see Theorem 2.5.4) is a metric space and metrizes convergence in the weak sense. For Euclidean space $\mathbb{R}^n$ one also has tightness, though this requires some justification.

Lemma 6.3.1 (tightness of $(\mu\mathcal{P}^k)$ in $\mathbb{R}^n$)**.** *Let (G,d) be the Euclidean space $(\mathbb{R}^n, \|\cdot\|)$ and $T_i : G \to G$ be nonexpansive, $i \in I$ and let $\operatorname{inv}\mathcal{P} \neq \emptyset$ for the corresponding Markov operator. The sequence $(\mu\mathcal{P}^k)$ is tight for any $\mu \in \mathscr{P}(G)$.*

Proof. First, let $\mu = \delta_x$ for $x \in G = \mathbb{R}^n$. We know, that $(\delta_s\mathcal{P}^k)$ is tight for $s \in S$ by Theorem 6.2.1. So for $\epsilon > 0$ there is a compact $K \subset \mathbb{R}^n$ with $p^k(s,K) \geq 1-\epsilon$ for all $k \in \mathbb{N}$. Since a.s. holds $d(X_k^x, X_k^s) \leq d(x,s)$, 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 any $\mu \in \mathscr{P}(G)$. For given $\epsilon > 0$ there is a compact $K_\epsilon^\mu \subset \mathbb{R}^n$ with $\mu(K_\epsilon^\mu) > 1-\epsilon$. From the special case established above, there exists a compact $K_\epsilon \subset \mathbb{R}^n$ 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 $\|X_k^x - X_k^0\| \leq \|x\| \leq M$. Hence $\mu\mathcal{P}^k(\overline{\mathbb{B}}(K_\epsilon, M)) > (1-\epsilon)^2$, which implies tightness of $(\mu\mathcal{P}^k)$. □

Remark 6.3.2 (tightness of (ν_k^μ) in $\mathbb{R}^n$)**:** The tightness of (ν_k^μ) for any $\mu \in \mathscr{P}(\mathbb{R}^n)$ follows immediately from tightness of $(\mu\mathcal{P}^k)$ as in Remark 6.2.2.

Lemma 6.3.3 (properties for nonexpansive mappings)**.** *Let (G,d) be a Polish space and $T_i : G \to G$ be nonexpansive, $i \in I$. 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.*

(i) If $\pi \in \operatorname{inv}\mathcal{P}$ then $\operatorname{dist}(X_{k+1}, S_\pi) \leq \operatorname{dist}(X_k, S_\pi)$ a.s. for all $k \in \mathbb{N}$.

Denote $\nu_n := \frac{1}{n}\sum_{i=1}^n \mu\mathcal{P}^i$ and assume that the sequence (ν_n) has a clusterpoint π. Then,

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

(iii) clusterpoints of (ν_n) have the same support.

(iv) clusterpoints of $(\mu\mathcal{P}^k)$ are in $\mathscr{P}(S_\pi)$ (if they exist).

Proof. (i) By Lemma 6.1.8, the sets $N_k \subset \Omega$ on which not holds $T_{\xi_k} S_\pi \subset S_\pi$ are $\mathbb{P}$-nullsets and so is their union, denoted as N. So, except for $\omega \in N$ holds for all $s \in S_\pi$

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

and hence

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

(ii) Since the function $f = \min(M, \operatorname{dist}(\cdot, S_\pi))$ for some $M > 0$ is bounded and continuous, we have for a subsequence (ν_{n_k}) of (ν_n) converging to π, that $\nu_{n_k} f = \frac{1}{n_k}\sum_{i=1}^{n_k} \mu\mathcal{P}^i f \to \pi f = 0$ as $k \to \infty$. Now ((i)) and

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

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

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

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

(iii) Let π_1, π_2 be two clusterpoints of (ν_n) with support S_1, S_2 respectively, then these probability measures are invariant for $\mathcal{P}$ by Theorem 2.8.2. By Corollary 6.2.5 $S_1 \cap S_2 = \emptyset$ is not possible, so $S_1 \cap S_2 \neq \emptyset$. 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_n) with $\nu_{n_k} f \to \pi_1 f$ and $\nu_{\tilde{n}_k} f \to \pi_2 f$ as $k \to \infty$. For the former sequence we have, for k large enough,

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

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

(iv) Let ν be a clusterpoint of $(\mu\mathcal{P}^k)$, 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 ((ii)) we find, that

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

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

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

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

We show now the convergence of the Cesáro average (ν_n) of $(\mu\mathcal{P}^k)$, where $\nu_n := \frac{1}{n}\sum_{k=1}^n \mu\mathcal{P}^k$, to an invariant probability measure for $\mathcal{P}$ for arbitrary initial measure. We restrict ourselves to Polish spaces with finite dimensional metric in order to apply a differentiation theorem. We begin with the next technical fact.

Lemma 6.3.4 (characterization of balls in $(\mathcal{E}, d_P)$). *Let G be a Polish space and $T_i : G \to G$ be nonexpansive, $i \in I$. Let $\pi, \tilde{\pi} \in \mathcal{E}$, then*

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

for $\epsilon \in (0, 1)$.

Proof. By Lemma 6.2.9 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 6.2.5, 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 2.6.2((i)) and Lemma 6.1.8, $\operatorname{supp} \mathcal{L}(X_k^s) \subset S_\pi$ 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 $k \in \mathbb{N}$. From nonexpansiveness of T_i, $i \in I$, 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

$$d(X_k^s, X_k^{\tilde{s}}) = d(s, \tilde{s}) \qquad \text{a.s. } \forall k \in \mathbb{N}. \tag{6.3}$$

Now, letting $c := \min(1, d(s, \tilde{s}))$, we show that $d_P(\pi, \tilde{\pi}) = c$, where d_P denotes the Prokhorov-Levi metric (see Theorem 2.5.4). Therefore take $(X, Y) \in C(\mathcal{L}(X_k^s), \mathcal{L}(X_k^{\tilde{s}}))$. Again, by Lemma 2.6.2((i)) and Lemma 6.1.8 $\operatorname{supp} \mathcal{L}(X) \subset S_\pi$ 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 Theorem 2.5.4(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 (6.3)

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

Altogether we find that $d_P(\mathcal{L}(X_k^s), \mathcal{L}(X_k^{\tilde{s}})) = c$ by Theorem 2.5.4((i)). Since also $\operatorname{supp} \nu_n^s \subset S_\pi$ and $\operatorname{supp} \nu_n^{\tilde{s}} \subset S_{\tilde{\pi}}$, where $\nu_n^x = \frac{1}{n} \sum_{k=1}^n \mathcal{L}(X_k^x)$ for any $x \in G$, it follows that

$$c \leq d_P(\nu_n^s, \nu_n^{\tilde{s}}) \leq \max_{k=1,\ldots,n} d_P(\mathcal{L}(X_k^s), \mathcal{L}(X_k^{\tilde{s}})) = c \tag{6.4}$$

by Theorem 2.5.4(v). Now taking the limit $n \to \infty$ of (6.4) and using Theorem 6.2.3, it follows that $d_P(\pi, \tilde{\pi}) = c$.

The proves the assertion. □

Definition 6.3.5 (Besicovitch family). A family $\mathcal{B}$ of 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 6.3.6 (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 $K \geq 1$ and $0 < r \leq \infty$ such that $\operatorname{Card} \mathcal{B} \leq K$ 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 6.3.7 (differentiation theorem, Theorem 5.12 in [44])**.** *Let (G, d) be a complete separable metric space. For every locally finite regular measure λ over (G, d), it holds that*

$$\lim_{r \to 0} \frac{1}{\lambda(\mathbb{B}(x, r))} \int_{\mathbb{B}(x,r)} f(y)\lambda(\mathrm{d}y) = f(x) \qquad \textit{for } \lambda\textit{-a.e. } x \in G, \tag{6.5}$$

for all $f \in L^1_{\mathrm{loc}}(G, \lambda)$, if and only if d is σ-finite dimensional.

Proposition 6.3.8 (Besicovitch covering property in $\mathcal{E}$)**.** *Let (G, d) be a Polish 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 $(\mathcal{E}, d_P)$ 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 $(\mathcal{E}, d_P)$, where $\pi_B \in \mathcal{E}$ and $\epsilon_B > 0$. Note that if $\epsilon_B \geq 1$, then $|\mathcal{B}| = 1$, since in that case $B = \mathcal{E}$ 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 6.3.4 into

$$\pi_B \notin B', \quad \forall B' \in \mathcal{B} \setminus B \qquad \Longleftrightarrow \qquad S_{\pi_B} \cap \overline{\mathbb{B}}(S_{\pi_{B'}}, \epsilon_{B'}) = \emptyset, \qquad \forall B' \in \mathcal{B} \setminus B \tag{6.6}$$

and

$$\bigcap_{B \in \mathcal{B}} B \neq \emptyset \qquad \Longleftrightarrow \qquad \bigcap_{B \in \mathcal{B}} \overline{\mathbb{B}}(S_{\pi_B}, \epsilon_B) \neq \emptyset. \tag{6.7}$$

Now fix π in the latter intersection in (6.7) 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 6.2.9). 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 (6.6) 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. $\square$

Remark 6.3.9 (Euclidean metric on $\mathbb{R}^n$ is finite dimensional)**:** The cardinality of any Besicovitch family in $\mathbb{R}^n$ is uniformly bounded depending on n [39, Lemma 2.6].

Lemma 6.3.10 (equality around support of ergodic measures implies equality of measures)**.** *Let (G, d) be a Polish 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*

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

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

Proof. From Theorem 6.1.2 follows the existence of probability measures Q_1, Q_2 on the set $\mathcal{E}$ of ergodic measures for $\mathcal{P}$ such that one has

$$\pi_i(A) = \int_{\mathcal{E}} \pi(A) Q_i(\mathrm{d}\pi), \qquad A \in \mathcal{B}(G),\, i = 1, 2.$$

If we set $Q = \frac{1}{2}(Q_1 + Q_2)$, then by Radon-Nikodyms theorem, there are densities $f_1, f_2 \geq 0$ on $\mathcal{E}$ with $Q_i = f_i \cdot Q$ and hence

$$\pi_i(A) = \int_{\mathcal{E}} \pi(A) f_i(\pi) Q(\mathrm{d}\pi), \qquad A \in \mathcal{B}(G),\, i = 1, 2.$$

For Q-m.b. subsets $E \subset \mathcal{E}$, one can define a probability measure on $\mathcal{E}$ via

$$\tilde{\pi}_i(E) := \int_{\mathcal{E}} \mathbb{1}_E(\pi) f_i(\pi) Q(\mathrm{d}\pi), \qquad i = 1, 2.$$

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

$$\pi_i(\overline{\mathbb{B}}(S_\pi, \epsilon)) = \tilde{\pi}_i(\overline{\mathbb{B}}(\pi, \epsilon)), \qquad i = 1, 2, \tag{6.8}$$

where $\overline{\mathbb{B}}(\pi, \epsilon) := \{\tilde{\pi} \in \mathcal{E} \,|\, d_P(\tilde{\pi}, \pi) \leq \epsilon\}$. This is due to Lemma 6.3.4, from which follows

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

We want to employ Proposition 6.3.7 to show that $f_1 = f_2$ Q-a.s., which would imply that $\pi_1 = \pi_2$. From [44, Theorem 5.12] the differentiation theorem is applicable for any probability measure Q on $(\mathcal{E}, \mathcal{B}(\mathcal{E}))$ if d_P is finite dimensional, i.e. the weak Besicovitch property is satisfied. So, according to Proposition 6.3.8 [44, Theorem 5.12] is applicable and differentiation of $\tilde{\pi}_i$ with respect to Q then gives Q-a.s.

$$\lim_{\epsilon \to 0} \frac{\tilde{\pi}_i(\overline{\mathbb{B}}(\pi, \epsilon))}{Q(\overline{\mathbb{B}}(\pi, \epsilon))} = f_i(\pi).$$

And since $\tilde{\pi}_1(\overline{\mathbb{B}}(\pi, \epsilon)) = \tilde{\pi}_2(\overline{\mathbb{B}}(\pi, \epsilon))$ by (6.8) and the assumption, we have $f_1 = f_2$ Q-a.s. □

Remark 6.3.11: In the assertion of Lemma 6.3.10, it is enough to claim the existence of a sequence $(\epsilon_i^\pi)_{i \in \mathbb{N}} \subset \mathbb{R}_+$ with $\epsilon_i^\pi \to 0$ as $i \to \infty$ satisfying

$$\pi_1(\overline{\mathbb{B}}(S_\pi, \epsilon_i^\pi)) = \pi_2(\overline{\mathbb{B}}(S_\pi, \epsilon_i^\pi)) \qquad \forall \pi \in \mathcal{E},\, \forall i \in \mathbb{N},$$

because from Proposition 6.3.7 one has the existence of the limit in the last equation of the proof Q-a.s. So by the definition of the limit, any particular null-sequence needs to yield the same limit in the last equation of the proof of Lemma 6.3.10.

Theorem 6.3.12 (convergence of Cesáro average)**.** *Let (G, d) be a Polish space with finite dimensional metric d, let $T_i : G \to G$ be nonexpansive $(i \in I)$ and assume* $\operatorname{inv} \mathcal{P} \neq \emptyset$*. Let $\nu_n = \frac{1}{n}\sum_{i=1}^n \mu \mathcal{P}^i$ generated by Algorithm 1 for any $\mu \in \mathscr{P}(G)$. If (ν_n) is tight then this sequence converges to an invariant probability measure for $\mathcal{P}$.*

Proof. Since $\mathcal{P}$ is Feller and (ν_n) is tight it follows from Theorem 2.8.2 that clusterpoints of (ν_n) are invariant measures for $\mathcal{P}$. Let ν^1, ν^2 be such clusterpoints, then we need to show that $\nu^1 = \nu^2$.
This follows from Lemma 6.3.10, we just need to verify the assumptions. It is required that

$$(\forall \pi \in \mathcal{E})(\forall \epsilon > 0) \quad \nu^1(B_\epsilon) = \nu^2(B_\epsilon) \tag{6.9}$$

where $B_\epsilon := \overline{\mathbb{B}}(S_\pi, \epsilon)$. First, fix $x \in G$ and $\epsilon > 0$. Let $A_k := \{X_k^x \in B_\epsilon\}$. By nonexpansivity $A_k \subset A_{k+1}$ for $i \in \mathbb{N}$, since we have by Lemma 6.1.8 a.s. $d(X_{k+1}^x, S_\pi) = d(X_{k+1}^x, T_{\xi_k} S_\pi) \leq d(X_k^x, S_\pi) \leq \epsilon$. Hence $(p^k(x, B_\epsilon)) = (\mathbb{P}(A_k))$ 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 for the average by the Monotone Convergence Theorem

$$\nu_n(B_\epsilon) := \frac{1}{n} \sum_{k=1}^{n} \int p^k(x, B_\epsilon)\, \mathrm{d}\mu \to \int b_\epsilon^x\, \mathrm{d}\mu\,, \qquad n \to \infty.$$

So, in particular $\nu^1(B_\epsilon) = \nu^2(B_\epsilon) = \int b_\epsilon^x\, \mathrm{d}\mu$ for all $\epsilon > 0$. Application of Lemma 6.3.10 gives the uniqueness of the clusterpoint, implying convergence of (ν_n). $\square$

Corollary 6.3.13 (convergence in $\mathbb{R}^n$)**.** *Let $(G, d) = (\mathbb{R}^n, \|\cdot\|)$, let $T_i : G \to G$ be nonexpansive $(i \in I)$ and assume $\operatorname{inv} \mathcal{P} \neq \emptyset$. Let $\nu_n = \frac{1}{n}\sum_{i=1}^{n} \mu \mathcal{P}^i$ generated by Algorithm 1 for any $\mu \in \mathscr{P}(\mathbb{R}^n)$. Then this sequence converges to an invariant probability measure for $\mathcal{P}$.*

Proof. The Euclidean norm is finite dimensional by Remark 6.3.9 and (ν_n) is tight for any initial probability measure μ by Remark 6.3.2. Now Theorem 6.3.12 is applicable. $\square$

6.4. Convergence theory for averaged mappings

Continuing the development of the convergence theory under increasingly strong 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. We restrict ourselves to the Euclidean space $(\mathbb{R}^n, \|\cdot\|)$. A mapping $T : \mathbb{R}^n \to \mathbb{R}^n$ is said to be *averaged* when this can be written as the convex combination of the identity mapping Id and a nonexpansive mapping T', that is, T is averaged when there exists a $T' : \mathbb{R}^n \to \mathbb{R}^n$ nonexpansive such that

$$T = \alpha T' + (1 - \alpha)\operatorname{Id}$$

for some $\alpha \in (0,1)$. It is easy to see that averaged mappings are also nonexpansive. We will use the following equivalent characterization: a mapping T is averaged with constant $\alpha \in (0,1)$ if and only if

$$\|Tx - Ty\|^2 \leq \|x - y\|^2 - \frac{1-\alpha}{\alpha}\|(\operatorname{Id} - T)x - (\operatorname{Id} - T)y\|^2. \tag{6.10}$$

6.4.1 Convergence of $(\mathcal{L}(X_k))$

We begin with a technical Lemma that describes properties of sequences whose relative expected distances are invariant under T_ξ.

Lemma 6.4.1 (constant expected separation)**.** *Let $T_i : \mathbb{R}^n \to \mathbb{R}^n$ be α_i-averaged with $\alpha_i \leq \alpha < 1$, $i \in I$. Let $\mu, \nu \in \mathscr{P}(\mathbb{R}^n)$ and $X \sim \mu$, $Y \sim \nu$ independent of (ξ_k) satisfy*

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

Then for $\mathbb{P}^{(X,Y)}$-a.e. $(x,y) \in \mathbb{R}^n \times \mathbb{R}^n$ 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 Cesáro average of $(\delta_x \mathcal{P}^k)$ and π^y of the Cesáro average of $(\delta_y \mathcal{P}^k)$.

Proof. From averagedness, one has

$$\begin{aligned}
\mathbb{E}\left[\|X - Y\|^2\right] &\geq \mathbb{E}\left[\|T_{\xi_0}X - T_{\xi_0}Y\|^2\right] + \frac{1-\alpha}{\alpha}\mathbb{E}\left[\|(X - T_{\xi_0}X) - (Y - T_{\xi_0}Y)\|^2\right] \\
&\geq \cdots \\
&\geq \mathbb{E}\left[\left\|T_{\xi_{k-1}} \cdots T_{\xi_0}X - T_{\xi_{k-1}} \cdots T_{\xi_0}Y\right\|^2\right] \\
&\quad + \frac{1-\alpha}{\alpha}\sum_{i=0}^{k-1}\mathbb{E}\left[\left\|(T_{\xi_{i-1}} \cdots T_{\xi_{-1}}X - T_{\xi_i} \cdots T_{\xi_0}X) - (T_{\xi_{i-1}} \cdots T_{\xi_{-1}}Y - T_{\xi_i} \cdots T_{\xi_0}Y)\right\|^2\right.
\end{aligned}$$

where we used $T_{\xi_{-1}} := \mathrm{Id}$ for a simpler representation of the sum. We will denote $X_k^x = T_{\xi_{k-1}} \cdots T_{\xi_0}x$. 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 $i = 1, \ldots, k$ $\mathbb{P}$-a.s.

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

and hence by induction

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

By the disintegration theorem Theorem 2.3.2 and using $(X,Y) \perp\!\!\!\perp (\xi_k)$ we have $\mathbb{P}$-a.s.

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

which means for $\mathbb{P}^{(X,Y)}$-a.e. $(x,y) \in \mathbb{R}^n \times \mathbb{R}^n$, that $\mathbb{P}$-a.s. holds

$$X_k^x - X_k^y = x - y, \qquad \forall k \in \mathbb{N}.$$

So in particular for any $A \in \mathcal{B}(\mathbb{R}^n)$

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

and hence, denoting $f_h = f(\cdot + h)$ and $\nu_n^x = \frac{1}{n}\sum_{i=1}^n p^i(x, \cdot)$, one also has for $f \in C_b(\mathbb{R}^n)$ by Corollary 6.3.13

$$\begin{aligned} \nu_n^y f_{x-y} &\to \pi^y f_{x-y} = \pi_{x-y}^y f, \\ \nu_n^x f &\to \pi^x f \end{aligned}$$

as $n \to \infty$ and where $\pi_{x-y}^y := \pi^y(\cdot - (x - y))$. So from $\nu_n^y f_{x-y} = \nu_n^x f$ for any $f \in C_b(\mathbb{R}^n)$ and $n \in \mathbb{N}$ follows $\pi_{x-y}^y = \pi^x$. $\square$

Theorem 6.4.2 (convergence of iterates for averaged mappings)**.** *Let $T_i : \mathbb{R}^n \to \mathbb{R}^n$ be α_i-averaged with $\alpha_i \le \alpha < 1$ ($i \in I$) and assume there exists an invariant probability distribution for $\mathcal{P}$. For any initial distribution $\mu \in \mathscr{P}(\mathbb{R}^n)$ the distributions of the iterates $\mu\mathcal{P}^k$ generated by Algorithm 1 converge to an invariant probability measure for $\mathcal{P}$.*

Proof. Let $x, y \in \mathbb{R}^n$ and define $d(x, y) := \|x - y\|^2$ and $f^M(x, y) := \min(M, d(x, y))$, where $M \in \mathbb{R}$. For given $g : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ define a sequence of functions (g_n) on $\mathbb{R}^n \times \mathbb{R}^n$ via

$$g_k(x, y) = \mathbb{E}\left[g(X_k^x, X_k^y)\right], \qquad k \in \mathbb{N},$$

where $X_k^z := T_{\xi_{k-1}} \cdots T_{\xi_0} z$ for any $z \in \mathbb{R}^n$. Note that $g_k \in C_b(\mathbb{R}^n)$ for all $k \in \mathbb{N}$ if $g \in C_b(\mathbb{R}^n)$ by continuity of T_i, $i \in I$ and Lebesgue's Dominated Convergence Theorem. From averagedness of T_i, $i \in I$, we get that a.s. for all $k \in \mathbb{N}$

$$\|X_k^x - X_k^y\|^2 \ge \left\|X_{k+1}^x - X_{k+1}^y\right\|^2 + \frac{1-\alpha}{\alpha}\left\|(X_k^x - X_{k+1}^x) - (X_k^y - X_{k+1}^y)\right\|^2. \tag{6.11}$$

After taking expectation, this is the same as writing

$$d_k(x, y) \ge d_{k+1}(x, y) + \frac{1-\alpha}{\alpha}\mathbb{E}\left[\left\|(X_k^x - X_{k+1}^x) - (X_k^y - X_{k+1}^y)\right\|^2\right].$$

We conclude that $(d_k(x, y))$ is a monotonically nonincreasing sequence for any $x, y \in G$. Let $s, \tilde{s} \in S_\pi$ for some $\pi \in \mathcal{E}$ and define the sequence of measures

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

for any measurable function $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. Note that due to nonexansiveness $(X_k^s, X_k^{\tilde{s}})$ a.s. takes values in $D_r := \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n : ||x - y||^2 \le r\}$ 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 6.3.1 and likewise $(X_k^{\tilde{s}})$ we know from Lemma 2.6.3 that the sequence (γ_k) is tight as well. Let γ be a clusterpoint of (γ_k), which is again concentrated on $D_{||s-\tilde{s}||^2}$, and consider a subsequence (γ_{k_n}) such that $\gamma_{k_n} \to \gamma$. By Lemma 2.6.3 we also know that $\gamma \in C(\nu_1, \nu_2)$ where ν_1 and ν_2 are the distributions of the weak limit of $(X_{k_n}^s)$ and $(X_{k_n}^{\tilde{s}})$. For any $f \in C_b(\mathbb{R}^n)$ we have

$\gamma_{k_n} f \to \gamma f$. We now consider $f = f^M$ and use that for $M \geq \|s-\tilde{s}\|^2$ we have γ_{k_n} and γ a.s. that $d = f^M$ in order to conclude that

$$\gamma_{k_n} d = \gamma_{k_n} f^M \to \gamma f^M = \gamma d.$$

However, by the monotonicity in (6.11) we now also obtain the convergence

$$\gamma_k d = \gamma_k f^M \searrow \gamma f^M = \gamma d$$

for the entire sequence. 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 $\tilde{X}_k^x := T_{\tilde{\xi}_{k-1}} \cdots T_{\tilde{\xi}_0} x$, $x \in \mathbb{R}^n$. Since $f_k^M \in C_b(\mathbb{R}^n)$ we have for $M \geq \|s-\tilde{s}\|^2$ that

$$\begin{aligned}
\gamma d_k = \gamma f_k^M &= \mathbb{E}\left[\min\left(M, \left\|\tilde{X}_k^X - \tilde{X}_k^Y\right\|^2\right)\right] = \lim_{n\to\infty} \gamma_{k_n} f_k^M \\
&= \lim_{n\to\infty} \mathbb{E}\left[\min\left(M, \left\|\tilde{X}_k^{X_{k_n}^s} - \tilde{X}_k^{X_{k_n}^{\tilde{s}}}\right\|^2\right)\right] \\
&= \lim_{n\to\infty} \mathbb{E}\left[\min\left(M, \left\|X_{k+k_n}^s - X_{k+k_n}^{\tilde{s}}\right\|^2\right)\right] \\
&= \lim_{n\to\infty} \gamma_{k+k_n} f^M = \gamma f^M = \gamma d.
\end{aligned}$$

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 6.2.3. An application of Lemma 6.4.1 yields then that $\pi(\cdot) = \pi(\cdot - (x-y))$, i.e. $x = y$. Hence $X = Y$ a.s. implying $\nu_1 = \nu_2 =: \nu$ and $\gamma d = 0$. That means

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

which implies convergence W_2 Wasserstein metric of the corresponding probability measures $\delta_s \mathcal{P}^k$ and $\delta_{\tilde{s}} \mathcal{P}^k$ and thus also in the Prohorov metric, namely

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

Hence, by the triangle inequality, if we have convergence of $\delta_s \mathcal{P}^{k_n} \to \nu$ then also $\delta_{\tilde{s}} \mathcal{P}^{k_n} \to \nu$ for any $\tilde{s} \in S_\pi$:

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

We conclude that for any $f \in C_b(\mathbb{R}^n)$ and $\mu \in \mathscr{P}(S_\pi)$ it holds that

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

by Lebesgue's Dominated Convergence Theorem. This means that $\mu\mathcal{P}^{k_n} \to \nu$ and since we may set $\mu = \pi$, we have $\nu = \pi$. Thus, all clusterpoints of $(\delta_s\mathcal{P}^k)$ for all $s \in S_\pi$ have the same distribution π and hence the convergence $\delta_s\mathcal{P}^k = p^k(x,\cdot) \to \pi$ follows due to tightness.

Now, let $\mu \in \mathscr{P}(S)$, where $S = \bigcup_{\pi\in\mathcal{E}} S_\pi$. 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, by Lebesgue's Dominated Convergence Theorem, one has for any $f \in C_b(\mathbb{R}^n)$,

$$\mu\mathcal{P}^k f = \int f(y)p^k(x,\mathrm{d}y)\mu(\mathrm{d}x) \to \int f(y)\pi^x(\mathrm{d}y)\mu(\mathrm{d}x) =: \pi^\mu f \tag{6.12}$$

as $k \to \infty$ and the measure π^μ is again invariant for $\mathcal{P}$ by invariance of π^x for all $x \in S$. Now, let $\mu = \delta_x$, $x \in \mathbb{R}^n \setminus S$. We obtain the tightness of $(\delta_x\mathcal{P}^k)$ from the tightness of $(\delta_s\mathcal{P}^k)$ for $s \in S$: For $\epsilon > 0$ there exists a compact $K_\epsilon \subset \mathbb{R}^n$ with $p^k(s,K_\epsilon) > 1-\epsilon$ for all $k \in \mathbb{N}$. Combining this with nonexpansiveness of T_i, $i \in I$ implying $\|X_k^x - X_k^s\| \le \|x-s\|$ for all $k \in \mathbb{N}$ leads to $p^k(x,\overline{\mathbb{B}}(K_\epsilon,\|x-s\|)) > 1-\epsilon$. Tightness implies the existence of a clusterpoint ν of the sequence $(\delta_x\mathcal{P}^k)$. From Corollary 6.3.13 we know, that $\nu_n^x = \frac{1}{n}\sum_{i=1}^n \delta_x\mathcal{P}^i \to \pi^x$ 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 6.3.3((iv)). So by (6.12) 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 $\nu = \pi^x$, i.e. π^x is the unique clusterpoint of $(\delta_x\mathcal{P}^k)$ and hence convergence follows by Theorem A.0.16. This is possible by showing that $\pi^\nu = \pi^x$, since then, as $k \to \infty$

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

To see that $\nu = \pi^x$, 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 $i \in \mathbb{N}$, since we have by Lemma 6.1.8 a.s.

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

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 for all $\pi \in \operatorname{inv}\mathcal{P}$ that

$$b_\epsilon^x = \lim_k p^k(x,\overline{\mathbb{B}}(S_\pi,\epsilon)) = \lim_n \frac{1}{n}\sum_{i=1}^n p^i(x,\overline{\mathbb{B}}(S_\pi,\epsilon)). \tag{6.13}$$

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 [33, Chapter 3, Example 1.3]). With the same argument as in (6.13) 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 also

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

needs to hold for all except countably many ϵ, which implies since $\pi^\nu \in \operatorname{inv}\mathcal{P}$ that $\pi^\nu = \pi^x$ by Lemma 6.3.10 combined with Remark 6.3.11.
For a general initial measure $\mu \in \mathscr{P}(\mathbb{R}^n)$, one has by Lebesgue's Dominated Donvergence Theorem that

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

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

6.4.2 Structure of ergodic measures for averaged mappings

Theorem 6.4.3 (structure of ergodic measures)**.** *Let $T_i : \mathbb{R}^n \to \mathbb{R}^n$ be α_i-averaged with $\alpha_i \leq \alpha < 1$ ($i \in I$) and assume there exists an invariant probability distribution for $\mathcal{P}$. Any two ergodic measures $\pi, \tilde{\pi}$ are shifted versions of each other, i.e. there exist $s \in S_\pi$ and $\tilde{s} \in S_{\tilde{\pi}}$ with $\pi = \tilde{\pi}(\cdot - (s - \tilde{s}))$.*

Proof. Since we can find for any $s \in S_\pi$ a closest point $\tilde{s} \in S_{\tilde{\pi}}$, i.e. $\operatorname{dist}(S_\pi, S_{\tilde{\pi}}) = d(s, \tilde{s})$, by Lemma 6.2.9, the assertion follows from

$$\operatorname{dist}(S_\pi, S_{\tilde{\pi}}) \leq \sqrt{\mathbb{E}[\|X_k^s - X_k^{\tilde{s}}\|^2]} \leq \|s - \tilde{s}\| \qquad \forall k \in \mathbb{N}, \tag{6.14}$$

where we also used that $\operatorname{supp}\mathcal{L}(X_k^s) \subset S_\pi$, $\operatorname{supp}\mathcal{L}(X_k^{\tilde{s}}) \subset S_{\tilde{\pi}}$. □

Proposition 6.4.4 (specialization to projectors)**.** *Let the mappings $T_i = P_i : \mathbb{R}^n \to \mathbb{R}^n$ be projectors onto nonempty closed and convex sets ($i \in I$). If there exist two ergodic measures π_1, π_2, then there exist infinitely many ergodic measures π_λ with $\pi_\lambda := \pi_2(\cdot - \lambda a)$ for all $\lambda \in [0,1]$, where a is the shift such that $\pi_1 = \pi_2(\cdot - a)$.*

Proof. For any pair $(s_1, s_2) \in S_{\pi_1} \times S_{\pi_2}$ of closest neighbors it holds that $a = s_1 - s_2$ by Theorem 6.4.3. From (6.14) and Lemma 6.4.1 follows $P_\xi s_1 = P_\xi s_2 + a$ a.s. Hence, $a \perp (s_i - P_\xi s_i)$, $i = 1, 2$, and then $P_\xi(s_2 + \lambda a) = P_\xi s_2 + \lambda a$ for $\lambda \in [0,1]$. Hence $X_k^{s_2+\lambda a} = X_k^{s_2} + \lambda a$ and $\lim_k \mathcal{L}(X_k^{s_2+\lambda a}) = \pi_2(\cdot - \lambda a)$. Note, that if $\mathcal{P}$ is Feller and the limit $(\mu\mathcal{P}^k)$ exists for some $\mu \in \mathbb{R}$, then it is also an invariant measure. □

6.5. Embedding into existing work

Random function iterations are often understood as a finite system of functions and probabilities thereon. There are also dynamical systems, where the function space is infinite, but these are mostly considered for continuous time. We understand random function systems as infinite (possibly uncountable) function systems. These were in the literature mostly analyzed with contraction properties of the mappings or the expected contraction property, see e.g. the review [48]. In that review also the terminology of an equicontinuous Markov chain comes into play. Equicontinuous Markov chains have been analyzed thoroughly in Meyn and Tweedie [40], there called e-chains. There equicontinuity is understood as equicontinuity for $(\mathcal{P}^k f)$ for any $f \in C_c(\mathbb{R})$. This concept is appropriate for only locally-compact metric spaces. Still their concept is not suitable for us, it is too demanding on the properties of the Markov operator $\mathcal{P}$. But there is another concept, the e-property concept by Lasota and Szarek [35]. Here equicontinuity is understood as equicontinuity of the sequence $(\mathcal{P}^k f)$ for any Lipschitz continuous $f : G \to \mathbb{R}$. This concept is working also on general Polish spaces. Any Markov operator that is constructed via nonexpansive functions fits exactly into that framework. There is a lot known and done in the case of Markov operators that satisfy the e-property. For example there is an extraordinary dissertation by Worm [54], who covers Theorem 6.3.12 in his Theorem 7.3.1, where he can show that even on general Polish spaces for any Markov operator satisfying the weaker Cesàro e-property and for which the sequence $(\delta_x \mathcal{P}^k)$ is tight for some $x \in G$, this sequence of Cesàro averages is converging to an invariant measure for the Markov operator. (We found out about that after our proof was done, so the proof in this thesis is independent of that). But Worm is also able to show a much finer structure of the ergodic set S even without the e-property. Our work is based on his results on general Polish spaces as well as on the work by Szarek [35,51] (and later work), especially when dealing with Markov operators satisfying the e-property.

In this thesis we are just interested in $\mathbb{R}^n$. Work on properties of Markov operators on locally compact spaces are the monographs by Duflo [18], Hernandez-Lerma and Lasserre [26] and Zaharopol [55]. Neither of these monographs uses the e-property, but Meyn and Tweedie [40, Theorem 12.0.1] and Zaharopol [55, Theorem 4.3.1] can show convergence of Cesàro averages for e-chains. In [26, Theorem 5.2.2] there are more general results than we have found in Corollary 6.1.7 about the ergodic behavior of our Markov chain in locally compact metric spaces, in particular this theorem is generalized to locally compact metric spaces.

In compact metric spaces a strong law of large numbers and convergence of the Cesàro averages could be shown under the assumption that there exists a unique probability measure, or that the Markov operator is equicontinuous in the sense of e-chains, see for example [12], [27]. These results were generalized in the popular monograph [40] by Meyn and Tweedie. Under the assumption of uniqueness of the invariant measure, there could be shown a law of large numbers [40, chapter 17] for e-chains. For the even stronger condition of positive Harris recurrence, there could be shown also a strong law of large numbers and a central limit theorem [40, chapter 18]. See also [18, chapter 8]. It would be interesting to know, if these results are also true in our case.

CHAPTER 7

Geometric Convergence - Inconsistent Feasibility

For a general metric space G, we have the property that the sequence $(\mu\mathcal{P}^k) \subset \mathscr{P}(G)$ for some $\mu \in \mathscr{P}(G)$ is Fejér monotone with respect to $\operatorname{inv}\mathcal{P}$ (i.e. $D(\mu\mathcal{P}^{k+1}, \pi) \leq D(\mu\mathcal{P}^k, \pi)$ for all $\pi \in \operatorname{inv}\mathcal{P}$ and $k \in \mathbb{N}$, and some metric D on $\mathscr{P}(G)$) in the TV-norm by [46, Proposition 3(d)]. Under the assumption that $\{T_i\}_{i \in I}$ is a family of nonexpansive mappings, Fejér monotonicity is given also for the Wasserstein metric directly by the definition and with help of Lemma 2.8.1.

In [30] they introduce the concept of the modulus of regularity to be able to speak of a convergence rate of the Fejér monotone sequence under a regularity assumption on the sequence. In particular, a function $F : \mathscr{P}(G) \to \mathbb{R}$ with $F^{-1}(0) = \operatorname{inv}\mathcal{P}$ is said to have a *modulus of regularity* ϕ w.r.t. to $\overline{\mathbb{B}}(z, r)$ if for $\epsilon > 0$

$$|F(x)| < \phi(\epsilon) \qquad \Longrightarrow \qquad \operatorname{dist}(x, F^{-1}(0)) < \epsilon$$

for a function $\phi : (0, \infty) \to (0, \infty)$ and for all $x \in \overline{\mathbb{B}}(z, r)$ with $z \in F^{-1}(0)$ and $r > 0$.

Theorem 7.0.1 (Theorem 4.1 in [30])**.** *Let (X, d) be a complete metric space and $F : X \to \overline{\mathbb{R}}$ with $F^{-1}(0) \neq \emptyset$. Suppose that (x_n) is a sequence in X which is Fejér monotone w.r.t. $F^{-1}(0)$, $b > d(x_0, z)$ for some $z \in F^{-1}(0)$ and there exists $\alpha : (0, \infty) \to \mathbb{N}$ such that*

$$\forall \epsilon > 0 \, \exists n \leq \alpha(\epsilon) : \quad |F(x_n)| < \epsilon.$$

If ϕ is a modulus of regularity for F w.r.t. $\overline{\mathbb{B}}(z, b)$, then (x_n) is a Cauchy sequence, and if $F^{-1}(0)$ is closed then (x_n) converges to a zero of F with rate of convergence $\alpha(\phi(\epsilon/2))$.

In Theorem 7.0.1 it is shown how to get a rate of convergence for the sequence $(\mu\mathcal{P}^k)$. The only difficulty is to find an appropriate function that has a modulus of regularity for this sequence. A candidate could possibly be $F(\mu) = d(\mu, \mu\mathcal{P})$, but we were not able to show the regularity property in general. They also mention how to get a linear rate of convergence in an updated version of [30, Theorem 4.5].

We now focus on geometric rates of convergence and, in simple cases of the structure of invariant measures, show what regularity conditions lead to geometric convergence. We will just consider special structures on the set of ergodic measures, and were not able to give a statement for the general case, but at the end of this chapter there will be some thoughts, where to start.

Recall, the regularity condition for the consistent stochastic feasibility problem to show geometric convergence behavior was

$$\operatorname{dist}^2(x, C) \leq \kappa R(x) \qquad \forall x \in \mathbb{R}^n, \tag{7.1}$$

where R is defined in Eq. (4.5). First, we will give a regularity condition, that is equivalent to that in (7.1) in the consistent case: Let π be the unique invariant measure for $\mathcal{P}$ with compact support, assume

$$\|x - y\|^2 \leq \kappa \mathbb{E}[\|(T_\xi x - x) - (T_\xi y - y)\|^2], \qquad \forall x \in G, \forall y \in \operatorname{supp} \pi. \tag{7.2}$$

In the consistent case, that would correspond to $C = \{c\}$ for a $c \in \mathbb{R}^n$. Then (7.2) would read as $\operatorname{dist}^2(x, C) \leq \kappa R(x)$, which is (7.1).
To see the geometric convergence, note that from averagedness of T_i, $i \in I$ follows for all $x, y \in \mathbb{R}^n$ a.s., that (see (6.10))

$$\|x - y\|^2 \geq \|T_\xi x - T_\xi y\|^2 + \frac{1-\alpha}{\alpha}\|(T_\xi x - x) - (T_\xi y - y)\|^2.$$

Taking the expectation, letting $y \in \operatorname{supp} \pi$ and using (7.2) yields

$$\|x - y\|^2 \left(1 - \frac{1-\alpha}{\alpha}\kappa^{-1}\right) \geq \mathbb{E}\left[\|T_\xi x - T_\xi y\|^2\right].$$

Integration with respect to $\gamma \in C(\mu, \pi)$, where $\mu \in \mathscr{P}(\mathbb{R}^n)$ and γ is the optimal coupling for $W_2(\mu, \pi)$ yields (note that $\operatorname{supp} \gamma \subset \operatorname{supp} \mu \times \operatorname{supp} \pi$)

$$W_2^2(\mu, \pi)\left(1 - \frac{1-\alpha}{\alpha}\kappa^{-1}\right) \geq W_2^2(\mu\mathcal{P}, \pi).$$

A trivial generalization to the case that more invariant measures exist, is possible in the case, that $\mathbb{R}^n$ is decomposable into the closed sets D_π, such that $\delta_x \mathcal{P}^k \to \pi$ with $x \in D_\pi$ and $\pi \in \mathcal{E}$. So, one has that $\mathbb{R}^n = \bigcup_{\pi \in \mathcal{E}} D_\pi$, where the sets D_π are disjoint.

A regularity condition, that ensures geometric convergence of $(\mathcal{L}(X_k))$ to its limit and which is equivalent to (7.1) in the consistent case is:

$$\|x - y\|^2 \leq \kappa \mathbb{E}[\|(T_\xi x - x) - (T_\xi y - y)\|^2], \qquad \forall x \in D_\pi, \forall y \in \operatorname{supp} \pi. \tag{7.3}$$

Suppose the problem is consistent, i.e. $C \neq \emptyset$, then from the proof of [7, Theorem 5.12] we know that $\|x - X\|^2 \leq 2 \operatorname{dist}^2(x, C)$, where X is the a.s. limit of (X_k^x), i.e. $X \sim \pi$ for $x \in D_\pi$. So in particular for $x \in D_\pi$ and all $y \in S_{\pi^x}$ it holds, if (7.1) is satisfied, that

$$\|x - y\|^2 \leq 2\kappa \mathbb{E}[\|T_\xi x - x\|^2].$$

Clearly, if (7.3) is satisfied, then also (7.1) holds.
Now we will show, that the regularity condition in (7.3) implies in fact geometric convergence:

Theorem 7.0.2 (Geometric Convergence)**.** *Let $T_i : \mathbb{R}^n \to \mathbb{R}^n$ be α_i-averaged with $\alpha_i \le \alpha < 1$, $i \in I$. Suppose there exists an invariant measure for $\mathcal{P}$ and assume that (7.3) is satisfied (i.e. also a decomposition of $\mathbb{R}^n$ into sets D_π exists), then*

$$W_2^2(\mu\mathcal{P}^k, \pi^\mu) \le \left(1 - \frac{1-\alpha}{\alpha}\kappa^{-1}\right)^k \int W_2^2(\delta_x, \pi^x)\mu(\mathrm{d}x)$$

for any initial probability measure $\mu \in \mathscr{P}(\mathbb{R}^n)$ and corresponding limiting measure π^μ of $(\mu\mathcal{P}^k)$.

Proof. Letting $x \in \mathbb{R}^n$ and $Y_x \sim \pi_x$ with $Y_x \perp\!\!\!\perp (\xi_k)$, where π^x the unique ergodic measure such that $x \in D_{\pi^x}$, we get from averagedness (6.10) that $k \in \mathbb{N}$ holds

$$\left\|X_k^x - X_k^{Y_x}\right\|^2 \ge \left\|X_{k+1}^x - X_{k+1}^{Y_x}\right\|^2 + \frac{1-\alpha}{\alpha}\left\|(X_k^x - X_{k+1}^x) - (X_k^{Y_x} - X_{k+1}^{Y_x})\right\|^2,$$

where $X_k^x = T_{\xi_{k-1}} \cdots T_{\xi_0} x$. Taking the expectation and using (7.3) (possible, since $(X_k^x, X_k^{Y_x}) \in D_{\pi^x} \times \operatorname{supp} \pi^x$ a.s. by) yields

$$\mathbb{E}[\left\|X_k^x - X_k^{Y_x}\right\|^2]c \ge \mathbb{E}[\left\|X_{k+1}^x - X_{k+1}^{Y_x}\right\|^2],$$

where $c = 1 - \frac{1-\alpha}{\alpha}\kappa^{-1}$. So, by induction

$$\mathbb{E}[\left\|X_k^x - X_k^{Y_x}\right\|^2] \le c^k \mathbb{E}[\|x - Y_x\|^2].$$

For the Wasserstein distance between $\mu\mathcal{P}^k$ and π^x we find

$$W_2^2(\mu\mathcal{P}^k, \pi^x) \le \int \mathbb{E}[\left\|X_k^x - X_k^{Y_x}\right\|^2]\mu(\mathrm{d}x) \le c^k \int \mathbb{E}[\|x - Y_x\|^2]\mu(\mathrm{d}x) = c^k \int W_2^2(\delta_x, \pi^x)\mu(\mathrm{d}x).$$

In the first inequality we used that $\mu\mathbb{P}^{(X_k^\cdot, X_k^{Y_\cdot})} \in C(\mu\mathcal{P}^k, \pi^\mu)$, where $x \mapsto Y_x$ can be chosen to be measurable by Section 6.4.2, since ergodic supports are shifted, we let $Y_x = Y + s_x$, where for some fixed $\pi \in \mathcal{E}$ we let $Y \sim \pi$ and s_x be the shift between S_π and S_{π^x}. Note that $W_2^2(\delta_x, \pi^x)$ could be infinite in this theorem, which would make the assertion trivial. So it is just reasonable for $\pi \in \mathscr{P}_2(\mathbb{R}^n)$. □

Another criterion for geometric convergence is the minorization condition, i.e. the existence of a measure ν for which holds

$$p(x, \cdot) \ge \kappa\nu, \qquad \forall x \in \mathbb{R}^n.$$

Linear convergence with respect to the TV-norm, i.e.

$$\|\mu - \nu\|_{\mathrm{TV}} = \sup_{\|f\|_\infty \le 1, f \text{ mb.}} |\mu f - \nu f|$$

is then a consequence of Theorem 2.10.3. This is again for the case of a single invariant measure, but existence is implied by the theorem.
Again the simple generalization for the case of sets D_π that decompose $\mathbb{R}^n$, with

$$D_\pi = \left\{x \in \mathbb{R}^n \,\middle|\, p^k(x, \cdot) \to \pi\right\}, \qquad \pi \in \mathcal{E}$$

is possible. Theorem 2.10.3 is applicable on the closed set $\mathscr{P}(D_\pi)$ with minorization measure ν_π and uniform constant κ, so that

$$\left\|\delta_x \mathcal{P}^k - \pi^x\right\|_{\mathrm{TV}} \leq (1-\kappa)^k \|\delta_x - \pi^x\|_{\mathrm{TV}}$$

for all $x \in \mathcal{R}^n$ and all $k \in \mathbb{N}$. By Lemma 2.10.1((i)) immediately follows that

$$\left\|\mu \mathcal{P}^k - \pi^\mu\right\|_{\mathrm{TV}} \leq (1-\kappa)^k \int \|\delta_x - \pi^x\|_{\mathrm{TV}} \mu(\mathrm{d}x).$$

Remark 7.0.3 (Minorisation condition)**:** Note, that the condition $\nu\mathcal{P} \geq \kappa\nu_\pi$ for all $\nu \in \mathscr{P}(D_\pi)$ implies, that also $\pi \geq \kappa\nu_\pi$ by properties of measures on metric spaces [43, Theorem 1.2] and properties of the weak limit [43, Theorem 6.1 (c)]. This makes clear, that the minorisation condition is very strong, since knowledge of ν_π localizes π to some extend, e.g. its support.

It is important to note that still, the following simple problem is not describable in any of the previous frameworks, since the sets D_π do not form a partition of $\mathbb{R}^2$:

Example 7.0.4 (Geometric convergence, regularity condition?)**.** Define the following two convex sets in $\mathbb{R}^2$

$$\begin{aligned} C_1 &:= \{x = (x_1, x_2) \,|\, x_2 \geq |x_1|, \quad x_1 \leq 0\}\,, \\ C_2 &:= \{x = (x_1, x_2) \,|\, x_2 \geq 0, \quad x_1 \geq 1\}\,. \end{aligned}$$

The projectors satisfy $P_2P_1x \in \mathrm{Fix}(P_2P_1)$ and $P_1P_2x \in \mathrm{Fix}(P_1P_2)$ for all $x \in \mathbb{R}^2$. Letting $\mathbb{P}(\xi = 1) = \mathbb{P}(\xi = 2) = \frac{1}{2}$, the invariant measure that is limit of $(\mathcal{P}^k\delta_x)$ is given for $x = (x_1, x_2)$ with $-x_1 > x_2 > 0$ by

$$\pi_x = \frac{1}{4}\left(\delta_{(0,x_2)} + \delta_{(1,x_2)} + \delta_{(0,\tilde{x}_2)} + \delta_{(1,\tilde{x}_2)}\right),$$

for $\tilde{x}_2 := \frac{1}{2}(x_2 - x_1)$, where $P_1P_2x = (0, x_2)$, $P_2P_1x = (1, \tilde{x}_2)$, $P_1P_2P_1x = (0, \tilde{x}_2)$ and $P_2P_1P_2x = (1, x_2)$. Note that this measure is not ergodic for any of these x satisfying the above conditions. The convergence is linear (in the TV-norm), since for $k \geq 2$

$$\begin{aligned} \mathbb{P}(X_k^x \in A) = \frac{1}{2^k}\left(\mathbb{1}_A\{P_1x\} + \mathbb{1}_A\{P_2x\}\right) + \frac{1}{4}\left(\mathbb{1}_A\{P_2P_1x\} + \mathbb{1}_A\{P_1P_2x\}\right) \\ + \frac{1}{4}\left(1 - \tfrac{1}{2^{k-2}}\right)\left(\mathbb{1}_A\{P_1P_2P_1x\} + \mathbb{1}_A\{P_2P_1P_2x\}\right). \end{aligned}$$

A satisfying regularity conditions is yet to be found that would describe geometric convergence for a general structure of the ergodic measures. A potentially fitting framework would be to view $\mathcal{P}$ as a linear operator on the Banach space, that is generated as the closure of the span of all Dirac's delta measures, see [54, Chapter 2]. The set of all finite signed measures is densely contained in it, for details see [54, Corollary 2.3.10]. Another possibility is to consider the cone of all finite measures, see [54, Theorem 2.3.9]. Then hopefully the Fejér monotonicity and some regularity condition on $\mathcal{P}$ (metric regularity) can be formulated to get linear convergence of $(\mu\mathcal{P}^k)$ for any $\mu \in \mathscr{P}(G)$.

CHAPTER 8

Applications and Examples

We give several artificial and also practically relevant examples of consistent and inconsistent feasibility problems. These will show the richness and generality of our setup to model errors as selection of averaged mappings.

8.1. Consistent Feasibility

We specialize the framework above to several well-known settings: consistent convex feasibility, linear operator equations and in particular Hilbert-Schmidt operators (i.e. linear integral equations).

8.1.1 Feasibility and stochastic projections

There are many algorithms for solving convex feasibility problems. We focus on the (conceptually) simplest of these, namely stochastic projections. In the context of Algorithm 1, $T_i = P_i$ is a projector, $i \in I$, onto a nonempty closed and convex set $C_i \subset \mathcal{H}$, $i \in I$ and $\mathcal{H}$ a Hilbert space. Note that projectors are $\frac{1}{2}$-averaged operators [7, Proposition 4.8] (also referred to as *firmly nonexpansive* operators), so $\alpha_i = \frac{1}{2}$ for all $i \in I$, we then can choose the upper bound $\alpha = \frac{1}{2}$ as well. Also note that $\operatorname{Fix} P_i = C_i$, $i \in I$.

As a first assertion we give an equivalent characterization for the regularity property in Eq. (5.1) using just properties of R. This characterization, known as Kurdyka-Łojasiewicz (KL) property, eliminates the term with the distance to the usually unknown fixed point set C, but one needs to be able to compute the first derivative of the function R. For convex sets this is unproblematic since R is the expectation of the squared distances to the convex sets C_i, see Lemma A.0.23.

Definition 8.1.1 (KL property)**.** A convex, continuously differentiable function $f : \mathcal{H} \to \mathbb{R}$ with $\inf_x f(x) = 0$ and $S := \operatorname{argmin} f \neq \emptyset$ is said to have the global KL property, if there exists a concave continuously differentiable function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ with $\varphi(0) = 0$ and $\varphi' > 0$ such that

$$\varphi'(f(x))\|\nabla f(x)\| \geq 1 \qquad \forall x \in \mathcal{H} \setminus S.$$

The following theorem is a direct consequence of [11].

Proposition 8.1.2 (equivalent characterization of Eq. (5.1))**.** *Under the standing assumptions, let $T_i = P_i$ be projectors onto nonempty, closed and convex sets, $i \in I$. Then the regularity condition in Eq. (5.1) is satisfied with $\kappa > 0$ if and only if $R(x) \leq \frac{\kappa}{4}\|\nabla R(x)\|^2$ $\forall x \in \mathcal{H}$, i.e. R has the global KL property.*

Proof. Apply [11, Corollary 6] with $\varphi(s) := \sqrt{\kappa s}$ and $f = R$ and note that R is convex and differentiable (see Lemma A.0.23). □

Theorem 8.1.3 (uniform bounds)**.** *Under the standing assumptions, suppose the regularity condition in Eq. (5.1) is satisfied and that $\mathcal{H}$ is separable and $T_i = P_i$ are projectors onto nonempty, closed and convex sets, $i \in I$. Then the probability of any point being feasible is uniformly bounded, i.e. $\mathbb{P}(x \in C_\xi) \leq r < 1$ for all $x \in \mathcal{H} \setminus C$.*

Proof. It holds surely for all $x \in \mathcal{H}$

$$\operatorname{dist}(P_\xi x, C) \geq \operatorname{dist}(x, C) - \operatorname{dist}(x, C_\xi).$$

This, together with the expectation

$$\mathbb{E}[\operatorname{dist}(x, C_\xi)] = \mathbb{E}[\operatorname{dist}(x, C_\xi)\mathbb{1}_{\{x \notin C_\xi\}}] \leq \mathbb{E}[\operatorname{dist}(x, C)\mathbb{1}_{\{x \notin C_\xi\}}] = \operatorname{dist}(x, C)(1 - \mathbb{P}(x \in C_\xi))$$

yields, for $X_0 \sim \delta_x$,

$$\mathbb{E}[\operatorname{dist}(X_1, C)] \geq \mathbb{P}(x \in C_\xi) \operatorname{dist}(x, C).$$

Hence by Theorem 5.0.9

$$1 > r := \sup_{x \in \mathcal{H} \setminus C} \frac{\mathbb{E}[\operatorname{dist}(X_1, C)]}{\operatorname{dist}(x, C)} \geq \sup_{x \in \mathcal{H} \setminus C} \mathbb{P}(x \in C_\xi).$$

□

Theorem 8.1.4 (finite vs. infinite convergence)**.** *Under the standing assumptions, let $\mathcal{H}$ be separable and let $T_i = P_i$ be projectors ($i \in I$). Then one of the following holds:*

(i) $\mathbb{P}(X_1 \in C) = 1$ *and* $\mathbb{P}(X_n \in C) = 1$ *for all* $n \in \mathbb{N}$,

(ii) $\mathbb{P}(X_1 \in C) < 1$ *and* $\mathbb{P}(X_n \in C) < 1$ *for all* $n \in \mathbb{N}$.

Proof. (i) If $\mathbb{P}(X_1 \in C) = 1$, then $X_k = X_1$ a.s. for all $k \geq 1$.

(ii) From $\int p(x, C)\mu(\mathrm{d}x) = \mathbb{P}(X_1 \in C) < 1$ we get, that there is $x \in \operatorname{supp}\mu \setminus C$ with $p(x, C) < 1$, where μ is the initial distribution. Since $p(x, \mathcal{H} \setminus C) > 0$, there exists $y \in \operatorname{supp} p(x, \cdot) \setminus C$. Then by Theorem 2.4.1 this implies that $p(x, \overline{\mathbb{B}}(y, \epsilon)) > 0$ for all $\epsilon > 0$.
Furthermore, one has for any $\epsilon > 0$ that

$$(\forall z \in \overline{\mathbb{B}}(y, \epsilon)) \qquad p(z, \overline{\mathbb{B}}(y, 2\epsilon)) \geq p(x, \overline{\mathbb{B}}(y, \epsilon)) > 0. \tag{8.1}$$

To see this, note that, for $\omega \in M(\epsilon) := \left\{\omega \in \Omega \,\middle|\, P_{\xi(\omega)}x \in \overline{\mathbb{B}}(y, \epsilon)\right\}$, we have

$$\begin{aligned}\left\|P_{\xi(\omega)}z - y\right\| &\leq \left\|P_{\xi(\omega)}z - P_{\xi(\omega)}y\right\| + \left\|P_{\xi(\omega)}y - y\right\| \\ &\leq \|z - y\| + \left\|P_{\xi(\omega)}y - y\right\| \\ &\leq \|z - y\| + \left\|P_{\xi(\omega)}x - y\right\| \\ &\leq 2\epsilon.\end{aligned}$$

Here we have used nonexpansiveness of P_ξ and the definition of a projection. Now (8.1) follows from the identity $\mathbb{P}(M_k(\epsilon)) = p(x, \overline{\mathbb{B}}(y, \epsilon)) > 0$.
Furthermore, one has for $\epsilon > 0$ that

$$(\forall w \in \overline{\mathbb{B}}(x, \epsilon)) \qquad p(w, \overline{\mathbb{B}}(y, 2\epsilon)) \geq p(x, \overline{\mathbb{B}}(y, \epsilon)) > 0. \tag{8.2}$$

To see this, note that for $\omega \in M(\epsilon)$, we have

$$\begin{aligned}\left\|P_{\xi(\omega)}w - y\right\| &\leq \left\|P_{\xi(\omega)}w - P_{\xi(\omega)}x\right\| + \left\|P_{\xi(\omega)}x - y\right\| \\ &\leq 2\epsilon.\end{aligned}$$

Now, fix $\epsilon > 0$ such that both $\overline{\mathbb{B}}(y, \epsilon) \cap C = \emptyset$ and $\overline{\mathbb{B}}(x, \epsilon) \cap C = \emptyset$. We get for any $w \in \mathcal{H}$ and $n \in \mathbb{N}$

$$p^{n+1}(w, \overline{\mathbb{B}}(y, \epsilon)) \geq \int_{\overline{\mathbb{B}}(y, \frac{\epsilon}{2})} p(z, \overline{\mathbb{B}}(y, \epsilon))p^n(w, \mathrm{d}z) \geq p(x, \overline{\mathbb{B}}(y, \tfrac{\epsilon}{2}))p^n(w, \overline{\mathbb{B}}(y, \tfrac{\epsilon}{2})).$$

So iteratively, denoting $\epsilon_n := 2^{-n}\epsilon$, we arrive at

$$p^{n+1}(w, \overline{\mathbb{B}}(y, \epsilon)) \geq \prod_{i=1}^{n} p(x, \overline{\mathbb{B}}(y, \epsilon_i))p(w, \overline{\mathbb{B}}(y, \epsilon_n)).$$

The last probability can be estimated for $w \in \overline{\mathbb{B}}(x, \epsilon_{n+1})$ by (8.2) through

$$p(w, \overline{\mathbb{B}}(y, \epsilon_n)) \geq p(x, \overline{\mathbb{B}}(y, \epsilon_{n+1})).$$

Summarizing, we have that $p^n(w, \overline{\mathbb{B}}(y, \epsilon))$ is locally uniformly bounded from below for $w \in \overline{\mathbb{B}}(x, \epsilon_n)$. That implies

$$\begin{aligned}\mathbb{P}(X_n \in \mathcal{H} \setminus C) &= \int_{\mathcal{H}} p^n(w, \mathcal{H} \setminus C)\mu(\mathrm{d}w) \\ &\geq \int_{\overline{\mathbb{B}}(x, \epsilon_n)} p^n(w, \overline{\mathbb{B}}(y, \epsilon))\mu(\mathrm{d}w) \\ &\geq [p(x, \overline{\mathbb{B}}(y, \epsilon_n))]^n \mu(\overline{\mathbb{B}}(x, \epsilon_n)) > 0,\end{aligned}$$

i.e. $\mathbb{P}(X_n \in C) < 1$ for all $n \in \mathbb{N}$, as claimed.

□

Remark 8.1.5: Theorem 8.1.4 can be interpreted as a lower bound on the complexity of the RFI analogous to the deterministic case [36, Theorem 5.2], where the alternating projection algorithm converges either after one iteration or after infinitely many. Alternatively, the *stopping* or *hitting time* of a process is defined as

$$T := \inf\{n \mid X_n \in C\}.$$

In this context, Theorem 8.1.4 says that, either $\mathbb{P}(T = 1) = 1$ or $\mathbb{P}(T = n) < 1$ for all $n \in \mathbb{N}$. Note, it could happen that $\mathbb{P}(T = \infty) = 1$, in which case $\mathbb{P}(T = n) = 0$ for all $n \in \mathbb{N}$.

Example 8.1.6 (finite and infinite convergence)**.** With just two sets, the deterministic alternating projections algorithms can converge in finitely many steps. But when the projections onto the respective sets are randomly selected, convergence might only come after infinitely many steps. For example, let $C_1 = \mathbb{R}_+ \times \mathbb{R}$ and $C_2 = \mathbb{R} \times \mathbb{R}_+$ and $\mathbb{P}(\xi = 1) = 0.3$, $\mathbb{P}(\xi = 2) = 0.7$. Then $C = \mathbb{R}_+ \times \mathbb{R}_+$. Set $\mu = \delta_x$, where $x = \binom{-1}{-1}$. Then $\mathbb{P}(X_1 \in C) = 0$ or more generally $\mathbb{P}(X_n \in C) = 1 - 0.3^n - 0.7^n < 1$, $n \in \mathbb{N}$. Now let $\mathbb{P}(\xi = 1) = 1$ and $\mathbb{P}(\xi = 2) = 0$, then $C = C_1$ and for μ as above $\mathbb{P}(X_1 \in C) = 1$ and so $\mathbb{P}(X_n \in C) = 1$.

Example 8.1.7 (no uniform geometric convergence)**.** In this example we show a sublinear convergence rate for infinitely many overlapping intervals. This is in contrast to the convergence properties of finitely many intervals with nonempty interior, where one would expect a geometric rate.

Let $\xi \sim \mathrm{unif}[\epsilon - \frac{1}{2}, \frac{1}{2} - \epsilon]$ for some $\epsilon \in [0, \frac{1}{2})$. Define the nonempty and closed intervals $C_r = [r - \frac{1}{2}, r + \frac{1}{2}]$, $r \in \mathbb{R}$.

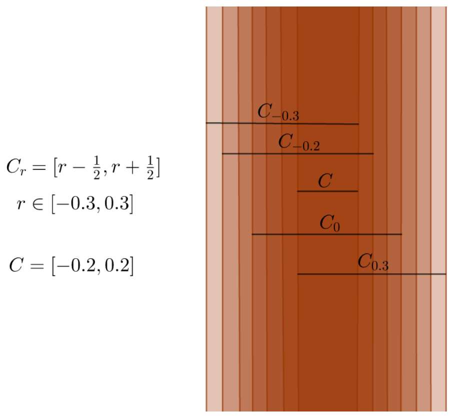

Figure 8.1:

The projector onto these intervals is given by

$$P_r x = \begin{cases} r + \frac{1}{2} & x \geq r + \frac{1}{2} \\ r - \frac{1}{2} & x \leq r - \frac{1}{2} \\ x & r - \frac{1}{2} \leq x \leq r + \frac{1}{2} \end{cases}.$$

The Lebesgue-density ρ_ϵ of ξ is

$$\rho_\epsilon(y) = \frac{1}{1-2\epsilon}\mathbb{1}_{[\epsilon-\frac{1}{2},\frac{1}{2}-\epsilon]}(y).$$

One can compute

$$\begin{aligned} R_\epsilon(x) = \mathbb{E}[|P_\xi x - x|^2] &= \int_{\mathbb{R}} |P_r x - x|^2 \rho_\epsilon(r)\,\mathrm{d}r = \frac{1}{1-2\epsilon}\int_{\epsilon-\frac{1}{2}}^{\frac{1}{2}-\epsilon} |P_r x - x|^2\,\mathrm{d}r \\ &= \frac{1}{1-2\epsilon}\mathbb{1}_{[\epsilon,\infty)}(|x|)\frac{(|x|-\epsilon)^3 + \min(1-|x|-\epsilon,0)^3}{3}. \end{aligned}$$

Now, let us examine regularity properties. For the case $\epsilon \in [0, \frac{1}{2})$, the problem is a consistent feasibility problem with $C = [-\epsilon, \epsilon]$. While the regularity condition in Eq. (5.1) is trivially satisfied for $|x| \leq \epsilon$, for $\epsilon \leq |x| \leq 1-\epsilon$ we find $R_\epsilon(x) = \frac{1}{1-2\epsilon}\frac{(|x|-\epsilon)^3}{3}$ and $\mathrm{dist}^2(x, C) = (|x|-\epsilon)^2$. So the regularity property in Eq. (5.1) is not satisfied for any $\kappa > 0$ here. That means by Theorem 5.0.9, that we cannot expect uniform geometric convergence (i.e. there is no $r \in [0, 1)$ with $\mathbb{E}[\mathrm{dist}(X_{k+1}, C)] \leq r\mathbb{E}[\mathrm{dist}(X_k, C)]$, where $X_0 \sim \delta_x$, $x \in \mathcal{H}$).

Example 8.1.8 (uniform geometric convergence)**.** We provide here a concrete example where geometric convergence of the RFI is achieved. This is somewhat surprising since the angle between the sets can become arbitrarily small. In the deterministic setting, this results in arbitrarily slow convergence of the algorithm. This provides some intuition for why stochastic algorithms can outperform deterministic variants.

Let $C_\alpha := \mathbb{R}e_\alpha$ with $e_\alpha = \binom{\cos(\alpha)}{\sin(\alpha)}$, $\alpha \in [0, 2\pi)$ and let $\xi \sim \mathrm{unif}[0, \beta]$, where $\beta \in (0, \frac{\pi}{2})$.

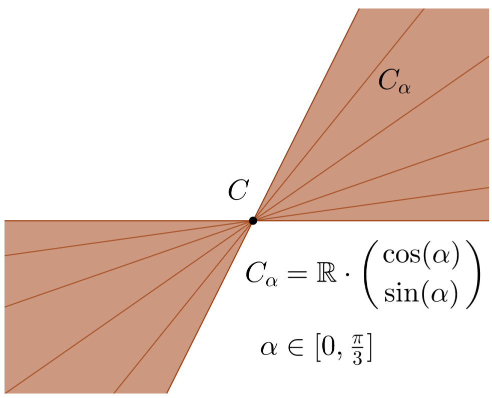

Figure 8.2:

We have $C = \{0\}$, so $\mathrm{dist}(x, C) = \|x\|$ and the density of ξ is $\rho_\beta(\alpha) = \frac{1}{\beta}\mathbb{1}_{[0,\beta]}(\alpha)$. The projector onto the linear subspace C_α of $\mathbb{R}^2$ is given by

$$P_\alpha(x) = x - \left\langle \begin{pmatrix}\sin(\alpha)\\ -\cos(\alpha)\end{pmatrix}, x\right\rangle \begin{pmatrix}\sin(\alpha)\\ -\cos(\alpha)\end{pmatrix}.$$

We find then

$$R_\beta(x) = \frac{1}{\beta}\int_0^\beta \|P_\alpha x - x\|^2\,\mathrm{d}\alpha = \frac{1}{\beta}\int_0^\beta (x_1\sin(\alpha) - x_2\cos(\alpha))^2\,\mathrm{d}\alpha$$
$$= \frac{1}{\beta}\left[x_1^2\left(\frac{\beta-\sin(\beta)\cos(\beta)}{2}\right) + x_2^2\left(\frac{\beta+\sin(\beta)\cos(\beta)}{2}\right) - x_1x_2\sin^2(\beta)\right].$$

Using that for $x = \lambda e_\alpha$ with $\lambda \geq 0$ holds $\operatorname{dist}(x, C) = \lambda$ and $R_\beta(x) = \lambda^2 R_\beta(e_\alpha)$ and employing trigonomertric calculation rules, we find the regularity constant in Eq. (5.1) not to be smaller than

$$\kappa = \sup_{x\in\mathbb{R}^2}\frac{\operatorname{dist}^2(x,C)}{R_\beta(x)} = \sup_{\alpha\in[0,2\pi)}\frac{8\beta}{2\beta-\sin(2\beta-2\alpha)-\sin(2\alpha)} = \frac{4\beta}{\beta-\sin(\beta)},$$

where the last supremum is attained at $\alpha = \frac{\beta}{2}$. So from Theorem 5.0.5 we get uniform geometric convergence.

Example 8.1.9 (disks on a circle)**.** This example illustrates Theorem 8.1.3. Let $C_\alpha := \overline{\mathbb{B}}(\rho e_\alpha, 1) \subset \mathbb{R}^2$, where $0 < \rho < 1$ and $e_\alpha = \binom{\cos(\alpha)}{\sin(\alpha)}$, $\alpha \in [0, 2\pi)$ and let $\xi \sim \operatorname{unif}[0, 2\pi]$.

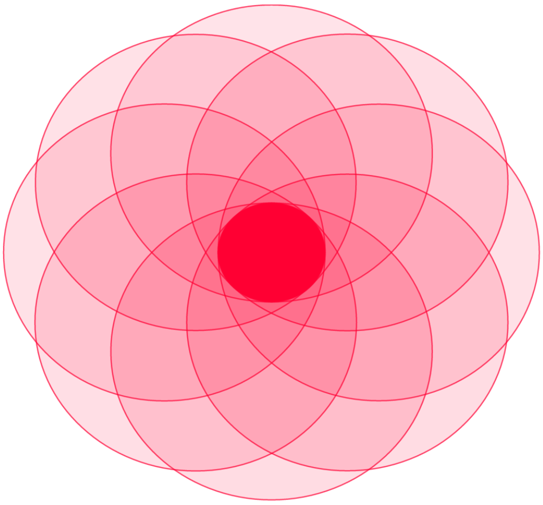

Figure 8.3:

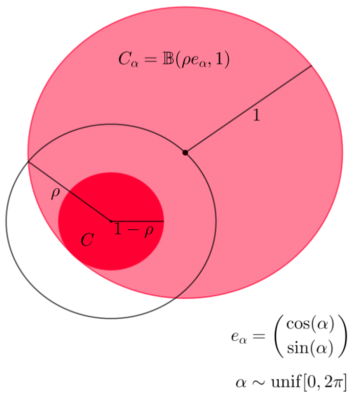

Figure 8.4:

The intersection is given by $C = \overline{\mathbb{B}}(0, 1-\rho)$. We show next that sets with this configuration do not satisfy (5.1). To see this we show that there is a sequence $(x_n)_n \subset \mathbb{R}^2$ with $\mathbb{P}(x_n \in C_\xi) \to 1$ as $n \to \infty$. By Theorem 8.1.3 we conclude that (5.1) cannot hold. Indeed, let $x = x(\lambda) = \lambda\binom{1}{0}$ with $\lambda \geq 1-\rho$, then

$$\mathbb{P}(x \in C_\xi) = \frac{1}{2\pi}\int_0^{2\pi} \mathbb{1}\{\|x-\rho e_\alpha\| \leq 1\}\,\mathrm{d}\alpha$$
$$= \frac{1}{2\pi}\int_0^{2\pi} \mathbb{1}\{\lambda^2+\rho^2-2\lambda\rho\cos(\alpha) \leq 1\}\,\mathrm{d}\alpha$$
$$= \frac{1}{2\pi}\int_{-\beta}^{\beta} 1\,\mathrm{d}\alpha$$
$$= \frac{\beta}{\pi},$$

where $\beta = \beta(\lambda) = \cos^{-1}\left(\frac{\lambda^2+\rho^2-1}{2\lambda\rho}\right)$, if $\lambda \leq 1+\rho$. We have $\beta(\lambda) \to \pi$ as $\lambda \to 1-\rho$, so $\mathbb{P}(x(\lambda) \in C_\xi) \to 1$ as $\lambda \to 1-\rho$.

In contrast to the case where $\rho \in (0,1)$, the extreme cases where $\rho = 0$ and $\rho = 1$ *do* satisfy (5.1). Indeed, if $\rho = 0$, since $C_\alpha = C = \overline{\mathbb{B}}(0,1)$ for all $\alpha \in [0, 2\pi)$, then we have $R(x) = \operatorname{dist}^2(x, C)$, i.e. $\kappa = 1$, and (5.1) holds.

On the other hand, if $\rho = 1$, i.e. $C = \{0\}$, one has for $x = \lambda e_\gamma$, where $\lambda > 0$ and $\gamma \in [0, 2\pi)$

$$R(x) = \left(1 - \frac{\beta}{\pi}\right)(\lambda^2 + 2) + 2\frac{\lambda}{\pi}\sqrt{1 - \frac{\lambda^2}{4}} - \frac{1}{\pi}\int_\beta^{2\pi-\beta} \sqrt{\lambda^2 + 1 - 2\lambda\cos(\alpha)}\, \mathrm{d}\alpha\,,$$

where $\beta = \beta(\lambda) = \cos^{-1}(\min(\lambda/2, 1))$. Note that this expression is rotationally symmetric, i.e. independent of γ. In order for this collection of sets to satisfy Eq. (5.1), we need $\sup_{x \notin C} \operatorname{dist}^2(x, C)/R(x)$ to be finite. Since by (i) and (ii) of Lemma 4.3.1 $R \geq 0$ is finite everywhere and 0 only in C we need just to consider the limits $\lambda \to 0$ and $\lambda \to \infty$ of $\operatorname{dist}^2(x, C)/R(x)$ with $x = \lambda e_\gamma$. The Taylor expansion of the square root expression at 0 with respect to λ yields

$$f(\lambda, \alpha) := \sqrt{\lambda^2 + 1 - 2\lambda\cos(\alpha)} = 1 - \cos(\alpha)\lambda + \frac{1}{2}\lambda^2(1 - \cos^2(\alpha)) + \mathcal{O}(\lambda^3), \quad \lambda \to 0$$

and hence

$$\frac{1}{\pi}\int_\beta^{2\pi-\beta} f(\lambda, \alpha)\, \mathrm{d}\alpha = \left(1 - \frac{\beta}{\pi}\right)\left(2 + \frac{\lambda^2}{2}\right) + 2\frac{\lambda}{\pi}\sqrt{1 - \frac{\lambda^2}{4}} + \mathcal{O}(\lambda^3), \qquad \lambda \to 0.$$

The error of the above expression is of order $\mathcal{O}(\lambda^3)$, because the integral of the error term of the Taylor approximation is

$$-\frac{1}{2\pi}\int_\beta^{2\pi-\beta} \frac{\partial^3}{\partial\lambda^3} f(\theta\lambda, \alpha)\, \mathrm{d}\alpha \leq \left(1 - \frac{\beta}{\pi}\right)\frac{1 + \theta\lambda}{|\theta\lambda - 1|^5},$$

where $\theta \in [0, 1]$. So in the limit $\lambda \to 0$, using $\beta \leq \frac{\pi}{2}$, $\operatorname{dist}^2(x, C)/R(x)$ is bounded from above by 4. In the case $\lambda \to \infty$, we may let $\lambda > 2$ and so $\beta = 0$. We get

$$R(x) = \lambda^2 + 2 - \frac{1}{\pi}\int_0^{2\pi} \sqrt{\lambda^2 + 1 - 2\lambda\cos(\alpha)}\, \mathrm{d}\alpha\,.$$

Since $\cos(\alpha) \geq -1$, it follows that

$$R(x) \geq \lambda^2 - 2\lambda,$$

yielding the finite limit 1 for $\kappa = \operatorname{dist}^2(x, C)/R(x)$. Since $x = \lambda e_\gamma$ is continuous in λ and $x \mapsto \operatorname{dist}^2(x, C)/R(x)$ is also nonnegative, continuous and bounded as a function of λ on $[0, +\infty)$, then this shows that κ is finite, hence (5.1) holds.

8.1.2 RFI with two families of mappings

The set feasibility examples above lead very naturally to the more general context of mappings $T_i : G \to G$, $i \in I$ and $S_j : G \to G$, $j \in J$ on a metric space (G, d), where I, J are arbitrary index sets. Here we envision the scenario where $C_T := \{x \in G \,|\, \mathbb{P}(x \in \operatorname{Fix} T_\xi) = 1\}$ and $C_S := \{x \in G \,|\, \mathbb{P}(x \in \operatorname{Fix} S_\zeta) = 1\}$ are distinctly different sets, possibly nonintersecting. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\xi : \Omega \to I$, $\zeta : \Omega \to J$ be two random variables. Let $(\xi_n)_{n\in\mathbb{N}}$ be an iid sequence with $\xi_n \overset{\mathrm{d}}{=} \xi$ and $(\zeta_n)_{n\in\mathbb{N}}$ iid with $\zeta_n \overset{\mathrm{d}}{=} \zeta$. The two sequences are assumed to be independent of each other. Let μ be a probability measure on $(G, \mathcal{B}(G))$. Consider the stochastic selection method for two families of mappings

Algorithm 2 RFI for two families of mappings

Initialization: $X_0 \sim \mu$
  for $k = 0, 1, 2, \dots$ **do**
    $X_{k+1} = S_{\zeta_k} T_{\xi_k} X_k$
  return $\{X_k\}_{k\in\mathbb{N}}$

Note, that this structure of two families of mappings is a special case of Algorithm 1, just set $\tilde{T}_{(i,j)} = S_j T_i$, where $(i, j) \in \tilde{I} = I \times J$ and $\tilde{\xi} = (\xi, \zeta)$. Also the Markov chain property is still satisfied, the transition kernel takes the form $p(x, A) = \mathbb{P}(S_\zeta T_\xi x \in A)$ for $x \in G$ and $A \in \mathcal{B}(G)$. An advantage of this formulation is that properties of the two families $\{S_j\}_{j\in J}$ and $\{T_i\}_{i\in I}$ can be analyzed more specifically, and independently. As long as the mapping $\tilde{T}$ satisfies the properties needed for convergence of the RFI, then convergence of the RFI for two families of mappings follows. At the very least, we need

$$C := \left\{x \in G \,\middle|\, \mathbb{P}(x \in \operatorname{Fix} \tilde{T}_{\tilde{\xi}}) = 1\right\} \neq \emptyset.$$

From this it is easy to see that for convergence the set C_T *could* be empty, but the set C_S *must* be nonempty.

Example 8.1.10 (consistent feasibility)**.** Revisit Example 8.1.6. We had $C_1 = \mathbb{R}_+ \times \mathbb{R}$ and $C_2 = \mathbb{R} \times \mathbb{R}_+$. Set $I = \{1\}$ and $J = \{2\}$, then the algorithm is the deterministic alternating projections method. One has $\mathbb{P}(X_1 \in C) = 1$ for all initial distributions.

Example 8.1.11 (inconsistent stochastic feasibility)**.** In this example we show that the framework established here is not exclusively limited to consistent feasibility. Consider the (trivially convex, nonempty and closed) set $S := \{(0, 10)\}$ together with the collection of balls in Example 8.1.9, $C_\alpha := \overline{\mathbb{B}}(\rho e_\alpha, 1) \subset \mathbb{R}^2$, where $0 \le \rho \le 1$ and $e_\alpha = \binom{\cos(\alpha)}{\sin(\alpha)}$, $\alpha \in [0, 2\pi)$ and let $\xi \sim \operatorname{unif}[0, 2\pi]$. The intersection of the disks is given by $C_T = \overline{\mathbb{B}}(0, 1-\rho)$ where $T_\alpha := P_{C_\alpha}$ is the metric projection onto C_α. Although $S \cap C_\alpha = \emptyset$ for all $\alpha \in [0, 2\pi)$, still the fixed point set for the mapping in Algorithm 2 (where $S_\zeta = P_S$) is $C = \{(0, 10)\}$, and this is found after one iteration for any initial probability distribution μ, where $X_0 \sim \mu$.

This is indeed a special example, but points to the richness of inconsistent stochastic feasibility, which will be studied in greater depth in Section 8.2.

8.1.3 Linear Operator equations

There are several applications of the RFI to the feasibility problem [15], [14]. We want to focus first on linear operator equations in the separable Hilbert space $\mathcal{H} = L_2([a,b])$. Let $T : \mathcal{H} \to \mathcal{H}$ be a bounded linear operator, we want to find $x \in \mathcal{H}$, such that

$$Tx = g,$$

for a given $g \in \mathcal{H}$. Clearly this is possible only if $g \in R(T)$. The idea in [14] to solve $Tx = g$ is to consider the family of evaluation mappings $\varphi_t : \mathcal{H} \to \mathbb{R}$, $t \in [a,b]$, which are given by

$$\varphi_t(x) := (Tx)(t).$$

Define the affine subspaces $C_t := \{x \in \mathcal{H} \,|\, \varphi_t(x) = g(t)\}$, $t \in [a,b]$. Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P}) = ([a,b], \mathcal{B}([a,b]), \frac{\lambda}{b-a})$, where λ is the Lebesgue-measure. Let $\xi : (\Omega, \mathcal{F}, \mathbb{P}) \to ([a,b], \mathcal{B}([a,b]))$ be a random variable with $\mathbb{P}^\xi = \mathbb{P} = \frac{\lambda}{b-a}$. Then for $g \in R(T)$, we have that

$$Tx = g \qquad \text{if and only if} \qquad x \in C := \{y \in \mathcal{H} \,|\, \mathbb{P}(y \in C_\xi) = 1\}.$$

So the linear operator equation becomes a stochastic feasibility problem.

In order to be able to compute projections onto the sets C_t, $t \in [a,b]$, we need the evaluation functionals φ_t to be continuous, i.e. $\|\varphi_t\| < \infty$ for almost all $t \in [a,b]$. By the Riesz representation theorem there exists a unique $u_t \in \mathcal{H}$ with $\varphi_t(x) = \langle u_t, x\rangle$ for all $x \in \mathcal{H}$ and almost all $t \in [a,b]$. We conclude that the projection onto C_t takes the form

$$P_t x = x + \frac{g(t) - (Tx)(t)}{\|u_t\|^2} u_t \qquad x \in L_2([a,b]).$$

Example 8.1.12 (linear integral equations)**.** Concretely, consider an integral equation of the first kind in the separable Hilbert space $L_2([a,b])$

$$(Tx)(t) = \int_a^b K(t,s)x(s)\,\mathrm{d}s = g(t) \qquad t \in [a,b],$$

with $g \in L_2([a,b])$. For $K \in L_2([a,b] \times [a,b])$, T is a continuous linear compact operator [1, Theorem 8.15] (a Hilbert-Schmidt operator). For the Riesz representation of the evaluation functionals we have that $\varphi_t(x) = (Tx)(t) = \langle u_t, x\rangle$, $t \in [a,b]$, as well as $u_t = K(t,\cdot)$ and hence $\|\varphi_t\| \le \|K(t,\cdot)\| < \infty$.

Example 8.1.13 (differentiation)**.** Let $K(t,s) = \mathbb{1}_{[a,t]}(s) = u_t(s)$, i.e. $(Tx)(t) = \int_a^t x(s)\,\mathrm{d}s$ and suppose $g \in C^1([a,b])$, then $Tx = g$ if and only if $x = g'$ almost surely and $g(a) = 0$. The projectors take the form

$$P_t x = x - \frac{g(t) - \int_a^t x(s)\,\mathrm{d}s}{t-a} \mathbb{1}_{[a,t]}.$$

The meaning of the operator equation can be extended for $g \in R(T) \oplus R(T)^\perp$ as done in inverse problems. In that case the optimization problem

$$\min_{x \in \mathcal{H}} \|Tx - g\|^2$$

is solvable and a solution $x \in \mathcal{H}$, called least squares solution, satisfies the normal equation

$$T^*Tx = T^*g,$$

which is again a linear operator equation in $\mathcal{H}$. There is a unique element in the set of least squares solution with minimal norm. One often denotes this solution the best approximate or minimal norm least squares solution $x^\dagger = T^\dagger g \in N(T)^\perp = \overline{R(T^*)}$, where $T^\dagger$ is the linear operator that maps g to the solution of the normal equation with minimal norm, it holds $D(T^\dagger) = R(T) \oplus R(T)^\perp$. Note that $T^\dagger$ is unbounded if the problem is ill-posed, i.e. the dependence of $x^\dagger$ on g is not continuous, which means small noise can lead to huge change for the solution $x^\dagger$. In that case $R(T) \neq \overline{R(T)}$ and $\left\|T^\dagger\right\| = \infty$. So since $\mathcal{H} = \overline{R(T)} \oplus R(T)^\perp$, there will not exist a solution to the normal equation for $g \in \overline{R(T)} \setminus R(T)$, and hence the least squares solution will not exist. So it is enough to consider operator equations of the form $Tx = g$, since the corresponding normal equation has the same structure, where T is replaced by T^*T and g with T^*g, where g is then allowed to be an element in $R(T) \oplus R(T)^\perp$, since $R(T)^\perp = N(T^*)$.

When solving such a problem numerically, it is necessary to discretize the problem in some sense. One idea is to solve the problem on a finite dimensional subspace $\mathcal{H}_m$ spanned by the basis $\{\varphi_1, \dots, \varphi_m\} \subset \mathcal{H}$. Instead of drawing points according to the distribution $\lambda/(b-a)$ on $[a,b]$ there are several alternatives, for example, choosing at random or deterministically m nodes $t_i \in [a,b]$, so that the approximate problem becomes

$$\text{Find } x = \sum_{i=1}^{m} x_i \varphi_i \in \mathcal{H}_m \qquad \text{satisfying} \qquad Tx(t_j) = g(t_j), \quad j = 1, \dots, m.$$

This is equivalent to solving the linear system

$$W\tilde{x} = \tilde{g}$$

with $W_{i,j} = \langle u_{t_i}, \varphi_j \rangle$, $\tilde{x}_j = x_j$ and $\tilde{g}_j = g(t_j)$, $i, j \in \{1, \dots, m\}$.

This approximation thus corresponds to solving a finite dimensional affine feasibility problem (solving a linear system). The projectors $P_i := P_{t_i}$ onto the sets C_i, $i = 1, \dots, m$ become the metric projections onto $C_i = \{x \mid \langle u_{t_i}, x \rangle = (W\tilde{x})_i = \tilde{g}_i\}$. The finite dimensional approximation to the stochastic feasibility problem would thus become to find

$$x \in \bigcap_{i \in \{1,\dots,m\}} C_i.$$

If W is invertible the solution $\tilde{x} \in \mathbb{R}^m$ is unique and the RFI would converge to this solution a.s. for any distribution on $\{t_1, \dots, t_m\}$ with $\mathbb{P}(\xi = t_i) > 0$ for $i = 1, \dots, m$.

Using the stochastic framework to its full potential is with this approach difficult, since the computation of $(Tx)(t)$ needs knowledge of x. A cheap way of approximating the solution x^* through solutions x_n on subspaces $\mathcal{H}_n$ has also to be thought of, which is not to be done in this thesis. Another example where the infinite feasibility problem gives a new model are convex optimization problems as pointed out in [14].

8.2. Inconsistent Feasibility

We give several examples on inconsistent feasibility problems that we will analyze for convergence and if possible for rates of convergence. Since we are interested in inexact problems or problems where the noise has a non negligible influence, we mention beforehand noise models that use to describe errors in general inconsistent convex feasibility problems. We restrict ourselves in the following to the Euclidean space $(G, d) = (\mathbb{R}^n, \|\cdot\|)$.

Definition 8.2.1 (error models)**.** Let $C \subset \mathbb{R}^n$ be convex, closed and nonempty and let P_C denote the metric projection onto C. Let ξ be a random variable in I and denote by $P_{C,\xi}$ the inexact or noisy projector onto C.

(i) We say $P_{C,\xi}$ is given by the *affine noise model*, if $I = \mathbb{R}^n$ and

$$P_{C,\xi}x = P_C x + \xi, \qquad x \in \mathbb{R}^n. \tag{8.3}$$

(ii) We say $P_{C,\xi}$ is given by the *generalized affine noise model*, if there are finite parameters $p \in \mathbb{R}^m$ such that $C = C(p)$, $I = \mathbb{R}^m$ and

$$P_{C,\xi}x = P_{C(p+\xi)}x, \qquad x \in \mathbb{R}^n. \tag{8.4}$$

(iii) We say $P_{C,\xi}$ is given by the *rotational and affine noise model*, if $\{O_\alpha\}_{\alpha\in A}$ is a family of rotation matrices in $\mathrm{SO}(n)$, where A is an index set, $\{c_\alpha\}_{\alpha\in A}$ is a family of points in C, $I = A \times \mathbb{R}^n$ and

$$P_\xi x := P_{R_{\xi_1}C}x + \xi_2, \qquad x \in \mathbb{R}^n,$$

where $(\alpha, x) \mapsto R_\alpha x := O_\alpha(x - c_\alpha) + c_\alpha$ is assumed to be measurable and where $\xi = (\xi_1, \xi_2)$ (also called r.a.n. random variable).

(iv) We say $P_{C,\xi}$ is a *random projector*, if there exist convex closed and nonempty sets $C_i \subset \mathbb{R}^n$, $i \in I \subset \mathbb{R}$, $\xi \sim \mathrm{unif}(I)$ and

$$P_{C,\xi}x = P_{C_\xi}x, \qquad x \in \mathbb{R}^n. \tag{8.5}$$

8.2.1 Contractions in expectation

In Theorem 2.9.2 we reviewed convergence of Markov chains under the often employed assumption that the mappings $\{T_i \,|\, i \in I\}$ are contractions in expectation, i.e. there exists $r \in (0,1)$ such that

$$\mathbb{E}[d(T_\xi x, T_\xi y)] \le r d(x,y) \qquad \forall x,y \in G.$$

Then it could be shown that the Markov operator is a contraction on $\mathscr{P}_1(G)$ equipped with the Wasserstein metric and application of Banach's fixed point theorem yields a nice result on geometric convergence of the distributions $(\mathcal{L}(X_n))$ of the RFI iterates (X_k) to the unique invariant measure for $\mathcal{P}$.

Example 8.2.2 (AR(1) process, affine noise)**.** Let $r \in (0,1)$ and $T : \mathbb{R} \to \mathbb{R}$, $Tx = rx$. Let ξ be a real-valued random variable with $\operatorname{supp} \xi = [a,b]$ for some $a, b \in \mathbb{R}$. Consider the contraction operator $T_\xi x := Tx + \xi$. Then there exists a unique distribution π on $\mathbb{R}$ with $\mathbb{P}^{X_k} = \mu\mathcal{P}^k \to \pi$ for all $\mu \in \mathscr{P}_1(G)$ (in fact we have geometric convergence, i.e. $W(\mu\mathcal{P}^k, \pi) \le r^k W(\mu, \pi)$). In particular for all Dirac's delta distributions δ_x, $x \in \mathbb{R}$ the sequence of distributions of the iterates $\mathbb{P}^{X_k} = \delta_x \mathcal{P}^k$ converges weakly to π.
Furthermore in this special case for the structure of T_ξ, we find for the variances of the iterates $\operatorname{Var} X_k = \frac{1-r^{2k}}{1-r^2} \operatorname{Var} \xi$, since $X_k = r^k x + \sum_{i=0}^{k-1} r^i \xi_{k-1-i}$, where $(\xi_i)_{i\in\mathbb{N}}$ is an i.i.d. sequence with $\xi_i \overset{\mathrm{d}}{=} \xi$. Hence, $\operatorname{Var} \pi = \frac{1}{1-r^2} \operatorname{Var} \xi$. Note that this is still true for any selection rule ξ not necessarily having compact support.

Example 8.2.3 (two lines, generalized affine noise)**.** Let $C_1 = \mathbb{R}\left(\begin{smallmatrix}1\\0\end{smallmatrix}\right)$ and $C_2(\alpha) = \mathbb{R}\begin{pmatrix}\cos(\alpha)\\ \sin(\alpha)\end{pmatrix}$ with $\alpha \in (0, \frac{\pi}{2})$. Then $C_1 \cap C_2 = \{0\}$. Assume now that the noise is: $\xi = (\xi_1, \xi_2)$ with $\xi_1 \in [-\epsilon, \epsilon]^2$ for some $\epsilon > 0$ and $\xi_2 \in [\beta_1, \beta_2]$ with $\beta_1 < \beta_2 \in (0, \frac{\pi}{2})$. Let $T_\xi x = P_{C_1}(P_{C_2(\xi_2)} x + \xi_1)$ for $x \in \mathbb{R}^2$.
The projectors onto the sets C_1, C_2 are linear operators and hence T_ξ is a contraction: From [16, Theorem 9.31] we have for $k \in \mathbb{N}$ that

$$\left\| (P_{C_1} P_{C_2(\alpha)})^k - P_{C_1 \cap C_2} \right\| = \cos^{2k-1}(\alpha) < 1,$$

that means in our case, for any $x, y \in \mathbb{R}^2$

$$\|T_\xi x - T_\xi y\| \le \left\| P_{C_1} P_{C_2(\xi_2)} \right\| \|x - y\| = \cos(\xi_2)\|x-y\| \le \cos(\beta_1)\|x-y\|.$$

It follows from Theorem 2.9.2 that there exists a unique invariant probability measure for $\mathcal{P}$ and the rate of convergence of the laws of the RFI iterates is geometric in the Wasserstein metric. Furthermore we can deduce that π has compact support that is bounded by $\frac{\sqrt{2}\epsilon}{1-\cos(\beta_1)}$ (see Lemma 8.2.4).

Denote by S_π the support of the unique invariant measure $\pi \in \mathscr{P}_1(G)$. We denote for $A \subset G$

$$\operatorname{diam} A := \sup_{x,y\in A} d(x,y) = \sup_{x\in A} \sup_{y \in A} d(x,y).$$

In a normed vector space holds $\max(\operatorname{diam} A, \operatorname{diam} B) \leq \operatorname{diam}(A+B) \leq \operatorname{diam} A + \operatorname{diam} B$. This is due to

$$\max\left(\sup_{a_1,a_2\in A}\|a_1-a_2\|, \sup_{b_1,b_2\in B}\|b_1-b_2\|\right) \leq \sup_{\substack{x,y\in A+B\\ x=a_1+b_1\\ y=a_2+b_2}}\|x-y\| \leq \sup_{\substack{a_1,a_2\in A\\ b_1,b_2\in B}}\|a_1-a_2\|+\|b_1-b_2\|$$

Lemma 8.2.4 (estimation of support)**.** *Consider the affine noise model $T_\xi x := Tx + \xi$ for a contraction $T : \mathbb{R}^n \to \mathbb{R}^n$ with constant $r \in (0,1)$, we will write* $\operatorname{diam}\xi := \operatorname{diam}\operatorname{supp}\mathcal{L}(\xi)$ *in the following. Then one has the following estimates for the diameter of the support*

$$\operatorname{diam}\xi \leq \operatorname{diam} S_\pi \leq \frac{1}{1-r}\operatorname{diam}\xi.$$

Proof. From Corollary 6.1.9 we have that $S_\pi = \overline{\bigcup_{x\in S_\pi}\mathcal{L}(T_\xi x)}$. Since $\mathcal{L}(T_\xi x) = Tx + \operatorname{supp}\xi$ by Lemma 2.4.2, one gets with the above estimates

$$\operatorname{diam}\xi \leq \operatorname{diam} S_\pi \leq \operatorname{diam} TS_\pi + \operatorname{diam}\xi.$$

Using $\operatorname{diam} TS_\pi \leq r\operatorname{diam} S_\pi$ and continuing inductively the estimation of $\operatorname{diam} S_\pi$ yields

$$\operatorname{diam} S_\pi \leq \sum_{i\in\mathbb{N}_0} r^i \operatorname{diam}\xi = \frac{1}{1-r}\operatorname{diam}\xi.$$

□

Note that in the case that the support of ξ is unbounded, also the limiting measure will have unbounded support. Of course, the usual quantity to measure an error of a variable is its variance. It would be interesting to know if similar to Example 8.2.2 an estimation as in Lemma 8.2.4 is possible for the variances instead of diameters. The estimation from below nevertheless is still possible by the trivial fact, that $\operatorname{Var} X_{k+1} = \operatorname{Var}[TX_k + \xi_k] \geq \operatorname{Var}\xi$ for all $k \in \mathbb{N}$ and so $\operatorname{Var}\pi \geq \operatorname{Var}\xi$.

Remark 8.2.5 (Numerical error of fixed point iteration)**:** For the fixed point iteration $x_{k+1} = Tx_k$, $k \in \mathbb{N}$, $x_0 \in \mathbb{R}$ with a contraction $T : \mathbb{R} \to \mathbb{R}$ with constant $r \in (0,1)$, we will presume the model $X_{k+1} = T_\xi X_k := TX_k + \xi_k$ for an i.i.d. sequence (ξ_k), $\xi_k \overset{\mathrm{d}}{=} \xi$, $k \in \mathbb{N}$ to describe the machine error $|\xi| \leq 10^{-15}$ made on a computer in every iteration through approximation of the binary representation of any real number.

For the diameter of the support of the invariant probability measure π, we have by the above estimation, that $\operatorname{diam} S_\pi \leq 2\cdot 10^{-15}/(1-r)$. It is worth to mention that boundedness of the infinite sequence of erroneous iterates is not true for nonexpansive mappings in general, e.g. the nonexpansive operator $x \mapsto T_\xi x = x + \xi$ can produce a divergent sequence, if $\mathbb{E}[\xi] \neq 0$ (this operator could be a model for summing up 0.1 on the computer N-times and subtracting $N/10$ and this sequence would diverge for $N \to \infty$, due to the non-exact binary representation of 0.1). For contractions however, as we have seen above, unboundedness can only occur, if the error ξ would be modeled as not finite.

In the case of the fixed point iteration it is also interesting to know, when $\lim_k x_k \in S_\pi$, because that is the actual value of interest. A sufficient, but not necessary condition for that assertion to be true is $0 \in \operatorname{supp} \mathcal{L}(\xi)$. Because then $\mathbb{P}(\{|\xi| \le \epsilon\}) > 0$ for all $\epsilon > 0$ and hence there is $(\omega_n) \subset \Omega$ with $\omega_n \in \{|\xi| \le \frac{1}{n}\}$, $n \in \mathbb{N}$ such that $TS_\pi + \xi(\omega_n) \subset S_\pi$ by Lemma 6.1.8, implying that $\operatorname{dist}(TS_\pi, S_\pi) = 0$, i.e. $TS_\pi \subset S_\pi$ and that means $T^k s \in S_\pi$ for all $s \in S_\pi$, $k \in \mathbb{N}$ and by closedness of S_π, also the unique limit $x = \lim_k T^k s \in S_\pi$, which is independent of $s \in S_\pi$, in fact even of $s \in G$ by Banach's theorem.

The following example shows that the RFI applied to projections onto arbitrary affine subspaces with a generalized affine noise model possesses an invariant measure and, moreover, the RFI converges geometrically. This is to be contrasted to the usual approach to model inexact computation or uncertainty with affine noise models. Affine noise models are not appropriate for the applications we have in mind (most of which are elementary, like solving systems of linear equations on machines with finite precision arithmetic); these examples show that, at least for existence of fixed points of the Markov operator, more realistic noise models pose no particular challenge. Indeed, simple affine noise models do not even have invariant measures, as demonstrated in Example 2.8.5.

Example 8.2.6 (invariant measure for subspaces)**.** We propose the following generalized affine noise model for a single affine subspace $H_{\xi,\zeta} = \{x \in \mathbb{R}^n \mid \langle a + \xi, x\rangle = b + \zeta\}$, where $a \in \mathbb{S}$ and $b \in \mathbb{R}$ and noise $(\xi, \zeta) \in \mathbb{R}^n \times \mathbb{R}$ is independent. The projector onto $H_{\xi,\zeta}$ is given by

$$P_{(\xi,\zeta)}x = x - \frac{\langle a + \xi, x\rangle - (\zeta + b)}{\|a + \xi\|^2}(a + \xi).$$

We show now that the projector is a contraction in expectation under some mild assumptions on the noise. In particular we assume

$$d := \sup_{a \in \mathbb{S}} \mathbb{E}\left[\frac{(b + \zeta)^2}{\|a + \xi\|^2}\right] < \infty,$$
$$c := \inf_{\substack{a \in \mathbb{S} \\ x \in \mathbb{S}}} \mathbb{E}\left[\langle a + \xi, x\rangle^2\right] > 0.$$

The latter condition is satisfied e.g. if ξ is isotropic or radially symmetric, the former e.g. if ζ has bounded variance and $\|\xi\|$ is bounded away from 1. We find for any $x, y \in \mathbb{R}^n$ that

$$\left\|P_{(\xi,\zeta)}x - P_{(\xi,\zeta)}y\right\|^2 = \|x - y\|^2 - \frac{\langle a_\xi, x - y\rangle^2}{\|a_\xi\|^2} = (1 - \cos^2(a_\xi, x - y))\|x - y\|^2,$$

where $a_\xi := a + \xi$. Taking the expectation and using the assumption yields

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

From Theorem 2.9.2 we get that there exists a unique invariant measure π for $\mathcal{P}$ (even $\pi \in \mathscr{P}_2$) and that it satisfies

$$W_2^2(\mu\mathcal{P}^k, \pi) \le (1 - c)^k W_2^2(\mu, \pi),$$

where $W_2^2(\mu, \pi) = \inf_{X\sim\mu, Y\sim\pi} \mathbb{E}[\|X - Y\|^2]$.
Note that for finitely many distorted affine subspaces (i.e. we are given m normal vectors $a_1, \ldots, a_m \in \mathbb{R}^n$ and displacement vectors $b_1, \ldots, b_m$), the linearity of the projector yields for a cyclic variant of the algorithm, that there also exists an invariant measure and the rate of convergence in the Wasserstein metric is also linear. We denote by P_j the inexact projection onto the j-th affine subspace, i.e.

$$P_j x = x - \frac{\langle a_j + \xi_j, x\rangle - (b_j + \zeta_j)}{\|a_j + \xi_j\|^2}(a_j + \xi_j),$$

where $(\xi_i)_{i=1}^m$ and $(\zeta_i)_{i=1}^m$ are i.i.d. and $(\xi_i) \perp\!\!\!\perp (\zeta_i)$. Furthermore we denote by

$$T_{(\xi,\zeta)}x = P_m \circ \ldots \circ P_1 x, \qquad x \in \mathbb{R}^n$$

the inexact cyclic projector, then

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

Hence, there exists a unique invariant measure and $(\mu\mathcal{P}^k)$ converges geometrically to it in the W_2 metric.

The next example is using random projections onto 3 affine subspaces with empty intersection. Again geometric convergence can only be shown in the Wasserstein metric.

Example 8.2.7 (Global geometric convergence: equilateral triangle)**.** The scenario presented here is a randomized cyclic projection algorithm (see the review [5]). Let $a_1, a_2, a_3 \in \mathbb{R}^2$ be given with $\|a_{i+1} - a_i\| = 1$ for $i = 1, 2, 3$, where $a_4 := a_1$. Define a chart $g_i : [0, 1] \to \mathbb{R}^2$ onto an edge via

$$g_i(\lambda) = a_i + \lambda(a_{i+1} - a_i), \qquad \lambda \in [0, 1] \quad i = 1, 2, 3.$$

Let $C_i := g_i([0, 1])$, and let ξ be a random variable with $\mathbb{P}(\xi = i) = \frac{1}{3}$ for all $i \in I := \{1, 2, 3\}$, so $C = \emptyset$. The projection onto the edges for $x \in \operatorname{conv} A := \{a_1, a_2, a_3\}$ (i.e. the convex hull of a_1, a_2, a_3) is given by

$$P_i x = x - \left\langle x - a_i, (a_{i+1} - a_i)^\perp \right\rangle (a_{i+1} - a_i)^\perp,$$

where we denote $x^\perp := \left(\begin{smallmatrix} -x_2 \\ x_1 \end{smallmatrix}\right)$ for $x = (x_1, x_2) \in \mathbb{R}^2$. Since $P_i x \in A$ for all $x \in \mathbb{R}^2$ and all $i \in I$, we get immediately from Corollary 2.8.6 that there exists an invariant probability measure π for $\mathcal{P}$. We show next that the invariant measure is $\pi = \frac{1}{3}\sum_{i=1}^3 \lambda_{C_i}$, where λ_{C_i} is the Lebesgue measure for the corresponding manifold, i.e.

$$\lambda_{C_i}(B) = \int_{g_i([0,1])} \mathbb{1}_B(x)\lambda_{C_i}(\mathrm{d}x) = \int_0^1 \mathbb{1}_B(g_i(\lambda))\,\mathrm{d}\lambda\,, \qquad B \in \mathcal{B}(\mathbb{R}^2).$$

First of all we have

$$\begin{aligned}\left\langle a_i - a_{i-1}, (a_{i+1}-a_i)^\perp \right\rangle (a_{i+1}-a_i)^\perp &= a_i - a_{i-1} + \frac{1}{2}(a_{i+1}-a_i)\\ &= a_{i+1} - a_{i-1} - \frac{1}{2}(a_{i+1}-a_i)\\ &= \left\langle a_{i+1} - a_{i-1}, (a_{i+1}-a_i)^\perp \right\rangle (a_{i+1}-a_i)^\perp,\end{aligned}$$

where $a_0 = a_3$. With that one can derive the following rules for composing the projections P_i with the charts g_i, $i = 1, 2, 3$

$$\begin{aligned} P_i \circ g_i &= \mathrm{Id}\\ P_i \circ g_{i+1}(\cdot) &= g_i\left(1 - \frac{\cdot}{2}\right)\\ P_i \circ g_{i-1}(\cdot) &= g_i\left(\frac{\cdot}{2}\right),\end{aligned}$$

where $g_0 := g_3$ and $g_4 := g_1$. An invariant measure is given by $\pi(A) = \frac{1}{3}\sum_{i=1}^3 \lambda_{C_i}(A)$, since with the transition kernel $p(x, A) = \frac{1}{3}\sum_{i=1}^3 \mathbb{1}_A(P_i x)$ we get that

$$\begin{aligned}\pi\mathcal{P}(A) &= \frac{1}{3}\sum_{i=1}^3 \int_{\mathbb{R}^2} \mathbb{1}_A(P_i x)\pi(\mathrm{d}x)\\ &= \frac{1}{9}\sum_{i,j=1}^3 \int_{\mathbb{R}^2} \mathbb{1}_A(P_i x)\lambda_{C_j}(\mathrm{d}x)\\ &= \frac{1}{9}\sum_{i,j=1}^3 \int_0^1 \mathbb{1}_A(P_i(g_j(\lambda)))\,\mathrm{d}\lambda\\ &= \frac{1}{9}\sum_{i}^3 \lambda_{C_i}(A) + 2\lambda_{C_i}(A \cap g_i([\tfrac{1}{2}, 1])) + 2\lambda_{C_i}(A \cap g_i([0, \tfrac{1}{2}]))\\ &= \frac{1}{3}\sum_{i=1}^3 \lambda_{C_i}(A)\\ &= \pi(A).\end{aligned}$$

To get a geometric rate we proceed as in Example 8.2.6 and get that

$$\|P_i x - P_i y\|^2 = \|x-y\|^2 - \left\langle (a_{i+1}-a_i)^\perp, x-y\right\rangle^2 = (1 - \cos^2((a_{i+1}-a_i)^\perp, x-y))\|x-y\|^2.$$

So, with

$$\cos^2\left(\left(\begin{smallmatrix}1\\0\end{smallmatrix}\right), b\right) + \cos^2\left(\begin{pmatrix}-\frac{1}{2}\\ \frac{\sqrt{3}}{2}\end{pmatrix}, b\right) + \cos^2\left(\begin{pmatrix}-\frac{1}{2}\\ -\frac{\sqrt{3}}{2}\end{pmatrix}, b\right) = \frac{3}{2}$$

for any $b \in \mathbb{R}^2 \setminus \{0\}$ and hence

$$\mathbb{E}\left[\cos^2((a_{\xi+1}-a_\xi)^\perp, x-y))\right] = \frac{1}{2}\mathbb{1}\{x \neq y\},$$

we get that P_ξ is a contraction in expectation, i.e. for all $x, y \in \mathbb{R}^2$

$$\mathbb{E}\left[\|P_\xi x - P_\xi y\|^2\right] = \frac{1}{2}\|x - y\|^2.$$

So, we get that

$$W_2^2(\mu\mathcal{P}^k, \pi) \leq \left(\frac{1}{2}\right)^k W_2^2(\mu, \pi)$$

for any $\mu \in \mathscr{P}(\mathbb{R}^2)$. (Note, that $W_2(\mu, \pi)$ could be infinite, but it is always finite for $\mu \in \mathscr{P}(A)$). Hence the rate of convergence is geometric in the Wasserstein metric and π is the unique invariant measure for $\mathcal{P}$. However, one can show that geometric convergence in the TV-norm can not occur for this example.

8.2.2 Convergence in TV-norm

The following example is the inconsistent instance of Example 8.1.7, where the consistent problem was shown to have uncountably many invariant measures and the RFI has no uniform geometric convergence. In comparison to Example 8.1.7, this example shows that the convergence properties of the RFI change drastically when the feasibility problem is inconsistent. Here uniform geometric convergence and a unique invariant measure are present.

Example 8.2.8 (Global geometric convergence: overlapping intervals)**.** Let $\xi \sim \mathrm{unif}[-\frac{1}{2}-\epsilon, \frac{1}{2}+\epsilon]$ for some $\epsilon > 0$. We apply here random projections on the nonempty and closed intervals $C_r = [r - \frac{1}{2}, r + \frac{1}{2}]$, where $r \in [-\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon]$. The projector onto these intervals is given by

$$P_r x = \begin{cases} r + \frac{1}{2} & x \geq r + \frac{1}{2} \\ r - \frac{1}{2} & x \leq r - \frac{1}{2} \\ x & r - \frac{1}{2} \leq x \leq r + \frac{1}{2} \end{cases}, \qquad x \in \mathbb{R}.$$

The density ρ of the uniform distribution on $[-\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon]$

$$\rho(y) = \frac{1}{1 + 2\epsilon}\mathbb{1}_{[-\frac{1}{2}-\epsilon, \frac{1}{2}+\epsilon]}(y)$$

determines the distribution function

$$F(x) = \begin{cases} 0 & x \leq -\frac{1}{2} - \epsilon \\ \frac{1}{1+2\epsilon}(x + \epsilon + \frac{1}{2}) & -\frac{1}{2} - \epsilon \leq x \leq \frac{1}{2} + \epsilon \\ 1 & x \geq \frac{1}{2} + \epsilon \end{cases}.$$

With this we can give an explicit expression for the transition kernel

$$\begin{aligned} p(x, A) = \mathbb{P}(P_\xi x \in A) = \quad & \mathbb{P}(\xi \in (A - \tfrac{1}{2}) \cap (-\infty, x - \tfrac{1}{2}]) + \mathbb{P}(\xi \in (A + \tfrac{1}{2}) \cap [x + \tfrac{1}{2}, \infty)) \\ & + \mathbb{P}(\xi \in [x - \tfrac{1}{2}, x + \tfrac{1}{2}])\mathbb{1}_A(x). \end{aligned} \tag{8.6}$$

Let

$$\rho_x^1(y) := \rho(y - \tfrac{1}{2})\mathbb{1}_{(-\infty,x]}(y), \quad \rho_x^2(y) := \rho(y + \tfrac{1}{2})\mathbb{1}_{[x,\infty)}(y),$$

and

$$\rho_x^3(y) := \mathbb{P}(\xi \in [x - \tfrac{1}{2}, x + \tfrac{1}{2}])\delta_x(y).$$

Then the transition kernel in (8.6) can be written equivalently as

$$p(x, A) = \int_A \rho_x^1(y) + \rho_x^2(y) + \rho_x^3(y)\,\mathrm{d}y\,.$$

An invariant distribution π for the corresponding Markov operator $\mathcal{P}$ determined by the mappings P_r in the sense of Eq. (2.6) is given by the density

$$\rho_\pi(x) = \frac{1}{2\epsilon}\mathbb{1}_{[-\epsilon,\epsilon]}(x),$$

since

$$\int_{\mathbb{R}} \rho_x^1(y)\rho_\pi(x)\,\mathrm{d}x = \rho(y - \tfrac{1}{2})\int_y^\infty \rho_\pi(x)\,\mathrm{d}x = \rho(y - \tfrac{1}{2})(\mathbb{1}_{(-\infty,-\epsilon]}(y) + (\epsilon - y)\rho_\pi(y)) = \frac{\epsilon - y}{1 + 2\epsilon}\rho_\pi(y),$$
$$\int_{\mathbb{R}} \rho_x^2(y)\rho_\pi(x)\,\mathrm{d}x = \rho(y + \tfrac{1}{2})\int_y^\infty \rho_\pi(x)\,\mathrm{d}x = \rho(y + \tfrac{1}{2})(\mathbb{1}_{[\epsilon,\infty)}(y) + (y + \epsilon)\rho_\pi(y)) = \frac{\epsilon + y}{1 + 2\epsilon}\rho_\pi(y),$$

and

$$\int_{\mathbb{R}} \rho_x^3(y)\rho_\pi(x)\,\mathrm{d}x = (F(y + \tfrac{1}{2}) - F(y - \tfrac{1}{2}))\rho_\pi(y) = \frac{1}{1 + 2\epsilon}\rho_\pi(y),$$

which, upon application of Fubini's Theorem, yields

$$\pi\mathcal{P}(A) = \int_{\mathbb{R}} p(x, A)\pi(\mathrm{d}x) = \int_A \int_{\mathbb{R}} \rho_x^1(y) + \rho_x^2(y) + \rho_x^3(y)\pi(\mathrm{d}x)\,\mathrm{d}y = \pi(A).$$

For convergence of the Markov chain (i.e. the distributions of the iterates), we will apply a result due to Doeblin about uniform ergodicity of the transition kernel (see Theorem 2.10.3) under the global minorization property that we will now show for the specific kernel at hand.

$$\begin{aligned}
x \geq \epsilon: &\quad p(x, A) \geq \int_A \rho_x^1(y)\mathbb{1}_{[-\epsilon,\epsilon]}(y)\,\mathrm{d}y = \frac{1}{1 + 2\epsilon}\int_A \mathbb{1}_{[-\epsilon,\epsilon]}(y)\,\mathrm{d}y = \frac{2\epsilon}{1 + 2\epsilon}\pi(A)\\
x \leq -\epsilon: &\quad p(x, A) \geq \int_A \rho_x^2(y)\mathbb{1}_{[-\epsilon,\epsilon]}(y)\,\mathrm{d}y = \frac{1}{1 + 2\epsilon}\int_A \mathbb{1}_{[-\epsilon,\epsilon]}(y)\,\mathrm{d}y = \frac{2\epsilon}{1 + 2\epsilon}\pi(A)\\
x \in [-\epsilon, \epsilon]: &\quad p(x, A) \geq \int_A (\rho_x^1 + \rho_x^2)\mathbb{1}_{[-\epsilon,\epsilon]}(y)\,\mathrm{d}y = \frac{1}{1 + 2\epsilon}\int_A \mathbb{1}_{[-\epsilon,\epsilon]}(y)\,\mathrm{d}y = \frac{2\epsilon}{1 + 2\epsilon}\pi(A).
\end{aligned}$$

Thus the global minorization property of p is satisfied for this setup with $\kappa = \frac{2\epsilon}{1+2\epsilon}$ and $\nu = \pi$. Theorem 2.10.3 then yields that π is the unique invariant measure for $\mathcal{P}$, and for any starting distribution μ one has geometric convergence of the distribution of X_n, i.e. $\mathcal{L}(X_n) = \mu\mathcal{P}^n$, to π in the TV-norm (see Lemma 2.10.1 for its definition and further properties).

Remark 8.2.9: When extending Example 8.2.8 into 2 dimensions, such that the sets do not change in the y-direction ($C_r^{2d} = C_r \times \mathbb{R}$), we see that the uniqueness of the invariant distribution fails, but still one can expect geometric convergence to the set of invariant distributions.

The following example shows that the support of an invariant measure does not have to be convex. It also shows that geometric convergence can occur just locally (here that means for all compactly supported initial distributions). Again the appropriate metric to measure distances of probability measures is here the TV-norm.

Example 8.2.10 (Local geometric convergence: half-lines in a circle)**.** We use here random projections onto the half-lines $C_\alpha = (\mathbb{R}_+ + 1)e_\alpha \subset \mathbb{R}^2$ with $e_\alpha = \binom{\cos(\alpha)}{\sin(\alpha)}$, $\alpha \in [0, 2\pi)$. Let $\xi \sim \operatorname{unif}[0, 2\pi)$.
The projector onto C_α is given by

$$P_\alpha x = \begin{cases} \langle x, e_\alpha \rangle \, e_\alpha, & \langle x, e_\alpha \rangle > 1 \\ e_\alpha, & \langle x, e_\alpha \rangle \le 1. \end{cases}$$

The transition kernel $p(x, A) = \mathbb{P}(P_\xi x \in A)$, $x \in \mathbb{R}^2$, $A \in \mathcal{B}(\mathbb{R}^2)$ is explicitly given by

$$p(x, A) = \frac{1}{\lambda_{\mathbb{S}}(\mathbb{S})}(\lambda_{\mathbb{S}}(A \cap H_x) + \lambda_{\mathbb{S}}(\mathbb{1}_A \circ \varphi_x \mathbb{1}_{H_x^c})),$$

where $\lambda_{\mathbb{S}}$ is the Lebesgue measure of the unit circle $\mathbb{S}$,

$$\varphi_x : \mathbb{R}^2 \to \mathbb{R}\,, \quad y \mapsto \langle x, y \rangle \, y$$

and

$$H_x := \left\{ y \in \mathbb{R}^2 \,\middle|\, \langle x, y \rangle \le 1 \right\}.$$

An invariant measure of $\mathcal{P}$ is given by normalizing $\lambda_{\mathbb{S}}$, i.e. $\pi = \lambda_{\mathbb{S}}/\lambda_{\mathbb{S}}(\mathbb{S})$ is invariant:

$$\begin{aligned} \pi\mathcal{P}(A) &= \int_{\mathbb{S}} \pi(A \cap H_x)\pi(\mathrm{d}x) + \int_{\mathbb{S}} \pi(f_{A,x}\mathbb{1}_{H_x^c})\pi(\mathrm{d}x) \\ &= \int_{\mathbb{S}}\int_{\mathbb{S}} \mathbb{1}_A(y)\, \underbrace{\mathbb{1}_{H_x}(y)}_{=1}\, \pi(\mathrm{d}y)\pi(\mathrm{d}x) + \int_{\mathbb{S}}\int_{\mathbb{S}} (\mathbb{1}_A \circ \varphi_x)(y)\, \underbrace{\mathbb{1}_{H_x^c}(y)}_{=0}\, \pi(\mathrm{d}y)\pi(\mathrm{d}x) \\ &= \pi(A). \end{aligned}$$

There are three cases to consider: First, let $\mu \in \mathscr{P}(\overline{\mathbb{B}}(0,1))$. When $x \in \overline{\mathbb{B}}(0,1)$, notice that $p(x, A) = \pi(A)$, so if $X_0 \sim \mu$, then $\mathcal{L}(X_n) = \pi$ for all $n \in \mathbb{N}$. This means that the probability measures $\mathcal{L}(X_n)$ converge after one step.
Consider next the case that $\mu \in \mathscr{P}(\overline{\mathbb{B}}(0, M))$ with $M > 1$. We claim that still the convergence is geometric. To see this, note that for $x, y \in \overline{\mathbb{B}}(0, M)$ we have $\overline{\mathbb{B}}(0, \frac{1}{M}) \subset H_x \cap H_y$, because $\langle x, z \rangle, \langle y, z \rangle \le 1$ for all $z \in \overline{\mathbb{B}}(0, \frac{1}{M})$. But that means the intersection

$H_x \cap H_y$ has nonempty interior around the origin uniformly in x and y, and hence $\lambda_{\mathbb{S}}(H_x \cap H_y) \geq \lambda(M) > 0$, where $\lambda(M)$ is independent of x and y. Now we observe that

$$\begin{aligned}
\|p(x,\cdot) - p(y,\cdot)\|_{\mathrm{TV}} &:= \frac{1}{2} \sup_{|f|\leq 1} |p(x,f) - p(y,f)| \\
&= \frac{1}{2} \sup_{|f|\leq 1} \left| \pi(f \cdot (\mathbb{1}_{H_x} - \mathbb{1}_{H_y})) + \pi(f \circ \varphi_x \mathbb{1}_{H_x^c} - f \circ \varphi_y \mathbb{1}_{H_y^c}) \right| \\
&\leq \frac{1}{2} \left(\pi \left| \mathbb{1}_{H_x} - \mathbb{1}_{H_y} \right| + \pi \left| \mathbb{1}_{H_x^c} - \mathbb{1}_{H_y^c} \right| \right),
\end{aligned}$$

where the suprema are taken over all measurable $f : \mathbb{R}^2 \to [-1,1]$, and

$$\begin{aligned}
\left| \mathbb{1}_{H_x} - \mathbb{1}_{H_y} \right|(z) = 1 - \mathbb{1}_{H_x \cap H_y}(z) - \mathbb{1}_{H_x^c \cap H_y^c}(z) \\
\left| \mathbb{1}_{H_x^c} - \mathbb{1}_{H_y^c} \right|(z) = 1 - \mathbb{1}_{H_x \cap H_y}(z) - \mathbb{1}_{H_x^c \cap H_y^c}(z).
\end{aligned}$$

From this it follows that

$$\begin{aligned}
\|p(x,\cdot) - p(y,\cdot)\|_{\mathrm{TV}} &\leq 1 - \pi(H_x \cap H_y) - \pi(H_x^c \cap H_y^c) \\
&\leq 1 - \lambda(M) < 1.
\end{aligned}$$

The local contraction coefficient $\beta(D)$ for a domain $D \subseteq \mathbb{R}^2$ satisfies (see Lemma 2.10.1)

$$\beta(D) := \sup_{x,y \in D} \|p(x,\cdot) - p(y,\cdot)\|_{\mathrm{TV}} = \sup_{\substack{\mu,\nu \in \mathscr{P}(D) \\ \mu \neq \nu}} \frac{\|\mu\mathcal{P} - \nu\mathcal{P}\|_{\mathrm{TV}}}{\|\mu - \nu\|_{\mathrm{TV}}}.$$

By completeness of $(\mathscr{P}(G), \|\cdot\|_{\mathrm{TV}})$ for any space G, application of Banach's fixed point theorem in the case $\beta(G) < 1$ (meaning $\mathcal{P}$ is a contraction) yields the existence of a unique element $\pi \in \mathscr{P}(G)$ with

$$\|\mu\mathcal{P}^n - \pi\|_{\mathrm{TV}} \leq \beta(G)^n \|\mu - \pi\|_{\mathrm{TV}}, \quad n \in \mathbb{N}.$$

So in our case $\beta(\overline{\mathbb{B}}(0,M)) \leq 1 - \lambda(M) < 1$ and hence for every $M > 1$ there is the same unique invariant measure π as defined above and the Markov chain converges geometrically for any initial probability measure μ with support in $\overline{\mathbb{B}}(0,M)$.
For a general initial measure $\mu \in \mathscr{P}(\mathbb{R}^2)$, the Markov chain still converges, but not necessarily geometrically. Indeed, we find for $f : \mathbb{R}^2 \to [-1,1]$, $x \in \mathbb{R}^2$, that

$$|p^n(x,f) - \pi f| \leq \beta(\|x\|)^n \underbrace{\|\delta_x - \pi\|_{\mathrm{TV}}}_{=1},$$

which implies

$$\|\mu\mathcal{P}^n - \pi\|_{\mathrm{TV}} \leq \int_{\mathbb{R}^2} \beta(\|x\|)^n \mu(\mathrm{d}x) =: \lambda_n,$$

where $\lambda_n \in [0,1]$ and $\lambda_{n+1} \leq \lambda_n$. As a monotone decreasing nonnegative sequence, it converges to a nonnegative number. We claim that $\lambda_n \to 0$, hence the convergence of the

Markov chain. To see this, choose any $\epsilon > 0$ and $M_\epsilon > 0$ such that $\mu(\overline{\mathbb{B}}(0, M_\epsilon)) \geq 1 - \epsilon$. Then

$$\lambda_n = \int_{\mathbb{R}^2} \beta(\|x\|)^n \mu(\mathrm{d}x) \leq \epsilon + \int_{\overline{\mathbb{B}}(0,M_\epsilon)} \beta(\|x\|)^n \mu(\mathrm{d}x) \leq \epsilon + \beta(M_\epsilon)^n.$$

Letting $n \to \infty$, we see that $\lim_n \lambda_n \leq \epsilon$. Since ϵ was arbitrary, this implies that $\lambda_n \to 0$ as $n \to \infty$. So $\mathcal{L}(X_n) = \mu\mathcal{P}^n$ converges to π in TV-norm with rate (λ_n), which is possibly non-geometric.

8.2.3 Other examples

The next example shows that two distorted random sets, where one of them is compact, have an invariant measure for the affine noise model of the RFI. These sets do not have to intersect for any observation. In Example 8.1.11 we considered two nonintersecting convex sets: the sets $S := \{(0, 10)\}$ together with the collection of balls $C_\alpha := \overline{\mathbb{B}}(\rho e_\alpha, 1) \subset \mathbb{R}^2$, where $0 \leq \rho \leq 1$ and $e_\alpha = \binom{\cos(\alpha)}{\sin(\alpha)}$, $\alpha \in [0, 2\pi)$ and let $\xi \sim \operatorname{unif}[0, 2\pi]$. Although $S \cap C_\alpha = \emptyset$ for all $\alpha \in [0, 2\pi)$, still the fixed point set for the composition $P_S P_{C_\alpha}$ is $C = \{(0, 10)\}$, and this is found after one iteration for any initial probability distribution μ, where $X_0 \sim \mu$. With the tools developed in the present study, we can expand this example considerably. We will employ the *rotational and affine noise model* for a set $C \subset \mathbb{R}^n$.

Example 8.2.11 (existence for compact set)**.** Let $C_1, C_2 \subset \mathbb{R}^n$ be closed, convex and nonempty sets. Let C_1 be compact. We will consider the rotational and affine noise model for C_1 and C_2. Let $\xi = (\xi_1, \xi_2)$ be a vector of r.a.n. random variables such that $T_\xi x = P_{\xi_1} P_{\xi_2} x$, $x \in \mathbb{R}^n$. Note that $R_\alpha C_1 \subset \overline{\mathbb{B}}(c, 2 \operatorname{diam} C_1) =: C$ for any $c \in C_1$. By tightness of any probability measure on $\mathbb{R}^n$ we can find an $M > 0$ with $\mathbb{P}(\xi_{1,2} \in \mathbb{B}(0, M)) \geq 1 - \epsilon$ and hence

$$\mathbb{P}(T_\xi x \in \mathbb{B}(C, M)) \geq \inf_{z \in \mathbb{R}^n} \mathbb{P}(P_{R_{\xi_{1,1}} C_1} z + \xi_{1,2} \in \mathbb{B}(C, M)) \geq \mathbb{P}(\xi_{1,2} \in \mathbb{B}(0, M)) \geq 1 - \epsilon$$

for arbitrary $x \in \mathbb{R}^n$. So we have by independence of (ξ_k) that

$$\mathbb{P}(X_{k+1} \in \mathbb{B}(C, M)) = \mathbb{E}[\mathbb{P}\left(T_{\xi_k} X_k \in \mathbb{B}(C, M) \mid X_k\right)] \geq 1 - \epsilon$$

for the RFI sequence (X_k) with arbitrary initial probability measure. In particular we can conclude with Proposition 2.8.3 that there exists an invariant probability measure for $\mathcal{P}$. In particular, if $\xi_{1,2}$ is bounded, then any invariant measure has compact support, see Remark 2.8.4. It is important to point out here that convexity is only used to ensure that T_ξ is single-valued: the projector onto any closed convex set is single-valued, but not so if the set is not convex. It is not difficult to envision a theory of Markov chains for *set-valued* mappings, but this requires building it from scratch, which is beyond the scope of this study.

The next example shows that a more complicated structure for the operator T_ξ still leads to an invariant measure for two compact convex sets and bounded noise.

Example 8.2.12 ($T_{\mathrm{DR}\lambda}$)**.** Let $C_1, C_2 \subset \mathbb{R}^n$ be compact, convex and nonempty sets. Define the operator T_ξ via

$$T_\xi x := T_{\mathrm{DR}\lambda,\xi} x := \frac{\lambda}{2}\left(R_{\xi_1} R_{\xi_2} x + x\right) + (1-\lambda) P_{\xi_2} x,$$

for $x \in \mathbb{R}^n$ and $\lambda \in (0,1)$, where $P_{\xi_i} x := P_{C_i} x + \xi_i$, $i = 1, 2$ and $R_{\xi_i} x := 2P_{\xi_i} x - x$ and where we let $\xi = (\xi_1, \xi_2)$ with $\mathbb{E}[\|\xi_1\|], \mathbb{E}[\|\xi_2\|] \le M_1 < \infty$. Let $M_2 > 0$ be such that $C_1, C_2 \subset \overline{\mathbb{B}}(0, M_2)$ and set

$$M_3 := (\lambda + |1 - 2\lambda|)(M_1 + M_2).$$

Then we have for any $x \in \mathbb{R}^n$ that

$$\begin{aligned}\mathbb{E}[\|T_\xi x\|] &= \mathbb{E}[\|\lambda P_{C_1}(R_{\xi_2} x) + (1-2\lambda) P_{C_2} x + \lambda x + \lambda \xi_1 + (1-2\lambda)\xi_2\|] \\ &\le M_3 + \lambda \|x\|.\end{aligned}$$

Choosing $X_0 \sim \delta_0$, inductively we get for the corresponding RFI sequence (X_k) that

$$\mathbb{E}[\|X_{k+1}\|] = \mathbb{E}\left[\mathbb{E}\left[\|T_{\xi_k} X_k\| \mid X_k\right]\right] \le M_3 + \lambda \mathbb{E}[\|X_k\|] \le M_3 \sum_{i=0}^{k} \lambda^i \le \frac{M_3}{1-\lambda}$$

for any $k \in \mathbb{N}_0$. Now Lemma 2.8.7 is applicable by uniform boundedness of the expectations and yields existence of an invariant measure for $\mathcal{P}$. Note that existence still follows when employing the rotational and affine noise model, since in that case $P_{R_{\xi_{i,1}}} x \in \overline{\mathbb{B}}(c_i, 2 \operatorname{diam} C_i)$, $i = 1, 2$ for any $c_i \in C_i$ and only the constants appearing in the above analysis might become larger.

CHAPTER 9

Conclusion

This thesis deals with analysis of convergence of the random function iteration (RFI) (see Algorithm 1). We analyzed the stochastic fixed point problem (see Eq. (3.1)) for the consistent and inconsistent case. Specializing to consistent feasibility, the characterizing strong type of convergence (almost sure convergence) enables one to analyze the RFI even on Hilbert spaces with a.s. weak convergence. A suitable description for a corresponding object for the inconsistent case was not applicable, so in that case an analysis of the RFI is only possible on the set of points that are contained in the support of any invariant measure.

We have shown that for averaged mappings convergence of the RFI in the weak sense is given as soon as an invariant measure exists for the corresponding Markov operator $\mathcal{P}$. That means that this is a reasonable question to ask, but yet to be done is a satisfying convergence rate analysis to get quantitative statements. This could only be done in special cases and several examples. The different examples show the generality of our approach and the different ways of modelling errors in a feasibility problem.

We showed that still a useful description of convergence of a noisy fixed point iteration is present. We hope that this description enables other researchers to formulate their problems in this setup to get a different understanding and that this work contributes to establishing a new view on inexact fixed point iterations and therefore in many branches of optimization.

Directions, in which further research is interesting, would be the following:

1. Go beyond convex setup, i.e. develop theory for set-valued Markov operators and almost averaged mappings, see [32] or [37].

2. We were able to characterize global geometric convergence in the consistent setup thoroughly, but for the inconsistent problem or general fixed point problem this is yet to be done. In particular, when the set of invariant measures is not a singleton. Or when the space is not decomposable into sets of points $x \in G$, where the limit measure of the sequence $(\delta_x \mathcal{P}^k)$ is ergodic.

3. Also an extension of weak convergence of the RFI in Hilbert spaces would be interesting to see.
4. Numerical experiments, to see if one can model several examples in our framework, like a model for the cut-off error in approximating real numbers on a computer.
5. Using the stochastic framework to get quantitative results on applying the RFI to solving linear operator equations or convex optimization problems. Maybe develop other (stochastic) algorithms for solving these problems.

APPENDIX A

Appendix

Theorem A.0.13 (Monotone Convergence Theorem). *Let (X_n) be a sequence of nonnegative real-valued random variables with $X_1 \leq X_2 \leq \ldots$ and let $X := \lim_n X_n \in [0, \infty]$ a.s. be the point-wise limit, we write in that case also $X_n \uparrow X$ a.s. as $n \to \infty$. Then X is a random variable and*

$$\int X \,\mathrm{d}\mu = \lim_n \int X_n \,\mathrm{d}\mu$$

for any measure μ on $(\Omega, \mathcal{F})$.

Theorem A.0.14 (Lebesgue's Dominated Convergence Theorem). *Let (X_n) be a sequence of real-valued random variables such that the point-wise limit $X := \lim_n X_n \in \mathbb{R}$ exists a.s. Suppose there exists a random variable $Y \geq 0$ with*

$$\int Y \,\mathrm{d}\mu < \infty \quad \textit{and} \quad |X_n| \leq Y \qquad \forall n \in \mathbb{N}.$$

Then

$$\int X \,\mathrm{d}\mu = \lim_n \int X_n \,\mathrm{d}\mu$$

for any measure μ on $(\Omega, \mathcal{F})$.

Theorem A.0.15 (Jensen's inequality). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{F}_0 \subset \mathcal{F}$ a σ-algebra. Let X be a real-valued random variable with $\mathbb{E}[X^-], \mathbb{E}[f(X)^-] < \infty$ and $f : \mathbb{R} \to \mathbb{R}$ convex, then*

$$f(\mathbb{E}[X \,|\, \mathcal{F}_0]) \leq \mathbb{E}[f(X) \,|\, \mathcal{F}_0] \qquad \text{a.s.}$$

Furthermore, if $\mathbb{E}[|X|] < \infty$, then

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

Proof. The proof in [29, Satz 8.20] is applicable using Theorem 2.2.8. □

Theorem A.0.16 (Convergence with subsequences)**.** *Let (G,d) be a metric space. Let $(x_n) \subset G$ be a sequence with the property that any subsequence has a convergent subsequence with the same limit $x \in G$. Then $x_n \to x$.*

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

Remark A.0.17: In a compact metric space, it is enough, that all clusterpoints are the same, because then every subsequence has a convergent subsequence.

Theorem A.0.18 (Regularity of measures, Proposition 2.3 and Corollary 2.5 in [21])**.** *Any finite measure μ on a Polish space is regular, that is, for every $A \in \mathcal{B}(G)$*

$$\begin{aligned} \mu(A) &= \sup\{\mu(K) \,|\, K \text{ is compact with } K \subset A\} \\ &= \sup\{\mu(B) \,|\, B \text{ is closed with } B \subset A\} \qquad \text{(inner regular)} \\ &= \inf\{\mu(U) \,|\, U \text{ is open with } A \subset U\} \qquad \text{(outer regular)} \end{aligned}$$

Lemma A.0.19 (slices of product σ-field, see Proposition 3.3.2 in [9])**.** *Let $(\Omega_i, \mathcal{F}_i)$, $i = 1,2$ be two measurable spaces and $M \in \mathcal{F}_1 \otimes \mathcal{F}_2$. Then for $\omega_1 \in \Omega_1$ holds $M_{\omega_1} := \{\omega_2 \in \Omega_2 \,|\, (\omega_1, \omega_2) \in M\} \in \mathcal{F}_2$.*

Theorem A.0.20 (dense sets in separable metric space)**.** *Let (G,d) be a Polish space (complete separable metric space). Then for any $A \subset G$, there is a dense countable subset $\{a_n\}_{n\in\mathbb{N}} \subset A$ and if A is closed then even $A = \operatorname{cl}\{a_n\}$ ($\operatorname{cl} U$ denotes the closure of the set $U \subset G$ w.r.t. the metric d).*

Proof. Since G is separable there exists a dense and countable subset $\{u_n\}_{n\in\mathbb{N}} \subset G$ with $G = \operatorname{cl}\{u_n\}_n$. By denseness of $\{u_n\} \subset G$, for any $x \in G$ and any $\epsilon > 0$, there is u_n, where n is depending on x and ϵ, with $d(u_n, x) < \epsilon$. Let $\epsilon > 0$ and choose $a_n^\epsilon \in \mathbb{B}(u_n, \epsilon) \cap A$, $n \in \mathbb{N}$, if the intersection is nonempty. The set $\tilde{A} := \{a_n^{1/m}\}_{n,m\in\mathbb{N}} \subset A$ is nonempty and countable as union of countable sets. It holds for any $a \in A$ and any $\epsilon > 0$ that $\exists n, m$ with $1/m < \epsilon$ and $d(a, u_n) < \epsilon$, hence

$$d(a, a_n^{1/m}) \leq d(a, u_n) + d(u_n, a_n^{1/m}) < 2\epsilon,$$

i.e. $\tilde{A} \subset A$ dense. So then $A \subset \operatorname{cl}\tilde{A}$ and if A is closed, then also $\operatorname{cl}\tilde{A} \subset A$. □

Lemma A.0.21 (measurability of integral of kernel)**.** *Let $(S, \mathcal{S})$, $(T, \mathcal{T})$ be measurable spaces and p a probability kernel from S to T. Let $f : S \times T \to \mathbb{R}_+$, then $s \mapsto \int f(s,t)p(s, \mathrm{d}t)$ is measurable.*

Lemma A.0.22. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be probability space. Let $(X_k)_{k\in\mathbb{N}_0}$, $(U_k)_{k\in\mathbb{N}_0}$ be sequences of nonnegative real-valued random variables with $X_k \in \mathcal{F}_k$, where $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots \subset \mathcal{F}$ are σ-algebras. Suppose for all $k \in \mathbb{N}_0$*

$$X_{k+1} \leq X_k - U_k \quad a.s.$$

Define $V_k := \mathbb{E}[U_k \,|\, \mathcal{F}_k]$ for $k \in \mathbb{N}_0$. Then $X_k \to X$ a.s. and $\sum_k U_k, \sum_k V_k < \infty$ a.s.

Proof. This is a special instance of the more general supermartingale convergence theorem in [45]. We will give a proof of this simpler result nevertheless. First define the $\mathbb{P}$-nullsets $N_k = \{X_k > X_{k-1} - U_{k-1}\}$, $k \in \mathbb{N}$ and $N = \bigcup_k N_k$. So $X_{k+1}(\omega) \leq X_k(\omega) - U_k(\omega)$ for all $k \in \mathbb{N}_0$ and all $\omega \in \Omega \setminus N$.
Since (X_k) is monotonically decreasing a.s., there exists a nonnegative random variable X with $X_k \to X$ a.s. as $k \to \infty$. So $\sum_k U_k \leq X_0 - X < \infty$ a.s. Since $0 \leq \sum_{k=1}^n U_k \uparrow \sum_k U_k$ and $0 \leq \sum_{k=1}^n V_k \uparrow \sum_k V_k$ as $n \to \infty$ we have by the Monotone Convergence Theorem for any $A \in \mathcal{F}_0$ that

$$\int_A \sum_k V_k \,\mathrm{d}\mathbb{P} = \sum_k \int_A V_k \,\mathrm{d}P = \sum_k \int_A U_k \,\mathrm{d}\mathbb{P} = \int_A \sum_k U_k \,\mathrm{d}\mathbb{P}\,.$$

We conclude from the fact that $\sum_k U_k \cdot \mathbb{P}$ determines a σ-finite measure, that also $\sum_k V_k \cdot \mathbb{P}$ does so. Then Lemma 2.2.5 implies $\sum_k V_k < \infty$ a.s. $\square$

Lemma A.0.23 (further properties of R)**.** *If $T_i = P_i$ are projectors onto nonempty, closed and convex sets, $i \in I$, then:*

1. *R is convex.*

2. *R is continuously differentiable, $\frac{1}{2}\nabla R(x) = x - \mathbb{E}[P_\xi x]$ for all $x \in \mathcal{H}$.*

3. *∇R is globally Lipschitz continuous with constant not larger than 4.*

4. *$C = \{\nabla R = 0\}$.*

Proof. 1. The function $x \mapsto \operatorname{dist}(x, C_i)$ is convex for all $i \in I$, since $C_i = \operatorname{Fix} P_i$ is convex, nonempty and closed. On $[0, \infty)$ the function $x \mapsto x^2$ is increasing and convex, so $x \mapsto \operatorname{dist}^2(x, C_i)$ is convex, $i \in I$. The convexity of R follows by linearity of the expectation.

2. We need to show that

$$\lim_{0 \neq \|y\| \to 0} \frac{|R(x+y) - R(x) - 2\mathbb{E}[\langle x - P_\xi x, y\rangle]|}{\|y\|} = 0.$$

Let $(y_n) \subset \overline{\mathbb{B}}(0, \epsilon) \subset \mathcal{H}$ with $y_n \to 0$. Define a sequence of functions on Ω via

$$f_n = \frac{\left|\operatorname{dist}^2(x + y_n, C_\xi) - \operatorname{dist}^2(x, C_\xi) - 2\left\langle x - P_\xi x, y_n\right\rangle\right|}{\|y_n\|}.$$

Then for fixed $\omega \in \Omega$ we have $f_n(\omega) \to 0$ as $n \to \infty$ since the function $x \mapsto \operatorname{dist}^2(x, C_i)$ is Fréchet differentiable for all $i \in I$ [7, Corollary 12.30]. Furthermore, we find for any $n \in \mathbb{N}$ that

$$\begin{aligned}
f_n &= \frac{\left| \|x + y_n - P_\xi(x + y_n)\|^2 - \|x - P_\xi x\|^2 - 2\left\langle x - P_\xi x, y_n \right\rangle \right|}{\|y_n\|} \\
&= \frac{\left| \|y_n - P_\xi(x + y_n) + P_\xi x\|^2 + 2\left\langle x - P_\xi x, P_\xi x - P_\xi(x + y_n) \right\rangle \right|}{\|y_n\|} \\
&= \frac{\left| \|y_n\|^2 + \|P_\xi x - P_\xi(x + y_n)\|^2 + 2\left\langle y_n + x - P_\xi x, P_\xi x - P_\xi(x + y_n) \right\rangle \right|}{\|y_n\|} \\
&\leq \frac{4\|y_n\|^2 + 2\operatorname{dist}(x, C_\xi)\|y_n\|}{\|y_n\|} \\
&\leq 4\epsilon + 2\operatorname{dist}(x, C_\xi) =: g,
\end{aligned}$$

where, in the first inequality, we used nonexpansivity of the projectors P_i, $i \in I$ and the Cauchy-Schwartz inequality. In particular with Hölder's inequality follows that $\mathbb{E}[g] \leq 4\epsilon + 2\sqrt{R(x)}$, i.e. g is integrable and hence Lebesgue's Dominated Convergence Theorem yields $\mathbb{E}[f_n] \to 0$, which gives us Fréchet differentiability of R with derivative $\nabla R(x) = 2\mathbb{E}[x - P_\xi x]$ (note that this integral exists in the sense of Bochner, since $\mathcal{H}$ is separable and $\langle P_\xi x, y\rangle$ is measurable for any $y \in \mathcal{H}$ [52]). Continuity of ∇R follows from

$$\begin{aligned}
\|\nabla R(x + y) - \nabla R(x)\| &= 2\|\mathbb{E}[y - P_\xi(x + y) + P_\xi x]\| \\
&\leq 2\mathbb{E}[\|y\| + \|P_\xi(x + y) - P_\xi x\|] \\
&\leq 4\|y\|,
\end{aligned}$$

where we used [52, Proposition 1.16] for the first inequality and nonexpansivity of the projectors P_i, $i \in I$ in the second inequality.

3. For any $x, y \in \mathcal{H}$ it holds that $\|\nabla R(x) - \nabla R(y)\| \leq 2(\|x - y\| + \|\mathbb{E}[P_\xi x - P_\xi y]\|)$. Applying [52, Proposition 1.16] and nonexpansivity, we arrive at the desired result.
4. Clearly if $x \in C$, then $x = P_\xi x$ a.s. and so $x = \mathbb{E}[P_\xi x]$, i.e. $\nabla R(x) = 0$ by 2. Now conversely, if $\nabla R(x) = 0$, then by convexity $R(x) - R(y) \leq \langle \nabla R(x), x - y\rangle = 0$ for all $y \in \mathcal{H}$. Since $C \neq \emptyset$ there is $y \in \mathcal{H}$ with $R(y) = 0$, so also $R(x) = 0$, i.e. $x \in C$.

□

APPENDIX B

PARACONTRACTIONS

Paracontractions include the set of averaged operators, but averaged mappings possess more useful regularity properties, e.g. when composing these operators, one stays in the set of averaged operators, whereas for nonaveraged operators this is not clear in general. An example of a nonaveraged paracontraction in $\mathbb{R}$ is a Huber function with parameter $\alpha > 0$ (see also [6, Example 2.3] for $\alpha = 1$)

$$f_\alpha(x) := \begin{cases} \frac{x^2}{2\alpha}, & |x| \le \alpha \\ |x| - \frac{\alpha}{2}, & |x| > \alpha \end{cases}, \qquad x \in \mathbb{R}.$$

We have that f_α is nonexpansive and paracontractive, but not averaged, since for $x = -2\alpha$ and $y = -\alpha$ one has $f(x) = \frac{3\alpha}{2}$ and $f(y) = \frac{\alpha}{2}$. Consequently

$$|f(x) - f(y)| = \alpha = |x - y|, \qquad \text{but} \quad |x - f(x) - (y - f(y))| = 2\alpha \ne 0.$$

In general metric spaces with nonlinear structure the averaged mappings are not defined or at least demand a different definition, but still the paracontraction framework applies here and exhibits a useful description of mappings which ensure that the RFI converges to a common fixed point. Paracontractions were used in Section 4.1 to guarantee Fejér monotonicity, yielding convergence; averagedness in this context would be too strong an assumption (and is not defined actually). In the following we provide an example of a class of paracontracting operators in $\mathbb{R}^n$, that are not in general averaged, resolvents of quasiconvex functions.

Definition B.0.24. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called quasiconvex, if the sublevel sets

$$\{x \in \mathbb{R}^n \,|\, f(x) \le \alpha\}$$

are convex for all $\alpha \in \mathbb{R}$. Equivalently, f satisfies

$$f(\lambda x + (1-\lambda)y) \le \max\{f(x), f(y)\} \qquad \forall x, y \in \mathbb{R}^n,\ \forall \lambda \in [0,1].$$

The proximity operator of a function $f : \mathbb{R}^n \to \mathbb{R}$ is given by the set-valued mapping

$$\operatorname{prox}_f(x) := \operatorname*{argmin}_{y\in\mathbb{R}^n} \left\{ f(y) + \frac{1}{2}\|x-y\|^2 \right\}, \qquad x \in \mathbb{R}^n.$$

Lemma B.0.25. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable and quasiconvex and satisfy $S := \operatorname{argmin} f \neq \emptyset$ and $\nabla f \neq 0$ on $\mathbb{R}^n \setminus S$, furthermore suppose that $\operatorname{Id} + \operatorname{Hess} f(x)$ is positive definit for all $x \in \mathbb{R}^n$, then prox_f is paracontracting.*

Proof. Denote $A := \operatorname{Id} + \nabla f$. Let $x, y \in \mathbb{R}^n$ with $f(x) \geq f(y)$, then

$$\|A(x) - y\|^2 = \|x-y\|^2 + \|\nabla f(x)\|^2 + \langle \nabla f(x), x-y\rangle \geq \|x-y\|^2,$$

where we used that in [2] it is shown, that a quasiconvex and differentiable function satisfies

$$f(x) \geq f(y) \qquad \Longrightarrow \qquad \langle \nabla f(x), x-y\rangle \geq 0,$$

for any $x, y \in \mathbb{R}^n$. Note that if $x \notin S$ then $\nabla f(x) \neq 0$ by assumption and hence for $y \in \mathbb{R}^n$ with $f(y) \leq f(x)$ it holds that

$$\|A(x) - y\| > \|x - y\|. \tag{B.1}$$

Moreover, the function

$$g(y) := f(y) + \frac{1}{2}\|x-y\|^2$$

for fixed $x \in \mathbb{R}^n$ is bounded from below, since $\inf_x f(x) > -\infty$ by assumption and coercive. From positive definitness of $\operatorname{Id} + \operatorname{Hess} f$ we have that g is also twice continuously differentiable and strictly convex, hence it possesses a unique minimizer $\bar{x}$ that satisfies

$$x = \nabla f(\bar{x}) + \bar{x} = A(\bar{x}),$$

it follows that $A(\mathbb{R}^n) = \mathbb{R}^n$, i.e. A is surjective. Furthermore, A is injective, since from uniqueness of the minimizer and sufficiency of the first order optimality criterion for $\bar{x}$ to be a minimizer (g is convex) it follows that, if $A(\bar{x}) = A(\bar{y})$, then $\bar{x} = \bar{y}$ is the minimizer for g and in particular $A(\bar{x}) = x \Leftrightarrow \operatorname{prox}_f(x) = \bar{x}$.

To show that also prox_f is continuous, fix $x \in \mathbb{R}^n$ and let $y \in \mathbb{B}(x, \epsilon)$. We can find a $z \in \overline{\mathbb{B}}(x, \epsilon)$ with $f(z) \leq f(y)$ for all $y \in \overline{\mathbb{B}}(x, \epsilon)$ by continuity of f, so we get with (B.1) that

$$\left\|\operatorname{prox}_f(x) - \operatorname{prox}_f(y)\right\| \leq \left\|\operatorname{prox}_f(x) - z\right\| + \left\|\operatorname{prox}_f(y) - z\right\| < \|x - z\| + \|y - z\| < 2\epsilon.$$

In particular, letting $y = \bar{x} \in S$ in (B.1), we have that

$$\left\|\operatorname{prox}_f(x) - \bar{x}\right\| < \|x - \bar{x}\| \qquad \forall x \in \mathbb{R}^n \setminus S,$$

where $S = \operatorname{argmin} f = \operatorname{Fix} \operatorname{prox}_f$. ☐

Example B.0.26 (non-averaged resolvent of quasiconvex function)**.** The function $f(x) := 1 - \exp\left(-\|x\|^2\right)$ for $x \in \mathbb{R}^n$ satisfies all the conditions in Lemma B.0.25. Its proximity operator has the derivative $\operatorname{prox}_f'(A(x)) = (A'(x))^{-1}$, where $A(x) = \left(1 + 2\exp\left(-\|x\|^2\right)\right) x$, i.e. $A'(x) = \left(1 + 2\exp\left(-\|x\|^2\right)\right) \operatorname{Id} - 4\exp\left(-\|x\|^2\right) x x^\mathsf{T}$. Since $\left\|\operatorname{prox}_f'(A(x))\right\| \geq \|y\| / \|A'(x)y\|$ for any $y \in \mathbb{R}^n \setminus \{0\}$, we have with $x = e_1 = (1, 0, \dots, 0)^\mathsf{T} = y$ that $\left\|\operatorname{prox}_f'(A(e_1))\right\| > 1$, which is in contradiction to nonexpansiveness of averaged mappings, that have derivative bounded by 1, if it exists.

Where paracontractions also occur are nonconvex feasibility problems, both consistent and inconsistent. As long as the fixed point set of the averaged projections operator consists of isolated points and the projectors are single-valued in a neighborhood of this fixed point, [37, Theorem 3.2] shows that these operators are paracontractions, whenever all assumptions of the theorem are met. Unfortunately, a statement on paracontractiveness, for the case that the fixed point set of the averaged projections operator does not consist of isolated points, is however not possible in general.

Furthermore, also nonconvex forward-backward operators appearing in structured optimization of nonconvex objective functions show the paracontractiveness property, see [37, Proposition 3.9], and these are not averaged in general, still the assumption that the fixed points are isolated is used.

Bibliography

[1] H.W. Alt, *Lineare funktionalanalysis: Eine anwendungsorientierte einführung*, Springer Lehrbuch, Springer, 2002.

[2] K. J. Arrow and Alain C. Enthoven, *Quasi-concave programming.*, Econometrica **29** (1961), 779–800 (English).

[3] 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** (1978), no. 1, 1–9.

[4] H. Bauer, *Maß- und integrationstheorie*, De Gruyter Lehrbuch, W. de Gruyter, 1992.

[5] H. H. Bauschke and J. M. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Rev. **38** (1996), no. 3, 367–426.

[6] Heinz H. Bauschke and Jonathan M. Borwein, *On projection algorithms for solving convex feasibility problems.*, SIAM Rev. **38** (1996), no. 3, 367–426 (English).

[7] Heinz H Bauschke and Patrick L Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer New York, 2011.

[8] P. Billingsley, *Convergence of Probability Measures*, 2nd ed., John Wiley and Sons, Inc, 1999.

[9] V.I. Bogachev, *Measure theory, part 1*, Measure Theory, Springer Berlin Heidelberg, 2007.

[10] ———, *Measure theory, part 2*, Measure Theory, Springer Berlin Heidelberg, 2007.

[11] Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W. Suter, *From error bounds to the complexity of first-order descent methods for convex functions*, Mathematical Programming **165** (2017), no. 2, 471–507.

[12] Leo Breiman, *The strong law of large numbers for a class of Markov chains.*, Ann. Math. Stat. **31** (1960), 801–803 (English).

[13] Felix E. Browder, *Existence and approximation of solutions of nonlinear variational inequalities*, Proceedings of the National Academy of Sciences **56** (1966), no. 4, 1080–1086.

[14] Dan Butnariu, *The expected-projection method: Its behavior and applications to linear operator equations and convex optimization*, Journal of Applied Analysis **1** (1995), no. 1, 93–108.

[15] Dan Butnariu and Sjur Didrik Flåm, *Strong convergence of expected-projection methods in hilbert spaces*, Numerical Functional Analysis and Optimization **16** (1995), no. 5-6, 601–636.

[16] Frank Deutsch, *Best approximation in inner product spaces*, Springer New York, 2001.

[17] Persi Diaconis and David Freedman, *Iterated Random Functions*, SIAM Review **41** (1999), no. 1, 45–76.

[18] Marie Duflo, *Random iterative models. Transl. from the French by Stephen S. Wilson.*, Berlin: Springer, 1997.

[19] Andreas Eberle, *Markov processes*, lecture notes, university Bonn, June 2017.

[20] M. Edelstein, *A remark on a theorem of M. A. Krasnoselski*, Amer. Math. Monthly **73** (1966), no. 5, 509–510.

[21] Onno Van Gaans, *Probability measures on metric spaces*, Lecture notes (2003), 1–29.

[22] L. Gubin, B. Polyak, and E. Raik, *The method of projections for finding the common point of convex sets*, USSR Comput. Math. and Math. Phys. **7** (1967), no. 6, 1–24.

[23] Martin Hairer, *Ergodic properties of Markov processes*, Lecture Notes in Mathematics **1881** (2006), 1–39.

[24] ———, *Convergence of Markov processes*, Lecture notes, University of Warwick (2016), 39.

[25] Neal Hermer, D. Russell Luke, and Anja Sturm, *Random Function Iterations for Consistent Stochastic Feasibility*, to appear, October 2018.

[26] Onésimo Hernández-Lerma and Jean Bernard Lasserre, *Markov chains and invariant probabilities.*, vol. 211, Basel: Birkhäuser, 2003.

[27] B. Jamison, *Asymptotic behavior of successive iterates of continuous functions under a Markov operator.*, J. Math. Anal. Appl. **9** (1964), 203–214 (English).

[28] O. Kallenberg, *Foundations of modern probability*, Probability and Its Applications, Springer New York, 1997.

[29] A. Klenke, *Wahrscheinlichkeitstheorie (german)*, Springer-Lehrbuch Masterclass, Springer Berlin Heidelberg, 2013.

[30] Ulrich Kohlenbach, Genaro López-Acedo, and Adriana Nicolae, *Moduli of regularity and rates of convergence for Fejér monotone sequences*, arXiv preprint, 2017.

[31] M. A. Krasnoselski, *Two remarks on the method of successive approximations*, Math. Nauk. (N.S.) **63** (1955), no. 1, 123–127, (Russian).

[32] Alexander Y. Kruger, D. Russell Luke, and Nguyen H. Thao, *Set regularities and feasibility problems*, Mathematical Programming **168** (2018), no. 1-2, 279–311.

[33] L. Kuipers and H. Niederreiter, *Uniform distribution of sequences*, John Wiley & Sons, 1974.

[34] T. Kurtz and S. Ethier, *Markov Processes, Characterization and Convergence*, John Wiley and Sons, Inc., 2005.

[35] Andrzej Lasota and Tomasz Szarek, *Lower bound technique in the theory of a stochastic differential equation.*, J. Differ. Equations **231** (2006), no. 2, 513–533 (English).

[36] D Russell Luke, Marc Teboulle, and Nguyen H Thao, *Necessary conditions for linear convergence of iterated expansive, set-valued mappings*, to appear, 2018.

[37] D Russell Luke, Nguyen H. Thao, and Matthew K. Tam, *Quantitative convergence analysis of iterated expansive , set-valued mappings*, to appear, 2018.

[38] W. R. Mann, *Mean value methods in iterations*, Proc. Amer. Math. Soc. **4** (1953), 506–510.

[39] P. Mattila, *Geometry of sets and measures in euclidean spaces: Fractals and rectifiability*, Cambridge Studies in Advanced Mathematics, Cambridge University Press, 1995.

[40] S P Meyn and R L Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag (1993), 792.

[41] Angelia Nedić, *Random projection algorithms for convex set intersection problems*, Proceedings of the IEEE Conference on Decision and Control (2010), 7655–7660.

[42] ______, *Random algorithms for convex minimization problems*, Mathematical Programming **129** (2011), no. 2, 225–253.

[43] K. R. Parthasarathy, *Probability Measures On Metric Spaces*, Academic Press, New York and London, 1967.

[44] Severine Rigot, *Differentiation of measures in metric spaces*, arXiv (2018), 1–22.

[45] H. Robbins and D. Siegmund, *A convergence theorem for non negative almost supermartingales and some applications.*, Optimizing Meth. Statist., Proc. Sympos. Ohio State Univ. 1971, 233-257 (1971)., 1971.

[46] Gareth O. Roberts and Jeffrey S. Rosenthal, *General state space Markov chains and MCMC algorithms.*, Probab. Surv. **1** (2004), 20–71 (English).

[47] Daniel Rudolf, *Wasserstein distance and Markov chains, Lecture notes, University of Goettingen*, 2017.

[48] Örjan Stenflo, *A survey of average contractive iterated function systems.*, Journal of Difference Equations and Applications **18** (2012), no. 8, 1355–1380.

[49] V. Strassen, *The Existence of Probability Measures with Given Marginals*, The Annals of Mathematical Statistics **36** (1965), no. 2, 423–439.

[50] D.W. Stroock, *Probability theory: An analytic view*, Cambridge University Press, 2010.

[51] Tomasz Szarek, *Feller processes on nonlocally compact spaces*, Annals of Probability **34** (2006), no. 5, 1849–1863.

[52] Jan van Neerven, *Stochastic evolution equations*, lecture notes (TU Delft), 2018.

[53] Cedric Villani, *Optimal transport, old and new*, 2008.

[54] Daniël Worm, *Semigroups on spaces of measures*, Ph.D. thesis, Universiteit Leiden, 2010.

[55] Radu Zaharopol, *Invariant probabilities of Markov-Feller operators and their supports.*, Basel: Birkhäuser, 2005.