1 Introductin

1.1 Motivation. Types of dynamical systems

Dynamical systems are used to build economic, medical, chemical, biological and physical models.

There are two main types of dynamical systems: continuous dynamical systems and discrete dynamical systems.

通常面对一些无解析解的DE.

Example of continuous dynamical system:

$$\dot{x} = F(x).$$

Example of discrete dynamical system:

$$x_{n+1} = f(x_n).$$

In our course we will focus on discrete dynamical systems. To be more precise, on one-dimensional discrete dynamical systems.

1.2 Examples of dynamical systems

1.2.1 An example from Finance

Consider the following situation. Suppose we deposit $1000 in a bank at 10 percent interest. We ask the question: if we leave this money untouched for n years, how much money will we have in our account at the end of this period? For simplicity, we assume that the 10 percent interest is added to our account once each year at the end of the year.

This is one of the simplest examples of an iterative process or dynamical system. Let's denote the amount we have in the bank at the end of the nth year by A_n. Our problem is to determine A_n for some given number of years n. We know that A_0, our initial deposit, is \$1000. After 1 year we add 10 percent to this amount to get our new balance. That is,

$$A_1 = A_0 + 0.1A_0 = 1.1A_0$$

In our specific case, $A_1 = \$1100$. At the end of the second year, we perform the same operation

$$A_2 = A_1 + 0.1A_1 = 1.1A_1,$$

so that $A_2 = \$1210$. Continuing,

$$\begin{aligned} A_3 &= 1.1A_2 \\ A_4 &= 1.1A_3 \\ &\vdots \\ A_n &= 1.1A_{n-1} \end{aligned}$$

Thus we can recursively determine the amount A_n once we know the previous year's balance.

The equation $A_n = 1.1A_{n-1}$ is an example of a (first-order) difference equation. In such an equation, we use information from the previous year (or other fixed time interval) to determine the current information, then we use this information to determine next year's amount, and so forth.

We solve this difference equation by the process of iteration. The iterative process involved is multiplication by 1.1. That is, if we define the function $F(x) = 1.1x$, then our savings balances are determined by repeatedly applying this function:

$$\begin{aligned} A_1 &= F\left(A_0\right) \\ A_2 &= F\left(A_1\right) \\ A_3 &= F\left(A_2\right) \end{aligned}$$

and so forth. Note that we may also write

$$\begin{aligned} A_2 &= F\left(F\left(A_0\right)\right) = F \circ F\left(A_0\right) \\ A_3 &= F\left(F\left(F\left(A_0\right)\right)\right) = F \circ F \circ F\left(A_0\right) \end{aligned}$$

to clearly indicate that we compose F with itself repeatedly to obtain the successive balances.

Since $F(x) = 1.1x$, we have

$$\begin{aligned} F(F(x)) &= (1.1)^2 x \\ F(F(F(x))) &= (1.1)^3 x \end{aligned}$$

and, in general, the nth iteration of the function yields

$$\underbrace{F \circ \cdots \circ F}_{n \text{ times}}(x) = (1.1)^n x$$

So to find A_n, we merely compute $(1.1)\ ^n$ and multiply by A_0. For example, using a calculator or computer, you may easily check that $A_{10} = \$2593.74$ and $A_{50} = \$117,390.85$.

This example is quite simple: the iterations we will encounter will in general yield much more complicated results.

1.2.2 An example from Ecology

Here is another difference equation which is essentially the same as our savings account example. Suppose we wish to predict the behavior of the population of a certain species which grows or declines as generations pass. Let's denote the population alive at generation n by P_n. So our question is: can we predict what will happen to P_n as n gets large? Will P_n tend to zero so that the species becomes extinct? Or will P_n grow without bound so that we experience a population explosion?

There are many mathematical models to predict the behavior of populations. By far the simplest (and most naive) is the exponential growth model. In this model we assume that the population in the succeeding generation is directly proportional to the population in the current generation. This translates to mathematics as another difference equation

$$P_{n+1} = rP_n$$

where r is some constant determined by ecological conditions. Thus, given the initial population P_0, we can recursively determine the population in the succeeding gencrations:

$$\begin{aligned} P_1 &= rP_0 \\ P_2 &= rP_1 = r^2P_0 \\ P_3 &= rP_2 = r^3P_0 \\ &\vdots \\ P_n &= rP_{n-1} = r^nP_0 \end{aligned}$$

As before, we determine the bchavior of the population via iteration. In this case, the function we iterate is $F(x) = rx$. So

$$P_n = \underbrace{F \circ \cdots \circ F}_{n \text{ times}} (P_0) = r^n P_0$$

Note that the ultimate fate of the population depends on r. If $r > 1$, then r^n tends to infinity with n, so we have unchecked population growth. If $r < 1, r^n$ tends to zero, so the species becomes extinct. Finally, if $r = 1$, $P_n = P_0$, so there is never any change in the population. Thus, we can achieve our goal of determining the fate of the species for any r. Of course, this simplified model is highly unrealistic in that real-life populations behave in a much more complicated fashion. For example, populations can never tend to infinity. To remedy this, we will add one assumption to our model that will take into account the possibility of overcrowding.

Specifically, we will discuss what ecologists call the logistic model of population growth. In this model, we assume at the outset that there is some absolute maximum population that can be supported by the environment. If the population ever reaches this number, then we have disastrous, overcrowding food supply becomes critically short, and the species immediately dies out. To keep the numbers manageable, let's assume that P_n now represents the fraction of this maximal population alive at generation n, so that $0 \leq P_n \leq 1$. The logistic model is then

$$P_{n+1} = \lambda P_n (1 - P_n)$$

加入环境承载力指标的
logistic. 模型.

As before, λ is a constant that depends on ecological conditions. For reasons that will become apparent later, we will always assume that $0 < \lambda \leq 4$.

Note that, in the absence of the $1 - P_n$ factor, we are left with the previous exponential growth model. If $P_n = 0$ (no individuals present), then $P_{n+1} = 0$ as well, as we would expect. If $P_n = 1$, then $P_{n+1} = 0$ as we have assumed.

Thus, to understand the growth and decline of the population under this model, we must iterate the logistic function $F_\lambda(x) = \lambda x(1-x)$. Unlike the previous examples, this function is quadratic rather than linear. We will see that this simple change gives rise to a very rich mathematical theory. Indeed, the behavior of this function under iteration is still far from being completely understood. We will spend most of this course analyzing the dynamical behavior of this and other similar functions.

1.2.3 Finding Roots and Solving Equations

How do you find $\sqrt{5}$ exactly? Believe it or not, the simplest method dates back to the time of the Babylonians 巴比伦人. and involves a simple iteration. We will make an initial guess x_0 for $\sqrt{5}$. We assume that x_0 is positive. Now, assuming that $x_0 \neq \sqrt{5}$, so we will use this guess to produce a new and better guess x_1.

Here is the procedure. If $x_0 \neq \sqrt{5}$, then we either have $x_0 < \sqrt{5}$ or $x_0 > \sqrt{5}$. In the former case, we have

$$\sqrt{5}x_0 < 5$$
$$\sqrt{5} < \frac{5}{x_0}$$

for $x_0 \neq 0$. On the other hand, if $x_0 > \sqrt{5}$, then

$$\sqrt{5} > \frac{5}{x_0}$$

Thus we have either

$$x_0 < \sqrt{5} < \frac{5}{x_0}$$

or

$$\frac{5}{x_0} < \sqrt{5} < x_0.$$

Therefore, if we take the average of x_0 and $5/x_0$, namely

$$x_1 = \frac{1}{2}\left(x_0 + \frac{5}{x_0}\right)$$

the resulting value will lie midway between x_0 and $5/x_0$ and so will, hopefully, be a better approximation to $\sqrt{5}$. So we use this value as our next "guess" for $\sqrt{5}$. Continuing, we form the successive averages

$$\begin{aligned} x_2 &= \frac{1}{2}\left(x_1 + \frac{5}{x_1}\right) \\ x_3 &= \frac{1}{2}\left(x_2 + \frac{5}{x_2}\right) \\ &\vdots \end{aligned}$$

Intuitively, the sequence of numbers $x_0, x_1, x_2, \ldots$ should eventually approach $\sqrt{5}$.

Let's see how this works in practice. Suppose we make the (somewhat silly) initial guess $x_0 = 1$ for $\sqrt{5}$. Then we have

$$\begin{aligned}
x_1 &= \frac{1}{2}(1+5) = 3 \\
x_2 &= \frac{1}{2}\left(3+\frac{5}{3}\right) = \frac{7}{3} = 2.333\ldots \\
x_3 &= \frac{1}{2}\left(\frac{7}{3}+\frac{15}{7}\right) = \frac{47}{21} = 2.238095\ldots \\
x_4 &= 2.236068\ldots \\
x_5 &= 2.236067\ldots \\
x_6 &= 2.236067\ldots
\end{aligned}$$

and we see that, very quickly, this sequence tends to the correct answer, as $\sqrt{5} = 2.236067\ldots$.

Clearly, to find other square roots, we need only replace the 5 in the formula for our initial guess. Thus, to "find" $\sqrt{9}$, we choose some initial guess, say $x_0 = 2$, and then compute

$$\begin{aligned}
x_1 &= \frac{1}{2}\left(2+\frac{9}{2}\right) = 3.25 \\
x_2 &= \frac{1}{2}\left(x_1+\frac{9}{x_1}\right) = 3.0096\ldots \\
x_3 &= 3.000015\ldots \\
x_4 &= 3.000000\ldots \\
x_5 &= 3.000000\ldots
\end{aligned}$$

and we see that this sequence quickly converges to $\sqrt{9} = 3$.

A related question that arises in all branches of science and mathematics is, how do you solve the equation $F(x) = 0$? For example, finding the square root of 5 is the same as solving the equation $x^2 - 5 = 0$. There are very few functions for which it is possible to write down the solutions to this equation explicitly. Even for polynomials, solving this equation is difficult when the degree is greater than 2, and generally impossible when the degree is 5 or greater. Yet the problem is extremely important, so we seek other possible methods.

One method, familiar from calculus, is Newton's method. This method involves the following procedure. Given a function F whose roots we are trying to find, we construct a new function, the Newton iteration function, given by

$$N(x) = x - \frac{F(x)}{F'(x)}.$$

Then we make an initial guess x_0 for a root. Newton's method is to iterate the function N, successively computing

$$\begin{aligned} x_1 &= N(x_0) \\ x_2 &= N(x_1) = N(N(x_0)) \\ x_3 &= N(x_2) = N(N(N(x_0))) \end{aligned}$$

and so forth. Often, though by no means always, this sequence of iterates converges to a root of F. Note that, in the special case where $F(x) = x^2 - 5$ (whose roots are $\pm\sqrt{5}$), the Newton iteration is simply

$$N(x) = x - \frac{x^2 - 5}{2x} = \frac{1}{2}\left(x + \frac{5}{x}\right)$$

as was discussed above. We will not take the time now to discuss why this happens or even where the Newton iteration function comes from. Rather, we will devote all of corresponding paragraph to this subject.

With the advent of accessible high-speed computation, iterative algorithms such as Newton's method are becoming an important and widespread area of interest in mathematics. No longer are such algorithms primarily of theoretical interest!

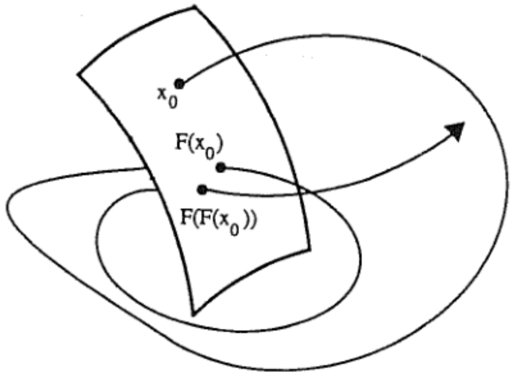
1.2.4 Differential Equations

Differential equations are also examples of dynamical systems. Unlike iterative processes where time is measured in discrete intervals such as years or generations, differential equations are examples of continuous dynamical systems wherein time is a continuous variable. Ever since the time of Newton, these types of systems have been of paramount importance.

Recently, however, discrete dynamical systems have also received considerable attention. This does not mean that continuous systems have declined in importance. Rather, mathematicians study discrete systems with an eye toward applying their results to the more difficult continuous case.

There are a number of ways that iterative processes enter the arena of differential equations. For example, a solution of a differential equation is a continuous function of time. If we look at this solution at discrete time intervals, say at times $t = 0, 1, 2, \ldots$, then we are really considering an iterative process as described above.

Most often, differential equations are impossible to solve explicitly. We must turn to the computer to generate numerical solutions. The numerical methods used to solve these equations are often iterative processes such as the Runge-Kutta method.



no explicit sol.

⇒ sol. on given surface.

A final way in which iteration arises in the study of differential equations occurs when a surface of section can be found. Suppose that we have a first-order differential equation in three dimensions whose independent variable is time. Then the solutions we seek are curves in space parameterized by time. Often, though not always, these curves intersect a given surface in space over and over again as depicted in Figure above. When this occurs, the study of solutions of the equation reduces to the study of an iterative process on the surface. Starting with any point on the surface x_0, we follow the solution through x_0 until it first reintersects the surface. Call this point of first return $F(x_0)$. The function F is called a first return map and the surface is a surface of section. Continuing to follow the solution curve, its next point of intersection is $F(F(x_0))$, so we see that determining the successive points of intersection is really a problem of iterating F.

Iteration of the first return map does not tell us all there is to know about the solutions of the differential equation, but it does give us a lot of qualitative information. In particular, we do not know the exact location of a solution at each moment of time, but we do gain long-term information about the behavior of the solution.

1.3 Preliminaries from calculus

In this section, we recall some elementary (and not-so-elementary) noions from single variable and multivariable calculus. In the sequel, we will also need a few notions from point-set topology, so we include them here as well. First, we fix some notation. $\mathbf{R}$ denotes the real numbers. I or J will always denote closed intervals in $\mathbf{R}$, i.e., all points x satisfying $a \leq x \leq b$ for some a and b . $\mathbf{R}^2$ denotes the Cartesian plane.

Let $f : \mathbf{R} \to \mathbf{R}$ be a function. We denote the derivative of f at x by $f'(x)$, the second derivative by $f''(x)$, and higher derivatives by $f^{(r)}(x)$. We say that f is of class C^r on I if $f^{(r)}(x)$ exists and is continuous at all $x \in I$. A function is said to be smooth if it is of class C^1. The function $f(x)$ is C^∞ if all derivatives exist and are continuous. Throughout this course, function means C^∞ function; occasionally we will use functions which are continuous but non-differentiable as examples, but in general, when we say function, we mean C^∞ function.

There are other classes of functions which are commonly studied in calculus. For example, analytic functions (i.e., those with convergent power series representations) are often encountered. For our purposes in this and next paragraph, these types of functions are too rigid in the following sense. We wish to allow small changes in our perturbations of the functions which will change the function in a certain interval but not everywhere. This is accomplished by the use of bump functions which we will introduce in the exercises. These small changes are impossible if we are restricted to analytic functions, for small change in any of the coefficients of the power series affects the behaivior of the function everywhere. Later, in chapter three, when we discuss complex analytic dynamical systems, we will restrict our attention solely to these types of functions.

There are some special classes of functions that often arise. The function $f(x)$ is linear if $f(x) = ax$ for some constant a; $f(x)$ is affine if $f(x) = ax + b$; $f(x)$ is piecewise linear if $f(x)$ is affine on a collection of intervals.

↙ injective

Definition 1. $f(x)$ is one-to-one if $f(x) \neq f(y)$ whenever $x \neq y$.

Clearly, increasing or decreasing functions are the only types of continuous one-to-one functions of a real variable. If $f : I \to J$ is one-to-one, then we may define the inverse of f, written $f^{-1}(x)$, by the rule $f^{-1}(x) = y$ if and only if $f(y) = x$. For example, if $f(x) = x^3$, then $f^{-1}(x) = \sqrt[3]{x}$ and if $g(x) = \tan x$, then $g^{-1}(x) = \arctan x$. Here $g : (-\pi/2, \pi/2) \to \mathbf{R}$ so $g^{-1} : \mathbf{R} \to (-\pi/2, \pi/2)$.

Definition 2. Let I and J be intervals and $f : I \to J$. The function f is onto if for any y in J there is an $x \in I$ such that $f(x) = y$.

↙ surjective.

Definition 3. Let $f : I \to J$. The function $f(x)$ is a homeomorphism if $f(x)$ is one-to-one, onto, and continuous, and $f^{-1}(x)$ is also continuous.

For example, $\tan x$ is a homeomorphism between $(-\pi/2, \pi/2)$ and $\mathbf{R}$. Thus we say the open interval $(-\pi/2, \pi/2)$ is homeomorphic to $\mathbf{R}$. Functions which are one-to-one are also said to be injective, while functions which are onto are also called surjective.

Definition 4. Let $f : I \to J$. The function $f(x)$ is a C^r-diffeomorphism if $f(x)$ is a C^r-homeomorphism such that $f^{-1}(x)$ is also C^r.

For example, it is easy to see that $\tan x$ is a C^∞ diffeomorphism from $(-\pi/2, \pi/2)$ to $\mathbf{R}$, whereas $f(x) = x^3$ is a homeomorphism which is not a diffeomorphism since $f^{-1}(x) = x^{1/3}$ and $\left(f^{-1}\right)'(0)$ does not exist.

We will see in subsequent chapters that diffeomorphisms on the real line are extremely simple, dynamically speaking. Therefore, in this and next paragraph, we will primarily consider non-invertible functions.

切值定理、 中值定理、 隐函数定理.

An important notion from elementary calculus is the Mean Value Theorem:

Theorem 1. Suppose $f : [a, b] \to \mathbf{R}$ is C^1. Then there exists $c \in [a, b]$ such that

$$f(b) - f(a) = f'(c)(b - a)$$

Another important result from calculus is the Intermediate Value Theorem:

Theorem 2. Suppose $f : [a, b] \to \mathbf{R}$ is continuous. Suppose that $f(a) = u$ and $f(b) = v$. Then for any z between u and v, there exists $c, a \leq c \leq b$, such that $f(c) = z$.

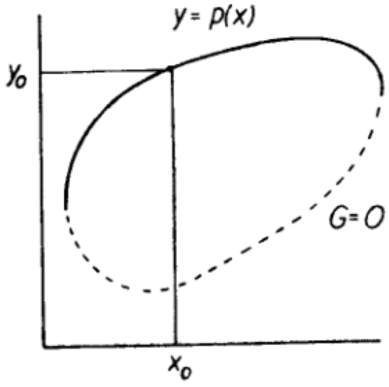
One of the most abstract and seemingly useless theorems from multivariable calculus is the Implicit Function Theorem. Most beginning students have no appreciation of the power of this Theorem when they encounter it in their first analysis course. We hope that the geometric results in bifurcation 分叉理论、 theory that we will encounter later will help dispel any misconceptions about the usefulness of this theorem.

Theorem 3. Suppose $G : \mathbf{R}^2 \to \mathbf{R}^1$ is a C^1-function (i.e., both partial derivatives of G exist and are continuous.) Suppose further that

1. $G(x_0, y_0) = 0$
2. $\frac{\partial G}{\partial y}(x_0, y_0) \neq 0$.

Then there exist open intervals I about x_0 and J about y_0 and a C^1-function $p : I \to J$ satisfying

1. $p(x_0) = y_0$
2. $G(x, p(x)) = 0$ for all $x \in I$.



Example 1. Let $G(x, y) = x^2 + y^2 - 1$. The level sets of G are clearly circles, and $G = 0$ defines the unit circle in the plane.

Suppose $G(x_0, y_0) = 0$ and $y_0 > 0$, i.e., (x_0, y_0) is a point on the upper or lower semicircle. Clearly,

$$\frac{\partial G}{\partial y}(x_0, y_0) = 2y_0 \neq 0$$

So the Implicit Function Theorem applies. The result is a function $p(x)$ which , satisfies $G(x, p(x)) = 0$ for all x sufficiently close to x_0. What is $p(x)$? In this case, we can construct $p(x)$ explicitly. Clearly, $p(x) = \sqrt{1 - x^2}$, which is C^∞ as long as $x \neq \pm 1$ (when $y = 0$). We have $G\left(x, \sqrt{1 - x^2}\right) = 0$ for $|x| < 1$, as the Implicit Function Theorem guarantees. If $y_0 < 0$, then we must choose $p(x) = -\sqrt{1 - x^2}$.

It is important to realize that, in practice, one cannot very often solve for the function $p(x)$ as we did here. Nevertheless, the Implicit Function Theorem guarantees its existence (whether or not we can explicitly write it down), and that is often exactly what we need.

Example 2. $G(x,y) = x^5y^4 - xy^5 - yx^2 + 1$ satisfies $G(1,1) = 0$ and

$$\frac{\dot{\partial} G}{\partial y}(1,1) = -2.$$

Hence there is a function $p(x)$ defined in some interval about $x = 1$ and which satisfies $G(x, p(x)) = 0$. Solving $G(x,y) = 0$ for $y = p(x)$ is impossible, however.

Fixed points for functions are points x which satisfy $f(x) = x$. These points will play a dominant role in the theory of dynamical systems. The following easy application of the Intermediate Value Theorem gives an important criterion for the existence of a fixed point.

Proposition 1. Let $I = [a, b]$ be an interval and let $f : I \to I$ be continuous. Then f has at least one fixed point in I.

Proof. Home task.

This theorem is a special case of a much more general theorem called Brouwer Fixed Point Theorem, which gives a similar sufficient condition for the existence of fixed points in higher dimensions. The following result is a special case of the Contraction Mapping Theorem.

Proposition 2. Let $f : I \to I$ and assume that $|f'(x)| < 1$ for all x in I. Then there exists a unique fixed point for f in I. Moreover

$$|f(x) - f(y)| < |x - y|$$

for all $x, y \in I, x \neq y$.

Proof. Home task.

We close this section with a few notions from general topology.

Deflnition 5. Let $S \subset \mathbf{R}$. A point $x \in \mathbf{R}$ is a limit point of S if there is a sequence of points $x_n \in S$ converging to $x.S$ is a closed set if it contains all of its limit points.

Clearly, closed intervals of the form $a \leq x \leq b$ are closed sets. Any finite union of closed sets is also closed. Infinite unions of closed sets, however, need not be closed, as the following example shows.

Example 3. Let $I_n = \left\{\frac{1}{n}, 1\right]$. Then

$$\bigcup_{n=1}^{\infty} I_n = (0, 1]$$

which is not closed, since 0 is a limit point of S which is not in S. Intersections of closed sets yield closed sets, however (the empty set is, by definition, a closed set.) Moreover, if I_n is a closed, non-empty, and bounded interval for each n and $I_{n+1} \subset I_n$, then $\cap_{n=1}^{\infty} I_n$ is a closed, non-empty set. The crucial word here is, of course, non-empty.

Definition 6. Let $S \subset \mathbf{R}$. S is an open set if, for any $x \in S$, there is an $\epsilon > 0$ such that all points t in the open interval $x - \epsilon < t < x + \epsilon$ are contained in S.

It is clear that the complement of a closed set is open and vice versa. Unlike closed sets, infinite unions of open intervals are open sets in R. However, infinite intersections of open intervals are not open sets. For example, if $J_n = \left(-\frac{1}{n}, \frac{1}{n}\right)$, then $\cap_{n=1}^{\infty} J_n = \{0\}$ which is closed.

For any set S, we denote the closure of S by $\bar{S}$. $\bar{S}$ consists of all points in S together with all limit points of S. For example, if S is the open interval $(0,1)$, then $\bar{S}$ is the closed interval $[0,1]$. Clearly, if S is closed, then $\bar{S}=S$.

For any set S, we denote the closure of S by $\bar{S}.\bar{S}$ consists of all points in S together with all limit points of S. For example, if S is the open interval $(0,1)$, then $\bar{S}$ is the closed interval $[0,1]$. Clearly, if S is closed, then $\bar{S}=S$.

Definition 7. A subset U of S is dense in S if $\bar{U}=S$. For example, any open set S is dense in its closure $\bar{S}$. A more interesting example is the set of rational numbers Q, which is dense in $\mathbf{R}$. Similarly, the irrationals are dense in $\mathbf{R}$. We caution the reader against thinking that dense subsets are necessarily large. Even open and dense sets may be quite small in the sense of total length. Here is an example in the unit interval I given by $0 \leq x \leq 1$. Since the rationals form a countable set in I, we may list thern in some order. One such ordering is

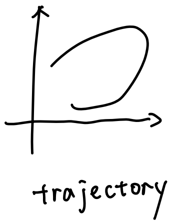
$$0,1,\frac{1}{2},\frac{1}{3},\frac{2}{3},\frac{1}{4},\frac{3}{4},\frac{1}{5},\frac{2}{5},\frac{3}{5},\frac{4}{5},\frac{1}{6},\ldots$$

Now let $\epsilon>0$ be small. Consider the open interval of length ϵ^{n} about the n^{th} rational in this list. The union of all of these intervals is an open set in I which is clearly dense since it contains all of the rationals in I. However, the total length of this set is quite small. Indeed, the length is given by

$$\sum_{n=1}^{\infty} \epsilon^n = \frac{\epsilon}{1-\epsilon}$$

This example shows clearly the difference between the topological approach to dynamics that we will adopt in the sequel and the measure theoretic approach. In a topological sense, an open, dense subset is considered "large." These sets may or may not be large in a measure theoretic sense, i.e., in the sense of total length.

2 Main Definitions



f $x, f(x), f^2(x)$

orbit

2.1 Orbits

The basic goal of the theory of dynamical systems is to understand the eventual or asymptotic behavior of an iterative process. If this process is a differential equation whose independent variable is time, then the theory attempts to predict the ultimate behavior of solutions of the equation in either the distant future ($t \to \infty$) or the distant past ($t \to -\infty$). If the process is a discrete process such as the iteration of a function, then the theory hopes to understand the eventual behavior of the points $x, f(x), f^2(x), \ldots, f^n(x)$ as n becomes large. That is, dynamical systems asks the somewhat nonmathematical sounding question: where do points go and what do they do when they get there? In this part of the course, we will attempt to answer this question at least partially for one of the simplest classes of dynamical systems, functions of a single real variable. Functions which determine dynamical systems are also called mappings, or maps, for short. This terminology connotes the geometric process of taking one point to another. As much of the sequel will in fact be geometric, we will use all of these terms synonymously.

Definition 1. The forward orbit of x is the set of points $x, f(x), f^2(x), \ldots$ and is denoted by $O^+(x)$. If f is a homeomorphism, we may define the full orbit of $x, O(x)$, as the set of points $f^n(x)$ for $n \in Z$, and the backward orbit of x, $O^-(x)$, as the set of points $x, f^{-1}(x), f^{-2}(x), \ldots$ Thus our basic goal is to understand all orbits of a map.

Orbits and forward orbits of points can be quite complicated sets, even for very simple nonlinear mappings. However, there are some orbits which are especially simple and which will play a central role in the study of the entire system.

Definition 2. The point x is a fixed point for f if $f(x) = x$. The point x is a periodic point of period n if $f^n(x) = x$. The least positive n for which $f^n(x) = x$ is called the prime period of x. We denote the set of periodic points of (not necessarily prime) period n by $\mathrm{Per}_n(f)$, and the set of fixed points by $\mathrm{Fix}\,(f)$. The set of all iterates of a periodic point form a periodic orbit.

Maps may have many fixed points. For example, the identity map $Id(x) = x$ fixes all points in $\mathbf{R}$, whereas the map $f(x) = -x$ fixes the origin, while all other points have period 2. These, however, are atypical dynamical systems; maps with intervals of fixed or periodic points are rare in a sense which will be made precise later. Most of the dynamical systems we will encounter will have isolated periodic points.

Example 1. The map $f(x) = x^3$ has 0,1 and -1 as fixed points and no other periodic points. The map $P(x) = x^2 - 1$ has fixed points at $(1 \pm \sqrt{5})/2$, while the points 0 and -1 lie on a periodic orbit of period 2 .

Example 2. Let S^1 denote the unit circle in the plane. We denote a point in S^1 by its angle θ measured in radians in the standard manner. Hence a point is determined by any angle of the form $\theta + 2k\pi$ for an integer k. Now let $f(\theta) = 2\theta$. (Note that $f(\theta + 2\pi) = f(\theta)$ on the circle so this map is well defined.) Now $f^n(\theta) = 2^n\theta$, so that θ is periodic of period n if and only if $2^n\theta = \theta + 2k\pi$ for some integer k, i.e., if and only if $\theta = 2k\pi/\left(2^n - 1\right)$ where $0 \leq k \leq 2^n$ is an integer. Hence the periodic points of period n for f are the $\left(2^{\bar{n}} - 1\right)^{\text{th}}$ roots of unity. It follows that the set of periodic points are dense in S^1.

Definition 3. A point x is eventually periodic of period n if x is not periodic but there exists $m > 0$ such that $f^{n+i}(x) = f^i(x)$ for all $i \geq m$. That is, $f^i(x)$ is periodic for $i \geq m$.

Example 3. Let $f(x) = x^2$. Then $f(1) = 1$ is fixed, while $f(-1) = 1$ is eventually fixed.

Example 4. Let $f(\theta) = 2\theta$ on the circle. Note that $f(0) = 0$ is fixed. If $\theta = 2k\pi/2^n$ then $f^n(\theta) = 2k\pi$ so that θ is eventually fixed. It follows that eventually fixed points are also dense in S^1. We remark that eventually periodic points cannot occur if the map is a homeomorphism.

$f^n(p) = p$. i.e. $p \in Per_n f$

Definition 4. Let p be periodic of period n. A point x is forward asymptotic to p if $\lim_{i\to\infty} f^{in}(x) = p$. The stable set of p, denoted by $W^s(p)$, consists of all points forward asymptotic to p.

If p is non-periodic, we may still define forward asymptotic points by requiring $\left|f^i(x) - f^i(p)\right| \to 0$ as $i \to \infty$. Also, if f is invertible, we may consider backward asymptotic points by letting $i \to -\infty$ in the above definition. The set of points backwards asymptotic to p is called the unstable set of p and is denoted by $W^u(p)$.

Example 5. Let $f(x) = x^3$. Then $W^s(0)$ is the open interval $-1 < x < 1$. $W^u(1)$ is the positive real axis, whereas $W^u(-1)$ is the negative real axis.

Definition 5. A point x is a critical point of f if $f'(x) = 0$. The critical point is non-degenerate if $f''(x) \neq 0$. The critical point is degenerate if $f''(x) = 0$.

For example $f(x) = x^2$ has a non-degenerate critical point at 0, but $f(x) = x^n$ for $n > 2$ has a degenerate critical point at 0. Note that degenerate critical points may be maxima, minima, or saddle points (as in the case of $f(x) = x^3$). But non-degenerate critical points must be either maxima or minima. Critical points cannot occur for diffeomorphisms, but their existence for non-invertible maps is one reason why these kinds of maps are more complicated.

The goal of dynamical systems is to understand the nature of all orbits, and to identify the set of orbits which are periodic, eventually periodic, asymptotic, etc. Generally, this is an impossible task. For example, if $f(x)$ is a quadratic polynomial, then finding explicitly the periodic points of period n necessitates solving the equation $f^n(x) = x$, which is a polynomial equation of degree 2^n. A computer does not help matters much, for numerical computations of periodic points are often misleading. Round-off errors tend to accumulate and make many periodic points invisible to the computer. Therefore we are left with only qualitative or geometric techniques to understand the dynamics of a given system. This means that we should look for a geometric picture of the behavior of all orbits of a system. This geometric picture is provided by the phase portrait which we now discuss.

The graph of a function on the reals provides information about its first iterate, but gives very little information about subsequent iterates. To understand higher iterates, we could attempt to sketch each of their graphs, but this is a cumbersome procedure. There is a much more efficient, geometric method for describing the orbits of a dynamical system, the phase portrait. This is a picture, on the real line itself, as opposed to the plane, of all orbits of a system. For example, to indicate that all non-zero orbits of $f(x) = -x$ have period 2, we could sketch the phase portrait as in Fig. 1-a. This figure also depicts the phase portraits of some other simple maps.

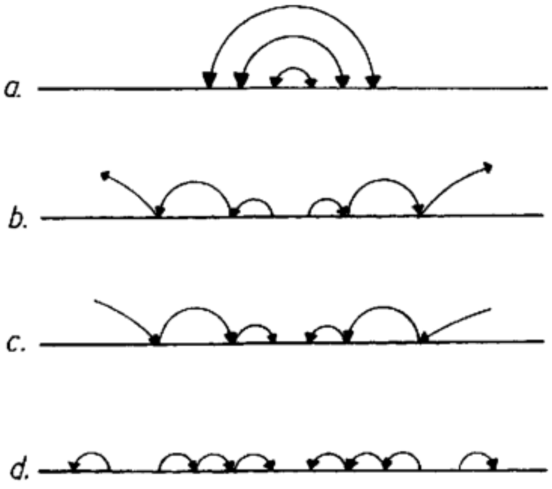


Figure 1: The phase portraits of a. $f(x) = -x$, b. $f(x) = 2x$, c. $f(x) = \frac{1}{2}x$, d. $f(x) = x^3$.

2.2 Graphical analysis

The graph of $f(x)$ does of course contain information about the first iteration of f. We may use it to gain insight into higher iterations and hence the phase portrait 相轨迹图 via the following procedure which we call graphical analysis. Identify the diagonal $\Delta = \{(x,x) \mid x \in \mathbf{R}\}$ with $\mathbf{R}$ in the obvious way. A vertical line from (p,p) to the graph of f meets the graph at $(p, f(p))$. Then a horizontal line from $(p, f(p))$ to Δ meets the diagonal at $(f(p), f(p))$. Hence a vertical line to the graph followed by a horizontal line back to Δ yields the image of the point p under f on the diagonal. We may thus visualize the phase portrait of a map as taking place on the diagonal rather than on the x-axis. Then an orbit is given by repeatedly drawing line segments vertically from Δ to the graph and then horizontally from the graph to Δ. Fig. 2 illustrates this procedure for $f(x) = 2x - x^2$.

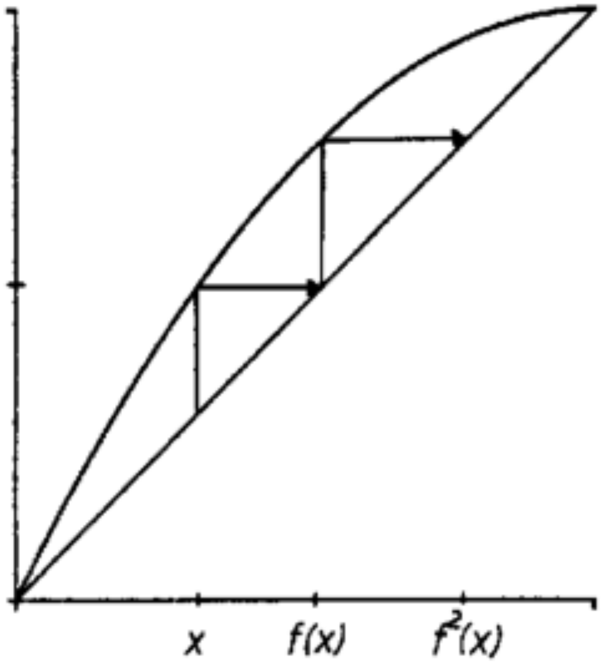


Figure 2: $f(x) = 2x - x^2$

Diffeomorphisms of the circle form an interesting class of maps which are somewhat different from maps of $\mathbf{R}$. The following example is typical.

Example 6. Let $f(\theta) = \theta + \epsilon \sin(2\theta)$ for $0 < \epsilon < 1/2$. Note that f has fixed points at $0, \pi/2, \pi$, and $3\pi/2$. 0 and π are repelling fixed points and $\pi/2$ and $3\pi/2$ are attracting. More generally, $f(\theta) = \theta + \epsilon \sin(N\theta)$ has N attracting and N repelling fixed points arranged alternately around the circle as long as $0 < \epsilon < 1/N$.

Another important class of circle maps are the translation maps.

Example 7. Translations of the circle. Let $\lambda \in \mathbf{R}$ and $T_\lambda(\theta) = \theta + 2\pi\lambda$. The maps T_λ behave quite differently depending upon the rationality or irrationality of λ. If $\lambda = p/q$, where p and q are integers, then $T_\lambda^q(\theta) = \theta +\ 2\pi p = \theta$ so that all points are fixed by T_λ^q. When λ is irrational, the situation is quite different. The following result is known as Jacobi's Theorem.

Theorem 1. Each orbit T_λ is dense in S^1 if λ is irrational.
Proof. Let $\theta \in S^1$. The points on the orbit of θ are distinct. For $T_\lambda^n(\theta) = T_\lambda^m(\theta)$ we would have $(n-m)\lambda \in \mathbf{Z}$, so that $n = m$. Any infinite set of points on the circle must have a limit point. Thus, given any $\epsilon > 0$, there must be integers n and m for which $|T_\lambda^n(\theta) - T_\lambda^m(\theta)| < \epsilon$. Let $k = n - m$. Then $\left|T_\lambda^k(\theta) - \theta\right| < \epsilon$.

Now T_λ preserves lengths in S^1. Consequently, T_λ^k maps the arc connecting θ to $T_\lambda^k(\theta)$ to the arc connecting $T_\lambda^k(\theta)$ and $T_\lambda^{2k}(\theta)$ which has length less than ϵ. In particular it follows that the points $\theta, T_\lambda^k(\theta), T_\lambda^{2k}(\theta), \ldots$ partition S^1 into arcs of length less than ϵ. Since ϵ was arbitrary. This completes the proof.

Exercises.

3 Hyperbolicity

Simple maps like $id(x) = x$ and $f(x) = -x$ are, unfortunately, atypical among dynamical systems. There are many reasons why this is so, but perhaps the most unusual feature of these maps is the fact that all points are periodic under iteration of these maps. Most maps do not have this type of behavior. Periodic points tend to be more spread out on the line. In this section we will introduce one of the main themes of this book, hyperbolicity. Maps with hyperbolic periodic points are the ones that occur typically in many dynamical systems and, moreover, they provide the simplest types of periodic behavior to analyze.

Definition 1. Let p be a periodic point of prime period n. The point p is hyperbolic if $\left|(f^n)'(p)\right| \neq 1$. The number $(f^n)'(p)$ is called the multiplier of the periodic point.

Example 1. Consider the diffeomorphism $f(x) = \frac{1}{2}\left(x^3 + x\right)$. There are 3 fixed points: $x = 0, 1,$ and -1 . Note that $f'(0) = 1/2$ and $f'(\pm 1) = 2$. Hence each fixed point is hyperbolic.

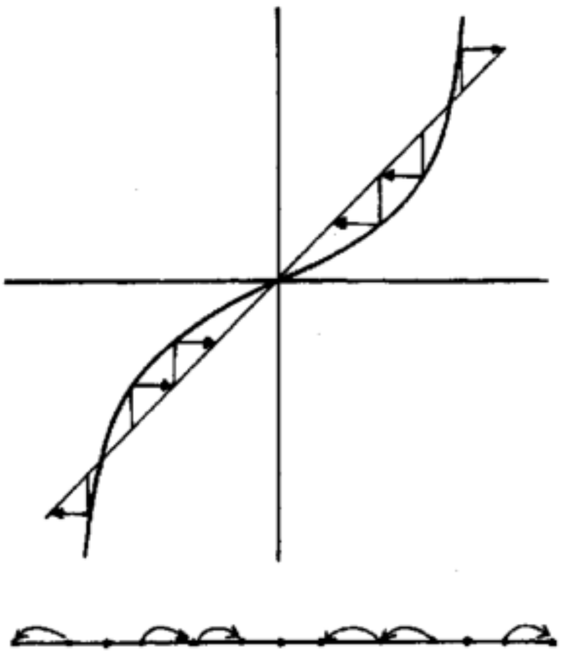


Figure 3: Example 1

Example 2. Let $f(x)=-\frac{1}{2}\left(x^{3}+x\right)$. 0 is a hyperbolic fixed point, with $f'(0)=-\frac{1}{2}$. The points ± 1 now lie on a periodic orbit of period 2 . We compute $\left(f^{2}\right)'(\pm 1)=f'(1)\cdot f'(-1)=4$ by the chain rule. Hence this periodic point is hyperbolic. Note that points in the interval $(-1,1)$ spiral toward 0 and away from ± 1.

We observe that, in the above two examples, we have $|f'(0)|<1$ and that points close to 0 are forward asymptotic to 0 . This situation occurs often.

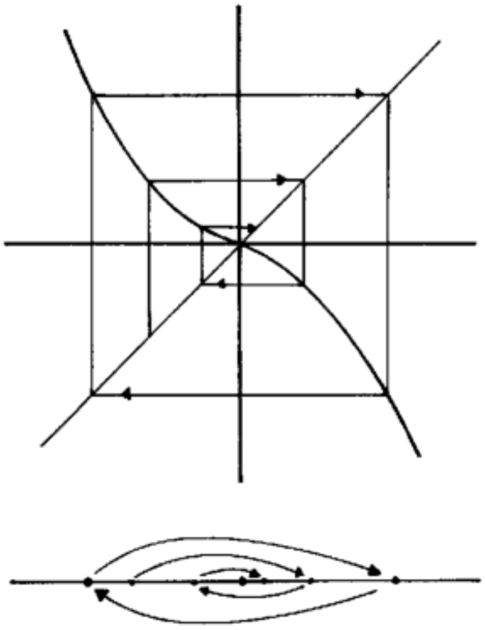
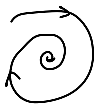


Figure 4: Example 2

Proposition 1. Let p be a hyperbolic fixed point with $|f'(p)| < 1$. Then there is an open interval U about p such that if $x \in U$, then

in graph



$$\lim_{n\to\infty} f^n(x) = p$$

Proof. Since f is C^1, there is $\epsilon > 0$ such that $|f'(x)| < A < 1$ for $x \in [p-\epsilon, p+\epsilon]$. By the Mean Value Theorem

$$|f(x) - p| = |f(x) - f(p)| \leq A|x-p| < |x-p| \leq \epsilon$$

Hence $f(x)$ is contained in $[p-\epsilon, p+\epsilon]$ and, in fact, is closer to p than x is. Via the same argument

$$|f^n(x) - p| \leq A^n |x - p|$$

so that $f^n(x) \to p$ as $n \to \infty$.

Remarks.

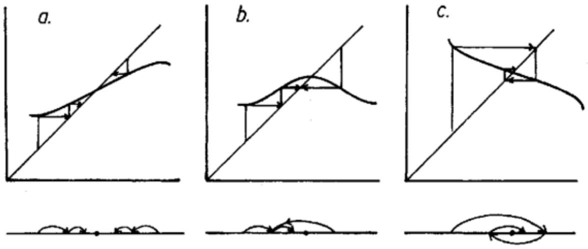
1. It follows that the interval $[p - \epsilon, p + \epsilon]$ is contained in the stable set associated to $p, W^s(p)$.

2. A similar result is true for hyperbolic periodic points of period n. In this case, we get an open interval U about p which is mapped inside itself by f^n. Of course, the assumption in this case is that $\left|(f^n)'(p)\right| < 1$.

Definition 2. Let p be a hyperbolic periodic point of period n with $\left|(f^n)'(p)\right| < 1$. The point p is called an attracting periodic point (an attractor) or a sink. (向 fix point “旋进”).

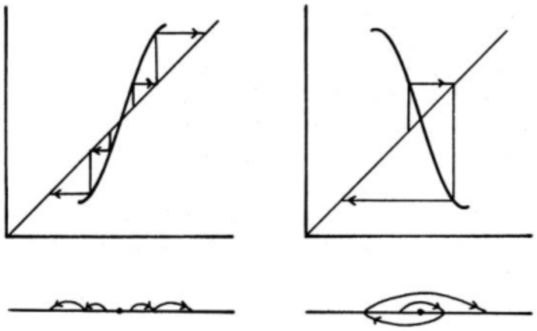
Attracting periodic points of period n thus have neighborhoods which are mapped inside themselves by f^n. Such a neighborhood is called the local stable set and is denoted by W^s_{loc}. We may actually distinguish three different types of attracting fixed points, namely those where $f'(p) = 0, 0 < f'(p) < 1$, and $-1 < f'(p) < 0$.

The behavior of a map near periodic points where the derivative is larger than one in absolute value is quite different from that of sinks.



Proposition 2. Let p be a hyperbolic fixed point with $|f'(p)| > 1$. Then there is an open interval U of p such that, if $x \in U, x \neq p$, then there exists $k > 0$ such that $f^k(x) \notin U$. （向外"跳跃"）、

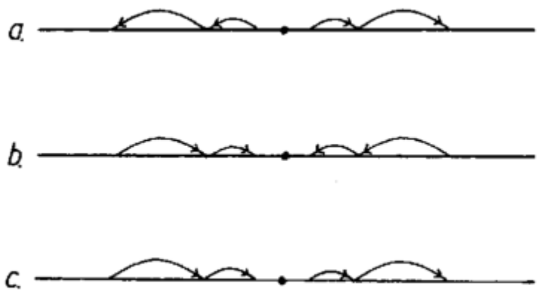
The proof is similar to the proof of the preceding proposition and is therefore left as an exercise. Graphically, the result is quite clear.



Definition 3. A fixed point p with $|f'(p)| > 1$ is called a repelling fixed point (a repellor) or source. The neighborhood described in the Proposition is called the local unstable set and denoted W^u_{loc}.

We remark that periodic points of period n exhibit similar behavior when $\left|(f^n)'(p)\right| > 1$. Hyperbolic periodic points therefore have local behavior which is governed by the derivative at the periodic point. This is not true when the point is indifferent or non-hyperbolic, as the following example shows.

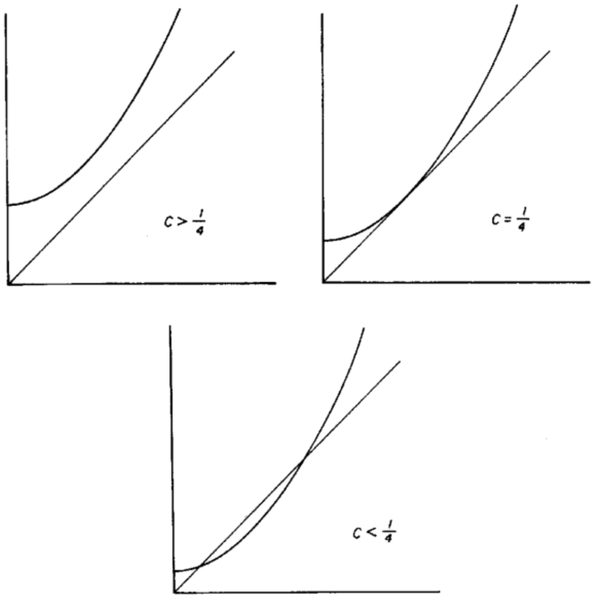
Example 3. Each of the maps in Figure satisfy $f(0) = 0$ and $f'(0) = 1$, but each have vastly different phase portraits near 0 . In a., the map $f(x) = x + x^3$ has a weakly repelling fixed point at 0 . In b., the map $f(x) = x - x^3$ has a weakly attracting fixed point at 0 . In c., the map $f(x) = x + x^2$ is weakly repelling from the right but weakly attracting from the left.



Most maps have only hyperbolic periodic points, as we shall see later. However, non-hyperbolic periodic points often occur in families of maps.

When this happens, the periodic point structure often undergoes a bifurcation. We will deal with bifurcation theory more extensively later, but for now we give several examples.

Example 4. Consider the family of quadratic functions $Q_c(x) = x^2 + c$, where c is a parameter. The graphs of Q_c assume three different positions relative to the diagonal depending upon whether $c > 1/4, c = 1/4$, or $c < 1/4$. Note that Q_c has no fixed points for $c > 1/4$. When $c = 1/4, Q_c$ has a unique non-hyperbolic fixed point at $x = 1/2$. And when $c < 1/4, Q_c$ has a pair of fixed points, one attracting and one repelling. Thus the phase portrait of Q_c changes as c decreases through 1/4. This change is an example of a bifurcation.

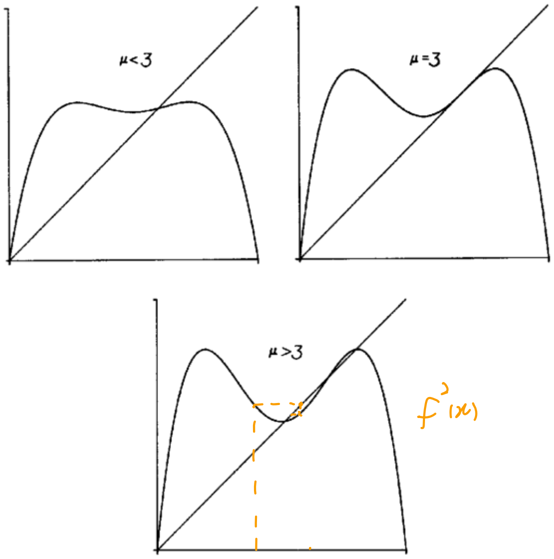


$\mu(x-x^2)$

$\mu - 2\mu x\left(\frac{\mu-1}{\mu}\right)$

Example 5. Let $F_\mu(x) = \mu x(1 - x)$ with $\mu > 1$. F_μ has two fixed points: one at 0 and the other at $p_\mu = (\mu - 1)/\mu$. Note that $F'_\mu(0) = \mu$ and

$F'_\mu(p_\mu) = 2 - \mu$. Hence 0 is a repelling fixed point for $\mu > 1$ and p_μ is attracting for $1 < \mu < 3$. When $\mu = 3, F'_\mu(p_\mu) = -1$. We sketch the graphs of F^2_μ for μ near 3. Note that 2 new fixed points for F^2_μ appear as μ increases through 3. These are new periodic points of period 2. Another bifurcation has occurred: this time we have a change in $\mathrm{Per}_2(F_\mu)$.



This quadratic family actually exhibits many of the phenomena that are crucial in the general theory. The next section is devoted entirely to this function.

Exercises

4 The quadratic family

In this paragraph, we will continue the discussion of the quadratic family $F_\mu(x) = \mu x(1-x)$. Actually, we will return to this example repeatedly throughout the remainder of this course, since it illustrates many of the most important phenomena that occur in dynamical systems.

$\mu - 2\mu x = 1$

Proposition 1. a. $F_\mu(0) = F_\mu(1) = 0$ and $F_\mu(p_\mu) = p_\mu$, where $p_\mu = \frac{\mu-1}{\mu}$.
b. $0 < p_\mu < 1$ if $\mu > 1$.

The proof of this proposition is straightforward.

From now on we will concentrate on the case $\mu > 1$. The following proposition shows that most points behave rather tamely under iteration of F_μ : all points which do not lie in the interval $[0,1]$ tend to $-\infty$.

Proposition 2. Suppose $\mu > 1$. If $x < 0$, then $F_\mu^n(x) \to -\infty$ as $n \to \infty$. Similarly, if $x > 1$, then $F_\mu^n(x) \to -\infty$ as $n \to \infty$.

Proof. If $x < 0$, then $\mu x(1-x) < x$ so $F_\mu(x) < x$. Hence $F_\mu^n(x)$ is a decreasing sequence of points. This sequence cannot converge to p, for then we would have $F_\mu^{n+1}(x) \to F_\mu(p) < p$, whereas $F_\mu^n(x) \to p$. Hence

$F_\mu^n(p) \to -\infty$ as required. If $x > 1$, then $F_\mu(x) < 0$ so $F_\mu^n(x) \to -\infty$ as well. q.e.d.

Graphical analysis yields the above results easily. As a consequence of this Proposition, all of the interesting dynamics of the quadratic family occur in the unit interval $I = \{x \mid 0 \leq x \leq 1\}$. For low values of μ, the dynamics of F_μ are not too complicated.

Proposition 3. Let $1 < \mu < 3$.

1. F_μ has an attracting fixed point at $p_\mu = (\mu - 1)/\mu$ and a repelling fixed point at 0.

2. If $0 < x < 1$, then

x

$$\lim_{n\to\infty} F_\mu^n(x) = p_\mu$$

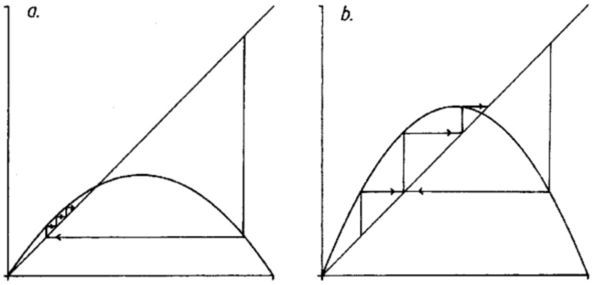
Proof. Part 1 was proved in Example 4 at the end of the last section. For part 2, we first deal with the case $1 < \mu < 2$. Suppose x lies in the interval $(0, 1/2]$. Then graphical analysis immediately shows that

$$|F_\mu(x) - p_\mu| < |x - p_\mu|$$

if $x \neq p_\mu$. Consequently, $F_\mu^n(x) \to p_\mu$ as $n \to \infty$. If, on the other hand, x lies in the interval $(1/2, 1)$, then $F_\mu(x)$ lies in $(0, 1/2)$, so that the previous argument implies

$$F_\mu^n(x) = F_\mu^{n-1}(F_\mu(x)) \to p_\mu$$

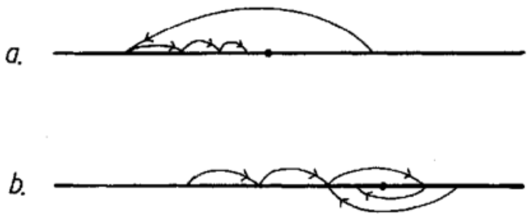
as $n \to \infty$.



The case when $2 < \mu < 3$ is more difficult. Graphical analysis shows what is different in this case. Note that $1/2 < p_\mu < 1$. Let $\hat{p}_\mu$ denote the unique point in the interval $(0, 1/2)$ that is mapped onto p_μ by F_μ. Then it's easy check that F_μ^2 maps the interval $[\hat{p}_\mu, p_\mu]$ inside $[1/2, p_\mu]$. It follows that $F_\mu^n(x) \to p_\mu$ as $n \to \infty$ for all $x \in [\hat{p}_\mu, p_\mu]$. Now suppose $x < \hat{p}_\mu$. Again graphical analysis shows that there exists $k > 0$ such that $F_\mu^k(x) \in [\hat{p}_\mu, p_\mu]$. Thus $F_\mu^{k+n}(x) \to p_\mu$ as $n \to \infty$ in this case as well. Finally, as before, F_μ maps the interval $(p_\mu, 1)$ onto $(0, p_\mu)$, so the result follows here as well. Since $(0, 1) = (0, \hat{p}_\mu) \cup [\hat{p}_\mu, p_\mu] \cup (p_\mu, 1)$, we are finished. We leave the intermediate case $\mu = 2$ as the exercises.

Hence for $1 < \mu < 3$, F_μ has only two fixed points and all other points in I are asymptotic to p_μ. Thus the dynamics of F_μ are completely understood for μ in this range.

As we showed in Example 4 in the previous section, as μ passes through 3, the dynamics of F_μ become slightly more complicated: a new periodic point of period 2 is born. This is the beginning of a long story: as μ continues to increase the dynamics of F_μ become increasingly more complicated until the phase portrait of F_μ is dramatically different from the



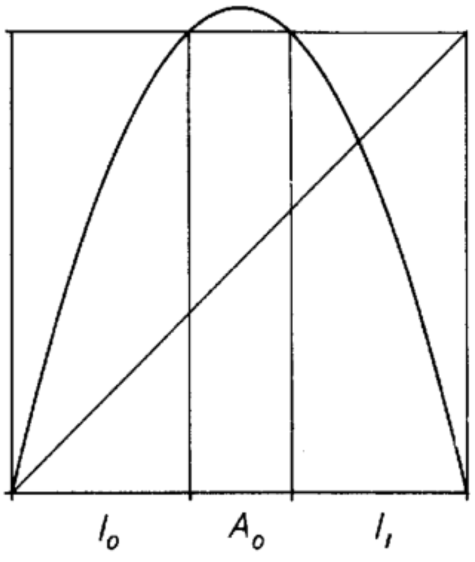
above picture. This is a scenario that we will investigate in much more detail later.

We now turn to the case when $\mu > 4$. For the remainder of this section, we will drop the subscript μ and write F instead of F_μ. As above, all of the interesting dynamics of F occur in the unit interval I. Note that, since $\mu > 4$, the maximum value $\mu/4$ of F is larger than one. Hence certain points leave I after one iteration of F. Denote the set of such points by A_0. Clearly, A_0 is an open interval centered at $\frac{1}{2}$ and has the property that, if $x \in A_0$, then $F(x) > 1$, so $F^2(x) < 0$ and $F^n(x) \to -\infty$. A_0 is the set of points which immediately escape from I. All other points in I remain in I after one iteration of F.

Let $A_1 = \{x \in I \mid F(x) \in A_0\}$. If $x \in A_1$, then $F^2(x) > 1, F^3(x) < 0$, and so, as before, $F^n(x) \to -\infty$.

Inductively, let $A_n = \{x \in I \mid F^n(x) \in A_0\}$.

That is, $A_n = \left\{x \in I \mid F^i(x) \in I \text{ for } i \leq n \text{ but } F^{n+1}(x) \notin I\right\}$, so that A_n consists of all points which escape from I at the $n + 1^{st}$ iteration. As above, if x lies in A_n, it follows that the orbit of x tends eventually to $-\infty$. Since we therefore know the ultimate fate of any point which lies in the A_n, it therefore remains only to analyze the behavior of those points which never escape from I, i.e., the set of points which lie in



Logistic map.

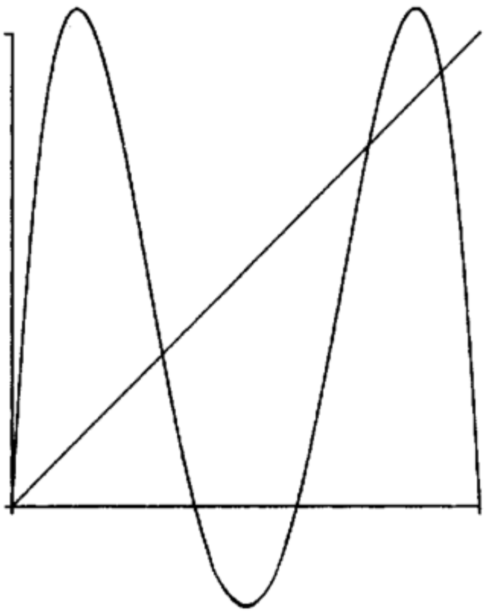
$$I - \left(\bigcup_{n=0}^{\infty} A_n \right).$$

Let us denote this set by Λ. Our first question is: what precisely is this set of points? To understand Λ, we describe more carefully its recursive construction.

Since A_0 is an open interval centered at $1/2, I - A_0$ consists of two closed intervals, I_0 on the left and I_1 on the right.

Note that F maps both I_0 and I_1 monotonically onto I; F is increasing on I_0 and decreasing on I_1. Since $F\left(I_0\right) = F\left(I_1\right) = I$, there are a pair of open intervals, one in I_0 and one in I_1, which are mapped into A_0 by F. Therefore this pair of intervals is precisely the set A_1.

Now consider $I - (A_0 \cup A_1)$. This set consists of 4 closed intervals and F maps each of them monotonically onto either I_0 or I_1. Consequently F^2 maps each of them onto I. We therefore see that each of the four intervals in $I-(A_0 \cup A_1)$ contains an open subinterval which is mapped by F^2 onto A_0. Therefore, points in these intervals escape from I upon the third iteration of F. This is the set we called A_2. For later use, we observe that F^2 is alternately increasing and decreasing on these four intervals. It follows that the graph of F^2 must therefore have two humps as shown in figure.



Continuing in this manner we note two facts. First, A_n consists of 2^n disjoint open intervals. Hence $I-(A_0 \cup \ldots \cup A_n)$ consists of 2^{n+1} closed intervals since

$$1+2+2^2+\ldots+2^n = 2^{n+1}-1$$

Secondly, F^{n+1} maps each of these closed intervals monotonically onto I. In fact, the graph of F^{n+1} is alternately increasing and decreasing on these intervals. Thus the graph of F^{n+1} has exactly 2^n humps on I, and it follows that the graph of F^n crosses the line $y = x$ at least 2^n times. This implies that F^n has at least 2^n fixed points or, equivalently, $\text{Per}_n(F)$ consists of 2^n points in I. Clearly, the structure of Λ is much more complicated when $\mu > 4$ than the earlier case $\mu < 3$.

The construction of Λ is reminiscent of the construction of the Cantor Middle Thirds set: Λ is obtained by successively removing open intervals from the "middles" of a set of closed intervals.

Definition 1. A set Λ is a Cantor set if it is a closed, totally disconnected, and perfect subset of I. A set is totally disconnected if it contains no intervals; a set is perfect if every point in it is an accumulation point or limit point of other points in the set.

Example 1. The Cantor Middle-Thirds Set. This is the classical example of a Cantor set. Start with I but remove the open "middle third," i.e., the interval $\left(\frac{1}{3}, \frac{2}{3}\right)$. Next, remove from what remains the two middle thirds again, i.e., the pair of intervals $\left(\frac{1}{9}, \frac{2}{9}\right)$ and $\left(\frac{7}{9}, \frac{8}{9}\right)$. Continue removing middle thirds in this fashion; note that 2^n open intervals are removed at the n^{th} stage of this process. Thus, this procedure is entirely analogous to our construction above.

Remark. The Cantor Middle-Thirds set is an example of a fractal. Intuitively, a fractal is a set which is self-similar under magnification. In the Cantor Middle-Thirds set, suppose we look only at those points which lie in the left-hand interval $\left[0, \frac{1}{3}\right]$. Under a microscope which magnifies this interval by a factor of three, the "piece" of the Cantor set in $\left[0, \frac{1}{3}\right]$ looks exactly like the original set. More precisely, the linear map $L(x) = 3x$ maps the portion of the Cantor set in $\left[0, \frac{1}{3}\right]$ homeomorphically onto the entire set. This process does not stop at the first level: one may magnify any piece of the Cantor set at the n^{th} stage of the construction by a factor of 3^n and obtain the original set.

To guarantee that our set Λ is a Cantor set, we need an additional hypothesis on μ. Let us assume that μ is large enough so that $|F'(x)| > 1$ for all $x \in I_0 \cup I_1$. It's easy to check that $\mu > 2 + \sqrt{5}$ suffices. Hence, for these values of μ, there exists $\lambda > 1$ such that $|F'(x)| > \lambda$ for all $x \in \Lambda$. By the chain rule, it follows that $\left|(F^n)'(x)\right| > \lambda^n$ as well. We claim that Λ contains no intervals. Indeed, if this were so, we could choose $x, y \in \Lambda$, $x \neq y$, with the closed interval $[x, y] \subset \Lambda$. But then, $\left|(F^n)'(\alpha)\right| > \lambda^n$ for all $\alpha \in [x, y]$. Choose n so that $\lambda^n|y - x| > 1$. By the Mean Value Theorem, it then follows that $|F^n(y) - F^n(x)| \geq \lambda^n|y - x| > 1$, which implies that at least one of $F^n(y)$ or $F^n(x)$ lies outside of I. This is a contradiction, and so Λ is totally disconnected.

Since Λ is a nested intersection of closed intervals, Λ is closed. We now prove that Λ is perfect. First note that any endpoint of an A_k is in Λ : indeed, such points are eventually mapped to the fixed point at 0, and so they stay in I under iteration. Now if $p \in \Lambda$ were isolated, every nearby point must leave I under iteration of F. Such points must belong to some A_k. Either there is a sequence of endpoints of the A_k converging to p, or else all points in a deleted neighborhood of p are mapped out of I by some power of F. In the former case, we are done as the endpoints of the A_k map to 0 and hence are in Λ. In the latter, we may assume that F^n maps p to 0 and all other points in a neighborhood of p into the negative real axis. But then F^n has a maximum at p so that $(F^n)'(p) = 0$. By the chain rule, we must have $F'\left(F^i(p)\right) = 0$ for some $i < n$. Hence $F^i(p) = 1/2$. But then $F^{i+1}(p) \notin I$ and so $F^n(p) \to -\infty$, contradicting the fact that $F^n(p) = 0$.

Hence we have proved

Theorem 1. If $\mu > 2 + \sqrt{5}$, then Λ is a Cantor set.

Remark. The theorem is true for $\mu > 4$, but the proof is more delicate.

We have now succeeded in understanding the gross behavior of orbits of F_μ when $\mu > 4$. Either a point tends to $-\infty$ under iteration of F_μ, or else its entire orbit lies in Λ. Hence we understand the orbit of a point under F_μ perfectly well as long as the point does not lie in Λ. In the next section, we will complete the analysis of the dynamics of F_μ by analyzing the dynamics of F_μ on Λ.

When $\mu > 2+\sqrt{5}$, we have shown that $\left|F_\mu'(x)\right| > 1$ on $I_0 \cup I_1$. This implies that $\left|F_\mu'(x)\right| > 1$ on Λ. This is a condition similar to the hyperbolicity condition, except that we require $\left|F_\mu'(x)\right| \neq 1$ on a whole set, not just at a periodic point. This motivates the definition of a hyperbolic set:

Definition 2. A set $\Gamma \subset \mathbf{R}$ is a repelling (resp. attracting) hyperbolic set for f if Γ is closed, bounded and invariant under f and there exists an

$N > 0$ such that $\left|(f^n)'(x)\right| > 1($ resp. $< 1)$ for all $n \geq N$ and all $x \in \Gamma$.

The Cantor set Λ for the quadratic map when $\mu > 2 + \sqrt{5}$ is of course a repelling hyperbolic set with $N = 1$.

5 Symbolic dynamics

Our goal in this section is to give a model for the rich dynamical structure of the quadratic map on the Cantor set Λ discussed in the previous section. To do this we will set up a model mapping which is completely equivalent to F. At first, this model may seem artificial and unintuitive. But, as we go along, it will become clear that such symbolic models describe the dynamics of F completely and also in the simplest possible way.

We need a "space" on which our model map will act. The points in this space will be infinite sequences of 0's and 1's. We don't worry about convergence of these sequences; rather, the difficult notion here is to imagine such an infinite sequence as representing a single "point" in space.

Definition 1. $\Sigma_2 = \{\mathbf{s} = (s_0 s_1 s_2 \ldots) \mid s_j = 0 \text{ or } 1\}$. Σ_2 is called the sequence space on the two symbols 0 and 1. More generally, we can consider the space Σ_n consisting of infinite sequences of integers between 0 and $n-1$. Elements of Σ_2 are infinite strings of integers, like $(000\ldots)$ or $(0101\ldots)$. We may make Σ_2 into a metric space as follows. For two sequences $\mathbf{s} = (s_0 s_1 s_2 \ldots)$ and $\mathbf{t} = (t_0 t_1 t_2 \ldots)$, define the distance between them by

$$d[\mathbf{s},\mathbf{t}] = \sum_{i=0}^{\infty} \frac{|s_i - t_i|}{2^i}$$

Since $|s_i - t_i|$ is either 0 or 1 , this infinite series is dominated by the geometric series

$$\sum_{i=0}^{\infty} \frac{1}{2^i} = 2$$

and therefore it converges. For example, if $\mathbf{s} = (000\ldots)$ and $\mathbf{t} = (111\ldots)$, then $d[\mathbf{s},\mathbf{t}] = 2$. If $\mathbf{r} = (1010...)$, then

$$d[\mathbf{s},\mathbf{r}] = \sum_{i=0}^{\infty} \frac{1}{2^{2i}} = \frac{1}{1-\frac{1}{4}} = \frac{4}{3}.$$

Proposition 1. d is a metric on Σ_2.

Proof. Clearly, $d[\mathbf{s},\mathbf{t}] \geq 0$ for any $\mathbf{s},\mathbf{t} \in \Sigma_2$, and $d[\mathbf{s},\mathbf{t}] = 0$ iff $s_i = t_i$ for all i. Since $|s_i - t_i| = |t_i - s_i|$, it follows that $d[\mathbf{s},\mathbf{t}] = d[\mathbf{t},\mathbf{s}]$. Finally, if $\mathbf{r},\mathbf{s}$, and $\mathbf{t} \in \Sigma_2$, then $|r_i - s_i| + |s_i - t_i| \geq |r_i - t_i|$ from which we deduce that $d[\mathbf{r},\mathbf{s}] + d[\mathbf{s},\mathbf{t}] \geq d[\mathbf{r},\mathbf{t}]$.

The metric d allows us to decide which subsets of Σ_2 are open and which are closed, as well as which sequences are close to each other.

Proposition 2. Let $\mathbf{s}, \mathbf{t} \in \Sigma_2$ and suppose $s_i = t_i$ for $i = 0, 1, \ldots, n$. Then $d[\mathbf{s}, \mathbf{t}] \leq 1/2^n$. Conversely, if $d[\mathbf{s}, \mathbf{t}] < 1/2^n$, then $s_i = t_i$ for $i \leq n$.

Proof. If $s_i = t_i$ for $i \leq n$, then

$$\begin{aligned} d[\mathbf{s}, \mathbf{t}] &= \sum_{i=0}^{n} \frac{|s_i - s_i|}{2^i} + \sum_{i=n+1}^{\infty} \frac{|s_i - t_i|}{2^i} \\ &\leq \sum_{i=n+1}^{\infty} \frac{1}{2^i} = \frac{1}{2^n} \end{aligned}$$

On the other hand, if $s_j \neq t_j$ for some $j \leq n$, then we must have

$$d[\mathbf{s}, \mathbf{t}] \geq \frac{1}{2^j} \geq \frac{1}{2^n}$$

consequently, if $d[\mathbf{s}, \mathbf{t}] < 1/2^n$, then $s_i = t_i$ for $i \leq n$.

The importance of this result is that we can decide quickly whether or not two sequences are close to each other. Intuitively, this result says that two sequences in Σ_2 are close provided their first few entries agree. We now define the most important ingredient in symbolic dynamics, the shift map on Σ_2.

Definition 2. The shift map $\sigma : \Sigma_2 \to \Sigma_2$ is given by $\sigma\left(s_0 s_1 s_2 \ldots\right) = (s_1 s_2 s_3 \ldots)$.

The shift map simply "forgets" the first entry in a sequence, and shifts all other entries one place to the left. Clearly, σ is a two-to-one map of Σ_2, as s_0 may be either 0 or 1 . Moreover, in the metric defined above, σ is a continuous map.

Proposition 3. $\sigma : \Sigma_2 \to \Sigma_2$ is continuous.

Proof. Let $\epsilon > 0$ and $\mathbf{s} = s_0 s_1 s_2 \ldots$. Pick n such that $1/2^n < \epsilon$. Let $\delta = 1/2^{n+1}$. If $\mathbf{t} = t_0 t_1 t_2 \ldots$ satisfies $d[\mathbf{s}, \mathbf{t}] < \delta$, then by Proposition 2 we have $s_i = t_i$ for $i \leq n+1$. Hence the i^{th} entries of $\sigma(\mathbf{s})$ and $\sigma(\mathbf{t})$ agree for $i \leq n$. Therefore $d[\sigma(\mathbf{s}), \sigma(\mathbf{t})] \leq 1/2^n < \epsilon$.

In the next section, we will show that the shift map is an exact model for the quadratic map F_μ when $\mu > 4$. Here we will simply show that the dynamics of σ can be understood completely.

For example, periodic points correspond exactly to repeating sequences, i.e., sequences of the form $\mathbf{s} = (s_0 \ldots s_{n-1}, s_0 \ldots s_{n-1}, s_0 \ldots s_{n-1} \ldots)$. Hence there are 2^n periodic points of period n for σ, each generated by one of the 2^n finite sequence of 0 's and 1's of length n.

Eventually periodic points are equally abundant and easy to recognize. For example, any sequence of the form $(s_0 \ldots s_n 1111 \ldots)$ is eventually fixed, while any eventually repeating sequence is eventually periodic for σ.

Another interesting fact about σ is that periodic points form a dense subset of Σ_2. Recall that a subset is dense in Σ_2 provided its closure is the entire space Σ_2. To prove that $\text{Per}(\sigma)$ is dense, we must produce a sequence of periodic points τ_n which converge to an arbitrary point $\mathbf{s} = (s_0 s_1 s_2 \ldots)$ in Σ_2. We define the sequence $\tau_n = (s_0 \ldots s_n, s_0 \ldots s_n, \ldots)$, i.e., τ_n is the repeating sequence whose entries agree with s up to the n^{th} entry. By Proposition 2, $d\,[\tau_n, \mathbf{s}] \leq 1/2^n$, so that we have $\tau_n \to \mathbf{s}$.

Of course, not all points in Σ_2 are periodic or eventually periodic. Any non-repeating sequence can never be periodic. In fact, the non-periodic sequences greatly outnumber the periodic sequences in Σ_2. Moreover, there are non-periodic orbits in Σ_2 which wind densely about Σ_2, i.e., the closure of the orbit is Σ_2 itself. Another way to say this is there are points in Σ_2 whose orbit comes arbitrarily close to any given sequence in Σ_2. To see this, consider

$$\mathbf{s}^* = (\ \underbrace{01}_{\text{1 blocks}}\ |\underbrace{00011011}_{\text{2 blocks}}|\underbrace{000001\cdots}_{\text{3blocks}}|\ \underbrace{\cdots}_{\text{4blocks}}\ )$$

$\mathbf{s}^*$ is constructed by successively listing all blocks of 0's and 1's of length n, then length $n+1$, etc. Clearly, some iterate of σ applied to $\mathbf{s}^*$ yields a sequence which agrees with any given sequence in an arbitrarily large number of places. Mappings which have dense orbits are called topologically transitive.

Let us list these properties of σ :

Proposition 4. a. The cardinality of $\mathrm{Per}_n(\sigma)$ is 2^n. b. $\mathrm{Per}(\sigma)$ is dense in Σ_2. c. There exists a dense orbit for σ in Σ_2.

In the next section, we will show that the shift map on Σ_2 is in fact the "same" map as f on Λ.

Symbolic dynamics is one of the main themes of this course. It will appear in various guises throughout, including later in this chapter when we introduce subshifts of finite type and also the kneading theory to describe the dynamics of F_μ when $\mu < 4$.