The map T_* is continuous and affine, so the set $\mathscr{M}^T(X)$ of T-invariant measures is a closed convex subset of $\mathscr{M}(X)$; in the next section[(48)] we will see that it is always non-empty.

4.1 Existence of Invariant Measures

The connection between ergodic theory and the dynamics of continuous maps on compact metric spaces begins with the next result, which shows that invariant measures can always be found.

Theorem 4.1. *Let $T : X \to X$ be a continuous map of a compact metric space, and let (ν_n) be any sequence in $\mathscr{M}(X)$. Then any weak*-limit point of the sequence (μ_n) defined by $\mu_n = \frac{1}{n}\sum_{j=0}^{n-1} T_*^j \nu_n$ is a member of $\mathscr{M}^T(X)$.*

An immediate consequence is the following important general statement, which shows that measure-preserving transformations are ubiquitous. It is known as the Kryloff–Bogoliouboff Theorem [214].

Corollary 4.2 (Kryloff–Bogoliouboff). *Under the hypotheses of Theorem* 4.1, *$\mathscr{M}^T(X)$ is non-empty.*

Proof. Since $\mathscr{M}(X)$ is weak*-compact, the sequence (μ_n) must have a limit point. □

Write $\|f\|_\infty = \sup\{|f(x)| \mid x \in X\}$ as usual.

Proof of Theorem 4.1. Let $\mu_{n(j)} \to \mu$ be a convergent subsequence of (μ_n) and let $f \in C(X)$. Then, by applying the definition of $T_*\mu_n$, we get

$$\begin{aligned}\left|\int f\circ T \,\mathrm{d}\mu_{n(j)} - \int f \,\mathrm{d}\mu_{n(j)}\right| &= \frac{1}{n(j)}\left|\int \sum_{i=0}^{n(j)-1}\left(f\circ T^{i+1} - f\circ T^i\right)\mathrm{d}\nu_{n(j)}\right| \\ &= \frac{1}{n(j)}\left|\int \left(f\circ T^{n(j)+1} - f\right)\mathrm{d}\nu_{n(j)}\right| \\ &\leqslant \frac{2}{n(j)}\|f\|_\infty \longrightarrow 0\end{aligned}$$

as $j \to \infty$, for all $f \in C(X)$. It follows that $\int f\circ T\,\mathrm{d}\mu = \int f\,\mathrm{d}\mu$, so μ is a member of $\mathscr{M}^T(X)$ by Lemma B.12. □

Thus $\mathscr{M}^T(X)$ is a non-empty compact convex set, since convex combinations of elements of $\mathscr{M}^T(X)$ belong to $\mathscr{M}^T(X)$. It follows that $\mathscr{M}^T(X)$ is an infinite set unless it comprises a single element. For many maps it is difficult to describe the space of invariant measures. The next example has very few ergodic invariant measures, and we shall see later many maps that have only one invariant measure.

Example 4.3 (North–South map). Define the stereographic projection π from the circle $X = \{z \in \mathbb{C} \mid |z - \mathrm{i}| = 1\}$ to the real axis by continuing the line from 2i through a unique point on $X \smallsetminus \{2\mathrm{i}\}$ until it meets the line $\Im(z) = 0$ (see Figure 4.1).

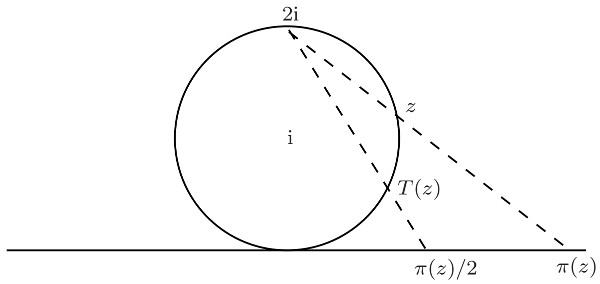

Fig. 4.1: The North-South map on the circle; for $z \neq 2\mathrm{i}$, $T^n z \to 0$ as $n \to \infty$.

The "North–South" map $T : X \to X$ is defined by

$$T(z) = \begin{cases} 2\mathrm{i} & \text{if } z = 2\mathrm{i}; \\ \pi^{-1}(\pi(z)/2) & \text{if } z \neq 2\mathrm{i} \end{cases}$$

as shown in Figure 4.1. Using Poincaré recurrence (Theorem 2.11) it is easy to show that $\mathscr{M}^T(X)$ comprises the measures $p\delta_{2\mathrm{i}} + (1-p)\delta_0$, $p \in [0,1]$ that are supported on the two points 2i and 0. Only the measures corresponding to $p = 0$ and $p = 1$ are ergodic.

It is in general difficult to identify measures with specific properties, but the ergodic measures are readily characterized in terms of the geometry of the space of invariant measures.

Theorem 4.4. *Let X be a compact metric space and let $T : X \to X$ be a measurable map. The ergodic elements of $\mathscr{M}^T(X)$ are exactly the extreme points of $\mathscr{M}^T(X)$.*

That is, T is ergodic with respect to an invariant probability measure if and only if that measure cannot be expressed as a strict convex combination of two different T-invariant probability measures. For any measurable set A, define $\mu\big|_A$ by $\mu\big|_A(C) = \mu(A \cap C)$. If T is not assumed to be continuous, then we do not know that $\mathscr{M}^T(X) \neq \varnothing$, so without the assumption of continuity Theorem 4.4 may be true but vacuous (see Exercise 4.1.1).

Proof of Theorem 4.4. Let $\mu \in \mathscr{M}^T(X)$ be a non-ergodic measure. Then there is a measurable set B with $\mu(B) \in (0,1)$ and with $T^{-1}B = B$. It follows that

$$\frac{1}{\mu(B)}\mu\big|_B, \frac{1}{\mu(X\smallsetminus B)}\mu\big|_{X\smallsetminus B} \in \mathscr{M}^T(X),$$

so

$$\mu = \mu(B)\left(\frac{1}{\mu(B)}\mu\big|_B\right) + \mu(X\smallsetminus B)\left(\frac{1}{\mu(X\smallsetminus B)}\mu\big|_{X\smallsetminus B}\right)$$

expresses μ as a strict convex combination of the invariant probability measures

$$\frac{1}{\mu(B)}\mu\big|_B$$

and

$$\frac{1}{\mu(X\smallsetminus B)}\mu\big|_{X\smallsetminus B},$$

which are different since they give different measures to the set B.

Conversely, let μ be an ergodic measure and assume that

$$\mu = s\nu_1 + (1-s)\nu_2$$

expresses μ as a strict convex combination of the invariant measures ν_1 and ν_2. Since $s > 0$, $\nu_1 \ll \mu$, so there is a positive function $f \in L^1_\mu$ (f is the Radon–Nikodym derivative $\frac{d\nu_1}{d\mu}$; see Theorem A.15) with the property that

$$\nu_1(A) = \int_A f \,\mathrm{d}\mu \tag{4.1}$$

for any measurable set A. The set $B = \{x \in X \mid f(x) < 1\}$ is measurable since f is measurable, and

$$\begin{aligned}\int_{B\cap T^{-1}B} f\,\mathrm{d}\mu + \int_{B\smallsetminus T^{-1}B} f\,\mathrm{d}\mu &= \nu_1(B)\\ &= \nu_1(T^{-1}B)\\ &= \int_{B\cap T^{-1}B} f\,\mathrm{d}\mu + \int_{(T^{-1}B)\smallsetminus B} f\,\mathrm{d}\mu,\end{aligned}$$

so

$$\int_{B\smallsetminus T^{-1}B} f\,\mathrm{d}\mu = \int_{(T^{-1}B)\smallsetminus B} f\,\mathrm{d}\mu. \tag{4.2}$$

By definition, $f(x) < 1$ for $x \in B\smallsetminus(T^{-1}B)$ while $f(x) \geqslant 1$ for $x \in T^{-1}B\smallsetminus B$. On the other hand,

$$\begin{aligned}\mu((T^{-1}B)\smallsetminus B) &= \mu(T^{-1}B) - \mu((T^{-1}B)\cap B)\\ &= \mu(B) - \mu((T^{-1}B)\cap B)\\ &= \mu(B\smallsetminus T^{-1}B)\end{aligned}$$

so equation (4.2) implies that $\mu(B\setminus T^{-1}B) = 0$ and $\mu((T^{-1}B)\setminus B) = 0$. Therefore $\mu((T^{-1}B)\triangle B) = 0$, so by ergodicity of μ we must have $\mu(B) = 0$ or 1. If $\mu(B) = 1$ then

$$\nu_1(X) = \int_X f \,\mathrm{d}\mu < \mu(B) = 1,$$

which is impossible. So $\mu(B) = 0$.

A similar argument shows that $\mu(\{x \in X \mid f(x) > 1\}) = 0$, so $f(x) = 1$ almost everywhere with respect to μ. By equation (4.1), this shows that

$$\nu_1 = \mu,$$

so μ is an extreme point in $\mathscr{M}^T(X)$. □

Write $\mathscr{E}^T(X)$ for the set of extreme points in $\mathscr{M}^T(X)$ – by Theorem 4.4, this is the set of ergodic measures for T.

Example 4.5. Let $X = \{1, \dots, r\}^{\mathbb{Z}}$ and let $T : X \to X$ be the left shift map. In Example 2.9 we defined for any probability vector $\mathbf{p} = (p_1, \dots, p_r)$ a T-invariant probability measure $\mu = \mu_{\mathbf{p}}$ on X, and by Proposition 2.15 all these measures are ergodic. Thus for this example the space $\mathscr{E}^T(X)$ of ergodic invariant measures is uncountable. This collection of measures is an inconceivably tiny subset of the set of all ergodic measures – there is no hope of describing all of them.

Measures μ_1 and μ_2 are called *mutually singular* if there exist disjoint measurable sets A and B with $A \cup B = X$ for which $\mu_1(B) = \mu_2(A) = 0$ (see Section A.4).

Lemma 4.6. *If $\mu_1, \mu_2 \in \mathscr{E}^T(X)$ and $\mu_1 \neq \mu_2$ then μ_1 and μ_2 are mutually singular.*

Proof. Let $f \in C(X)$ be chosen with $\int f \,\mathrm{d}\mu_1 \neq \int f \,\mathrm{d}\mu_2$ (such a function exists by Theorem B.11). Then by the ergodic theorem (Theorem 2.30)

$$\mathsf{A}_n^f(x) \to \int f \,\mathrm{d}\mu_1 \tag{4.3}$$

for μ_1-almost every $x \in X$, and

$$\mathsf{A}_n^f(x) \to \int f \,\mathrm{d}\mu_2$$

for μ_2-almost every $x \in X$. It follows that the set $A = \{x \in X \mid (4.3) \text{ holds}\}$ is measurable and has $\mu_1(A) = 1$ but $\mu_2(A) = 0$. □

Some of the problems for this section make use of the topological analog of Definition 2.7, which will be used later.