MATHEMATICAL MODELING

AND

ORDINARY DIFFERENTIAL EQUATIONS

I-Liang Chern

Department of Mathematics
National Taiwan University
2007, 2015

January 6, 2016

Contents

Chapter 1

First-Order Single Differential Equations

1.1 What is mathematical modeling?

In science, we explore and understand our real world by observations, collecting data, finding rules inside or among them, and eventually, we want to explore the truth behind and to apply it to predict the future. This is how we build up our scientific knowledge. The above rules are usually in terms of mathematics. They are called mathematical models. One important such models is the ordinary differential equations. It describes relations between variables and their derivatives. Such models appear everywhere. For instance, population dynamics in ecology and biology, mechanics of particles in physics, chemical reaction in chemistry, economics, etc.

As an example, an important data set is Tycho Brache's planetary motion data collected in 16th century. This data set leads Kepler's discovery of his three laws of planetary motion and the birth of Newton's mechanics and Calculus. The Newton law of motion is in terms of differential equation. Now-a-day, we have many advance tools to collect data and powerful computer tools to analyze them. It is therefore important to learn the theory of ordinary differential equation, an important tool for mathematical modeling and a basic language of science.

In this course, I will mainly focus on, but not limited to, two important classes of mathematical models by ordinary differential equations:

- population dynamics in biology
- dynamics in classical mechanics.

The first one studies behaviors of population of species. It can also be applied to economics, chemical reactions, etc. The second one include many important examples such as harmonic oscillators, pendulum, Kepler problems, electric circuits, etc. Basic physical laws such as growth laws, conservation laws, etc. for modeling will be introduced.

The goal of this lecture is to guide students to learn

(i) how to do mathematical modeling,

(ii) how to solve the corresponding differential equations,

(iii) how to interpret the solutions, and

(iv) how to develop general theory.

1.2 Relaxation and Equilibria

The most simplest and important example which can be modeled by ODE is a relaxation process. The system starts from some state and eventual reaches an equilibrium state. Such process is called a relaxation process. We use the following two examples to explain this relaxation process.

A falling object A object falling down from hight y_0. Let $v(t)$ be its velocity at time t. According to Newton's law,

$$\frac{dv}{dt} = -g, \tag{1.1}$$

where g is the gravitation constant. Usually the object experiences friction. The sign of the friction force should be opposite to the acceleration. If we treat the object as a particle, then this friction should also depend the speed of the object. The simplest friction model is $-\alpha v$. Adding this frictional force, the complete model becomes

$$\frac{dv}{dt} = -g - \alpha v, \tag{1.2}$$

where α is the frictional coefficient, which may depend on other physical parameters, for instance, the surface area of the object.

Cooling/Heating of an object An object is taken out of refrigerator to defrost. Let $y(t)$ be its temperature at time t. Let the room temperature be K and the initial temperature of the object is y_0. To model the cooling/heating process, we first notice that if the object has temperature K, then there is no variation of the object's temperature with the environment. Thus, the rate change of y is proportional to the difference between $y(t)$ and K. The simplest case is

$$\frac{dy(t)}{dt} = -\alpha(y(t) - K). \tag{1.3}$$

Here, α is a conductivity coefficient. It depends on the object. Sands has larger conductivity than water. This model is indeed called Newton's law of cooling/heating.

As you can see that these two models are mathematically identical. We can use one theory to cover them. This will be discussed below.

Methods and tools to solve the relaxation equation Let us solve the ODE by Calculus as the follows. The technique is called *separation of variables*. In this technique, The terms with same variable are moved to the same side. After that, we can integrate both sides. See the procedure below.

$$\frac{dy}{dt} = -\alpha(y - K).$$

$$\frac{dy}{y - K} = -\alpha dt$$

We integrate both sides to get

$$\log|y - K| = -\alpha t + C.$$

Here, C is an integration constant.

$$|y - K| = e^C \cdot e^{-\alpha t}$$

$$y(t) - K = \pm e^C \cdot e^{-\alpha t}$$

$$y(t) - K = C_1 e^{-\alpha t},$$

where $C_1 = \pm e^C$ is also a constant.

Alternatively, we can interpret the above procedure as a change-of-variable of integration, see below.

$$\frac{1}{y - K}\frac{dy}{dt} = -\alpha$$

We integrate both sides in t, use the change-of-variable $y'(t)dt = dy$ to get

$$\int \frac{y'}{y - K}\, dt = \int \frac{dy}{y - K} = -\int \alpha\, dt$$

$$\log|y - K| = -\alpha t + C.$$

Now, we plug the initial condition: $y(0) = y_0$. We then get $C_1 = y_0 - K$ and

$$y(t) = K + (y_0 - K)e^{-\alpha t}. \tag{1.4}$$

We observe that $y(t) \equiv K$ if $y_0 = K$. Furthermore, $y(t) \to K$ as $t \to \infty$ for any initial datum y_0. We call such K a *stable equilibrium*. In the example of the heating/cooling problem, this means that the temperature $y(t)$ will eventually *relax* to the room temperature K. In the falling object example, the velocity $v(t)$ will approach a termination velocity $K = -g/\alpha$. For any time $0 < t < \infty$, in fact, $y(t)$ is a linear interpolation between y_0 and K. That is,

$$y(t) = e^{-\alpha t}y_0 + (1 - e^{-\alpha t})K.$$

The time to reach half way (i.e. $(y_0 + K)/2$) is a typical time of this relaxation process. We denote it by t_{hf}. It satisfies

$$K + (y_0 - K)e^{-\alpha t_{hf}} = \frac{1}{2}(y_0 + K).$$

$$e^{-\alpha t_{hf}} = \frac{1}{2}.$$

This yields $t_{hf} = \log 2/\alpha$. We thus interpret $1/\alpha$ as the *relaxation time*. The solution $y(t)$ relaxes to its stable equilibrium K at time scale $1/\alpha$.

What we can learn from these two examples is that the ODE model of the form

$$\frac{dy}{dt} = \alpha(K - y)$$

can be used to model a system that tends to a constant state (equilibrium) in $O(1)$ time. Mathematically, the system tends to its equilibrium exponential fast with difference like $e^{-\alpha t}$.

Using mathematical software There are many mathematical software which can solve ODEs. We shall use Maple in this class. Let us type the following commands in Maple. To use the tool of differential equations, we need to include it by typing

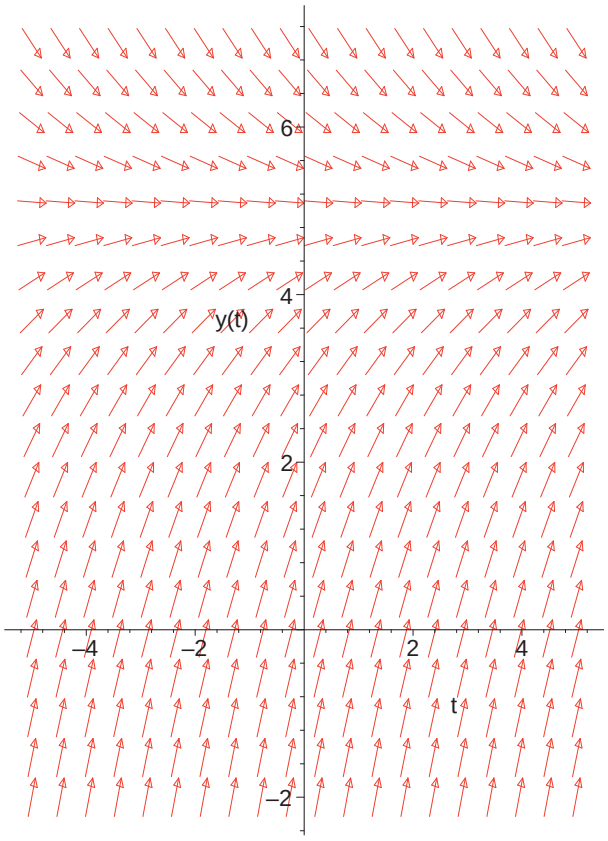
```
> with(DEtools):
```

```
> with(plots):
```

```
> Deq:= diff(y(t),t) = r*(K-y(t));
```

$$Deq := \frac{d}{dt}\,\mathrm{y}(t) = r\,(K - \mathrm{y}(t))$$

```
> dfieldplot(subs(r=0.5,K=5,Deq),y(t),t=-5..5,y=-2..7,arrows=slim):
```



In this example, we plot arrows at every grid points on the t-y plane to represent the vector field $(1, f(y))$. The ODE solution is a curve $(t, y(t))$ on this plane such that its tangent $(1, y'(t))$ is the vector field $(1, f(y(t))$.

Homeworks

1. A dead body is found at 6:30 AM with temperature $18°$. At 7:30 AM, the body temperature is

16°. Suppose the surrounding temperature is 15° and the alive people's temperature is about 37°. Use Newton's cooling/heating law to estimate the dead time.

2. Consider $y' = -a(y-K)^2$ with $a > 0$ and $y(0) = y_0 > K$, find its solution. Will the solution tends to K as $t \to \infty$? At what speed?

3. If the system is $y' = -a\sqrt{|y-K|}$ with $a > 0$ and $y(0) > K$, can the solution approach to K? at finite time? at what speed?

4. $y' = (y-y_0)(y_1-y)$

5. $y' = r(y-y_0)(y-y_1)(y-y_2)$

6. $y' = (y-y_0)^2(y_1-y)$

7. $y' = r\tanh(y)$

1.3 Modeling population dynamics of single species

Simple population growth model Let $y(t)$ be the population (say European population in U.S.) at time t. The census data are from 1790-2000 (every 10 years). We can build a model based on the following hypothesis:

$$\frac{dy}{dt} = \text{ births } - \text{ deaths } + \text{migration.} \tag{1.5}$$

It is natural to assume that the births and the deaths are propotition to the population. Let us neglect the migration for the moment. In terms of mathematical equations, this reads

$$\frac{dy}{dt} = ry \tag{1.6}$$

where r is called the net growth rate, which is the natural growth rate minus the death rate. We should have $r > 0$ if the population is growing. We can set the *initial value*

$$y(0) = y_0, \tag{1.7}$$

the population at year 1790. With (1.6) and (1.7), we can find its solution

$$y(t) = y_0 e^{rt}.$$

We can find the growth rate r by fitting the data, say the census at year 1800. This yields that $r = 0.03067$. We find it fits very well until 1820. From then on, the discrepancy becomes larger and larger. It suggests that

- the growth rate r is treated as a constant is only valid local in time;
- environmental limitation is not taken into account.

Logistic population model The above population model was proposed by Malthus (1766-1834), an economist and a mathematician. One criticism of the simple growth model is that it does not take the limit of environment into consideration. With this consideration, we should expect that there is a environmental carrying capacity K such that

- when $y < K$, the rate $y' > 0$,
- when $y > K$, the rate $y' < 0$.

A simple model with these considerations is the follows:

$$y' = ry\left(1 - \frac{y}{K}\right). \tag{1.8}$$

This is called the *logistic population model.* It was suggested by the Belgien mathematician Pierre Verhulst (1838). It is a nonlinear equation. There is another interpretation for the nonlinear term ry^2/K. Namely, y^2 represents the rate of pair-interaction. The coefficient r/K is the rate of this interaction to the change of y. The minus sign simply means that the pair-interaction decreases the population growth due to a competition of resource.

Exact solutions for the logistic equation We can solve this equation by the method of *separation of variable*.

$$\frac{y'(t)}{y(1-y/K)} = r.$$

Integrating in t yields

$$\int \frac{y'(t)}{y(1-y/K)}\,dt = rt + C.$$

By change-variable formula for integration, we have

$$\int \frac{1}{y(1-y/K)}\,dy = rt + C.$$

This yields

$$\int \left(\frac{1}{y} + \frac{1}{K-y}\right)\,dy = rt + C$$

$$\log\left|\frac{y}{K-y}\right| = rt + C.$$

$$\left|\frac{y}{K-y}\right| = \frac{1}{C_1 e^{-rt}}.$$

Here $C_1 = e^{-C}$ is another constant. When $0 < y < K$, we get

$$\frac{y}{K-y} = \frac{1}{C_1 e^{-rt}}.$$

This yields

$$y = \frac{K}{1 + C_1 e^{-rt}}.$$

When $y < 0$ or $y > K$, we get

$$\frac{y}{K - y} = -\frac{1}{C_1 e^{-rt}}.$$

This gives

$$y = \frac{K}{1 - C_1 e^{-rt}}.$$

When $t = 0$, we require $y(0) = y_0$. We find that in both cases, $C_1 = |1 - K/y_0|$. Thus, the solution is

$$y(t) = \begin{cases} \frac{K}{1-C_1 e^{-rt}} & y_0 < 0 \text{ or } y_0 > K \\ \frac{K}{1+C_1 e^{-rt}} & 0 < y_0 < K \end{cases}$$

and $y(t) \equiv 0$ if $y(0) = 0$, $y(t) \equiv K$ if $y(0) = K$.

Remarks.

1. We observe that

   - for initial y_0 with $y_0 > 0$, we have $y(t) \to K$;
   - the states $y \equiv 0$ and $y(t) \equiv K$ are constant solutions.

   These constant solutions are called the *equilibrium states*. Any solution with initial state near K will approach to K as t tends to infinity. We call K a *stable* equilibrium. On the other hand, if the initial state is a small perturbation of the 0 state, it will leave off the zero state and never come back. We call 0 an *unstable* equilibrium.

2. When $y_0 < 0$, the corresponding $C_1 > 1$, we observe that the solution $y(t) \to -\infty$ as $t \to t_1^*-$, where
   $$1 - C_1 e^{-rt_1^*} = 0.$$
   We call the solution blows up at finite time. This solution has no ecological meaning.

3. When $y_0 > K$, the corresponding $0 < C_1 < 1$ and $y(t) \to +\infty$ as $t \to t_2^*+$, where $1 - C_1 e^{-rt_2^*} = 0$.

Qualitative analysis for the logistic equation We can analyze the properties (equilibrium, stability, asymptotic behaviors) of solutions of the logistic equation by the phase portrait analysis. First, let us notice two important facts:

- For any point (t_0, y_0), there is a solution $y(\cdot)$ passing through (t_0, y_0). In other words, $y(t_0) = y_0$.
- No more than one solution can pass through (t_0, y_0).

These are the existence and uniqueness theorems of the ODE. Let us accept this fact for the moment. Next, we can use the equilibria to classify our general solutions.

- The first step is to find all equilibria of this system. Let us denote the right-hand side of (1.8) by $f(y)$, i.e.
$$f(y) = ry\left(1 - \frac{y}{K}\right).$$
An equilibrium is a constant solution $y(t) \equiv \bar{y}$, where $f(\bar{y}) = 0$. In our case, the equilibria are $y(t) \equiv 0$ and $y(t) \equiv K$.

- The second step is to classify all other solutions. On the t-y plane, we first draw the above two constant solutions. Now, by the uniqueness theorem, no solution can pass through these two constant solution. Therefore, the y-space (it is one dimension in the present case) is naturally partitioned into three regions:
$$I_1 = (-\infty, 0),\ I_2 = (0, K),\ I_3 = (K, \infty).$$
If $y(0) \in I_\ell$, then the corresponding $y(t)$ stays in I_ℓ for all t.

- The third step is to characterize all solutions in each regions. For any solution in I_2, we claim that $y(t) \to K$ as $t \to \infty$. From $f(y) > 0$ in I_2, we can conclude that $y(\cdot)$ is strictly increasing in I_2. We claim that $y(t) \to K$ as $t \to \infty$ for any solution in region I_2. Indeed, $y(t)$ is increasing and has an upper bound K. By the monotone convergence property of $\mathbb{R}$, $y(t)$ has a limit as t tends to infinity. Let us call this limit $\bar{y}$. We claim that $\bar{y} = K$. If not, $\bar{y}$ must be in $(0, K)$ and hence $f(\bar{y}) > 0$. By the continuity of f, there must be an $\epsilon > 0$ and a neighborhood $(\tilde{y}, \bar{y})$ such that $f(y) > \epsilon$ for all $y \in [\tilde{y}, \bar{y})$. Since $\lim_{t\to\infty} y(t) = \bar{y}$ monotonically, there must be a t_0 such that $\tilde{y} \le y(t) < \bar{y}$ for $t \ge t_0$. In this region, the corresponding $y'(t) = f(y(t)) \ge \epsilon$. Hence $y(t) \ge y(t_0) + \epsilon(t - t_0)$ for all $t \ge t_0$. This contradicts to $y(t)$ being bounded. Hence, we get $y(t) \to K$ as $t \to \infty$. Similarly, for solution $y(\cdot) \in I_3$, $y(t) \to K$ as $t \to \infty$.

- Using the same argument, we can show that for solution in $I_1 \cup I_2$, $y(t) \to 0$ as $t \to -\infty$. This means that 0 is *unstable*. Indeed, for $y(0) < 0$, we have $f(y) < 0$. This implies $y(\cdot)$ is decreasing for $t > 0$. If $y(t)$ has a low bound, then $y(t)$ will have a limit and this limit $\bar{y} < 0$ and must be a zero of f. This is a contradiction. Hence $y(t)$ has no low bound.

To summarize, we have the following theorem.

Theorem 1.1. *All solutions of (1.8) are classified into the follows.*

1. *equilibria:* $y(t) \equiv 0$ *and* $y(t) \equiv K$*;*
2. *If* $y(0) \in I_1 \cup I_2$*, then* $\lim_{t\to-\infty} y(t) = 0$*;*
3. *If* $y(0) \in I_2 \cup I_3$*, then* $\lim_{t\to\infty} y(t) = K$.

The biological interpretation is the follows.

- If $y(0) < K$, then $y(t)$ will increase to a saturated population K as $t \to \infty$.

- If $y(0) > K$, , then $y(t)$ will decrease to the saturated population K as $t \to \infty$.

- $y(t) \equiv K$ is the stable equilibrium, whereas $y(t) \equiv 0$ is an unstable equilibrium.

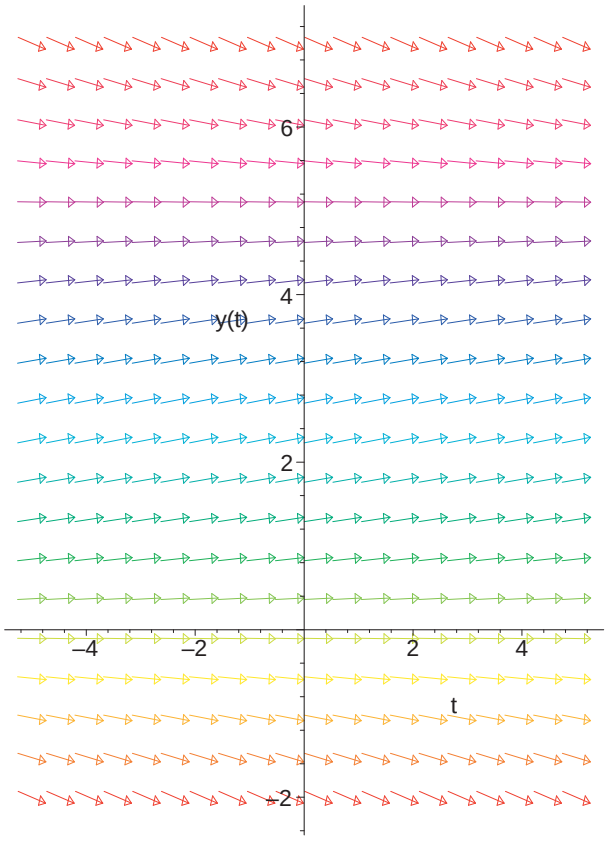
Maple Practice Below, we demonstrate some Maple commands to learn how to solve plot the solutions.

```
> with(plots):
> with(DEtools):
```

```
> DiffEq := diff(y(t),t)=r*y(t)*(1-y(t)/K);
```

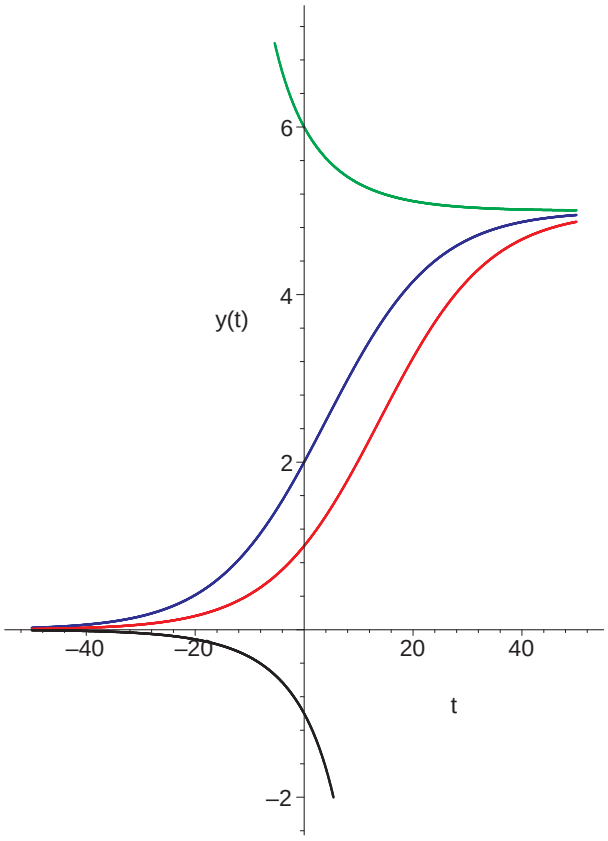
$$DiffEq := \frac{d}{dt}\,\mathrm{y}(t) = r\,\mathrm{y}(t)\,(1 - \frac{\mathrm{y}(t)}{K})$$

```
> dfieldplot(subs(r=0.1,K=5,DiffEq),y(t),t=-5..5,y=-2..7,arrows=slim,co
> lor=y/7);
```


6
4
y(t)
2
−4
−2
2
4
t
−2

```
> fig1 := DEplot(subs(r=0.1,K=5,DiffEq),y(t),
> t=-50..50,[[y(0)=1]],y=-2..7,stepsize=.05,arrows=none,linecolour=red):
> fig2 := DEplot(subs(r=0.1,K=5,DiffEq),y(t),
> t=-50..50,[[y(0)=2]],y=-2..7,stepsize=.05,arrows=none,linecolour=blue)
> :
> fig3 := DEplot(subs(r=0.1,K=5,DiffEq),y(t),
> t=-50..50,[[y(0)=6]],y=-2..7,stepsize=.05,arrows=none,linecolour=green
> ):
> fig4 := DEplot(subs(r=0.1,K=5,DiffEq),y(t),
> t=-50..50,[[y(0)=-1]],y=-2..7,stepsize=.05,arrows=none,linecolour=blac
> k):
```

```
> display({fig1,fig2,fig3,fig4});
```



Logistic population model with harvesting Suppose migration is considered. Let e be the migration rate. We should modify the model by

$$y' = ry\left(1 - \frac{y}{K}\right) - ey. \tag{1.9}$$

The migration rate e can be positive (migrate out) or negative (migrate in).

This model is often accepted in ecology for harvesting a renewable resources such as shrimps, fishes, plants, etc. In this case, $e > 0$ is the harvesting rate which measures the harvesting effort. The quantity ey is the amount of harvesting per unit time. It is called the harvesting yield per unit time.

This harvesting model is still a logistic equation

$$y' = (r-e)y\left(1 - \frac{ry}{(r-e)K}\right) \tag{1.10}$$

with new growth rate $r - e$. The new equilibrium is

$$K_h = K\left(1 - \frac{e}{r}\right),$$

which is the sustained population. Two cases:

- When $e < r$, we have $0 < K_h < K$. This means that the saturated population K_h decreases due to harvesting.
- When $e > r$, then the species will be extinct due to over-harvesting. Indeed, you can check that $y(t) \equiv 0$ is the stable equilibrium and $y(t) \equiv K_h$ is the unstable equilibrium now.

The quantity $Y(e) = eK_h(e)$ is called the sustained harvesting yield. It is the maximal amount of harvesting from this environment if the harvest yield rate is e. An ecological goal is to maximize this sustained harvesting yield at minimal harvesting effort. That is, $\max_e Y(e)$. We see that the maximum occurs at $e = r/2$. The corresponding sustained harvesting yield is

$$Y\left(\frac{r}{2}\right) = \frac{r}{2}\frac{K}{2} = \frac{rK}{4}.$$

There is another way to model harvesting of natural resources. We may use harvesting amount C instead of the harvesting rate e as our parameter. The model now reads

$$y' = ry\left(1 - \frac{y}{K}\right) - C := f_C(y). \tag{1.11}$$

The equilibrium (i.e. $f_C(y) = 0$) occurs at $f_C(y) = 0$. On the C-y plane, $f_C(y) = 0$ is a parabola. For $C \le rK/4$, there are two solutions for $f_C(y) = 0$:

$$y_\pm = \frac{K}{2} \pm \sqrt{\frac{K^2}{4} - \frac{CK}{r}}.$$

For $C > rK/4$, there is no real solution. For $C < rK/4$, we can draw arrows on the intervals $(-\infty, y_-)$, (y_-, y_+), (y_+, ∞) to indicate the sign of f_C in that interval. We conclude that y_+ is a stable equilibrium. We rename it K_h.

To have sustained resource, we need $K_h > 0$. That is,

$$K_h := \frac{K}{2} + \sqrt{\frac{K^2}{4} - \frac{CK}{r}} > 0.$$

So the maximal harvesting to maintain $K_h > 0$ is

$$C = \frac{rK}{4}.$$

For $C > rK/4$, $y(t) \to 0$ as t increases to some t^*.

The solution for $y' = ry(1 - \frac{y}{K}) - C$ with $y(0) = y_0$ is

$$y(t) = \frac{1}{2}\left(K + \frac{\Delta}{r}\tanh(\frac{\Delta}{2K}(t + C_0))\right)$$

where

$$\Delta = \sqrt{rK(rK - 4C)},\ C_0 = \frac{2K}{\Delta}\text{arctanh}(\frac{r}{\Delta}(2y_0 - K)).$$

In additional to the constraint $C \le rK/4$, we should also require $y(0) > 0$. Otherwise, there would be no harvesting at all. This would give another constraint on C. You may find it by yourself.

Homeworks

1. Solve the Gompertz equation for population growth

   $$y' = ry\ln(K/y).$$

   What are the equilibria? What are the asymptotic behaviors.

2. Solve the equation

   $$y' = \alpha y(1 - y^2).$$

   and discuss stability of its equilibria and the asymptotic behaviors of the solution at large time.

Abstract logistic population models We can use the following abstract model

$$y' = f(y) \tag{1.12}$$

to study the issues of equilibria and their stability. Here, the function f depends on y only. Such systems with f being independent of t are called *autonomous* systems. We consider the initial datum

$$y(0) = y_0. \tag{1.13}$$

Following the example of the logistic model, let us assume that $f(y)$ has the following qualitative properties:

- $f(y_0) = f(y_1) = 0$,
- $f(y) > 0$ for $y_0 < y < y_1$,
- $f(y) < 0$ for $y < y_0$, or $y > y_1$,

First, there are two equilibrium solutions:

$$y(t) \equiv y_0,\ y(t) \equiv y_1.$$

For general solutions, we integrate the equation

$$\frac{dy}{f(y)} = dt.$$

On the right, we integrate in t from 0 to t. On the left, with the change of variable: $t \to y(t)$, we get the integration domain of y is from y_0 to y, and We arrive at

$$\Phi(y) - \Phi(y_0) = t$$

where $\Phi(y) = \int dy/f(y)$. From the properties of f, we obtain that

$$\Phi(y) : \begin{cases} \text{decreasing}, & \text{for } y > y_1, y < y_0 \\ \text{increasing}, & \text{for } y_0 < y < y_1, \end{cases}$$

Therefore, the function is invertible in each of the three regions: $(-\infty, y_0)$, (y_0, y_1), and (y_1, ∞). The solution $y(t)$ with initial datum is precisely the inversion of Φ with $\Phi(y_0) = 0$.

A bistable model We consider the autonomous equation

$$y' = f(y),$$

where $f(y)$ has three zeros $y_1 < y_2 < y_3$. Assume the sign of f is $f(y) > 0$ for $y < y_1, y_2 < y < y_3$, and $f(y) > 0$ for $y_1 < y < y_2, y > y_3$. In this case, for $y(t)$ with initial data $y(0)$ satisfying $y(0) < y_2$, we have $y(t) \to y_1$ as $t \to \infty$. If $y(0) > y_1$, then $y(t) \to y_3$ as $t \to \infty$. The states y_1 and y_3 are the two stable states. Such a model is called a bistable model. It is usually used to model phase field of some material. A simple model is $f(y) = y(1-y)(1/2-y)$.

Maple tool: phase line analysis Use Maple to draw the function $f(y)$. The y-axis is partition into regions where $f(y) > 0$ or $f(y) < 0$. Those y^* such that $f(y^*) = 0$ are the equilibria. An equilibrium y^* is stable if f is increasing near y^* and unstable if f is decreasing there.

Asymptotic behaviors and convergent rates Let us focus to an autonomous system which has only one equilibrium, say $\bar{y} = 0$. That is, the rate function $f(0) = 0$. Let us consider two cases: $f(y) = -\alpha y$ and $f(y) = -\beta y^2$ with $y(0) > 0$. We need minus to have $y \equiv 0$ a stable equilibrium.

- Case 1: $y' = f(y) = -\alpha y$. In this case, we have seen that the solution is

  $$y(t) = y(0)e^{-\alpha t}$$

  We see that the solution tends to its equilibrium 0 exponentially fast. The physical meaning of $1/\alpha$ is the time that the difference of solution from its equilibrium is reduced by a fixed factor (e^{-1}). We say the convergent rate to its equilibrium to be $O(e^{-\alpha t})$.

- Case 2: $y' = f(y) = -\beta y^2$. In this case,

$$y(t) = \frac{1}{1/y(0) + \beta t}.$$

We observe that $y(t) \to 0$ as $t \to \infty$ with rate $O(1/t)$.

Homeworks

1. Construct an ODE so that $y(t) = (1 + t)$ is its asymptotic solution with convergent rate e^{-2t}.
2. Construct an ODE so that $y(t) = (1 + t)$ is its asymptotic solution with convergent rate t^{-1}.
3. Search for "bistability" in Wikipedia

1.4 Techniques to solve single first-order equations

1.4.1 Linear first-order equation

The linear first-order equation has the form:

$$y' = a(t)y + b(t). \tag{1.14}$$

The term $b(t)$ is called the source term, or the inhomogeneous term, whereas the part

$$y' = a(t)y$$

is called its homogeneous part. We first solve the homogeneous equation. We separate t and y and get

$$\frac{y'}{y} = a(t).$$

The left-hand side (LHS) is $d \log y(t)/dt$. We integrate it and get

$$\int \frac{d \log y(t)}{dt}\, dt = \int a(t)\, dt.$$

This yields

$$\log y(t) = A(t) + C_1, \text{ or } y(t) = Ce^{A(t)},$$

where $A'(t) = a(t)$, and C or C_1 is a constant. We may choose $A(0) = 0$. That is, $A(t) = \int_0^t a(s)\, ds$. The constant C is y_0 if we require $y(0) = y_0$. We conclude that the solution is

$$y(t) = y(0)e^{\int_0^t a(s)\, ds}.$$

Next, we study the inhomogeneous equation. We will introduce two methods.

Method of Variation of Constant We guess our solution to have the form

$$y(t) = C(t)e^{A(t)}.$$

Plugging it into (1.14), we obtain

$$C'(t)e^{A(t)} + a(t)C(t)e^{A(t)} = a(t)C(t)e^{A(t)} + b(t).$$

This yields

$$C'(t) = b(t)e^{-A(t)}.$$

Hence the solution is

$$C(t) = C(0) + \int_0^t b(s)e^{-A(s)}\,ds.$$

By plugging the initial datum, we obtain $C(0) = y(0)$. Hence, the general solution is given by

$$y(t) = y(0)e^{A(t)} + \int_0^t b(s)e^{-A(s)+A(t)}\,ds.$$

The idea behind the variation of constant is that the ansatz (i.e. the solution form that we guess)

$$y(t) = C(t)e^{A(t)}$$

has the property:

$$y'(t) = C(t)A'(t)e^{A(t)} + C'(t)e^{A(t)}.$$

In a short time, if C remains nearly unchanged, $e^{A(t)}$ behaves like solutions of $y' = A'(t)y$. By allowing $C(t)$ varying, the $C'(t)$ term can take care contribution of the source $b(t)$ pumping into the system.

It is important to notice that the integrand $b(s)e^{A(t)-A(s)}$ is the solution of $y' = a(t)y$ for $s < t$ with $y(s) = b(s)$. This means that the source term $b(s)$ generates a solution $b(s)e^{A(t)-A(s)}$ at time s. The total contribution of the source term from time 0 to t is the accumulation of these solutions, i.e. $\int_0^t b(s)e^{A(t)-A(s)}\,ds$. This is called the Duhamel principle. As a real world example, you imagine a tank with nutrition and bacteria. Suppose the growth of the bacteria satisfies $y' = 2y$, i.e. the population doubles per unit time. Suppose we also fill in $b(s)$ amount of bacteria at time s for $0 < s < t$. The bacteria $b(s)$ entering the tank at time s grows to $b(s)e^{2(t-s)}$ at time t. Thus, the total amount of the bacteria population filled-in and grow from time 0 to t is

$$\int_0^t b(s)e^{2(t-s)}\,ds.$$

This together with the grow of the initial population $y(0)$ give the total population at time t to be

$$y(0)e^{2t} + \int_0^t b(s)e^{2(t-s)}\,ds.$$

Method of Integration Factor Alternatively, we may multiply the equation

$$y' - a(s)y = b(s)$$

by $e^{-A(s)}$ on both sides. Then the left-hand side is

$$e^{-A(s)}(y' - a(s)y) = e^{-A(s)}(y' - A'(s)y) = \frac{d}{ds}\left(e^{-A(s)}y(s)\right).$$

Thus, we get

$$\frac{d}{ds}\left(e^{-A(s)}y(s)\right) = e^{-A(s)}b(s).$$

Since the left-hand side is a total differential, we can integrate it once to get

$$e^{-A(t)}y(t) - e^{-A(0)}y(0) = \int_0^t \frac{d}{ds}\left(e^{-A(s)}y(s)\right)\, ds = \int_0^t e^{-A(s)}b(s)\, ds.$$

From $A(0) = 0$, we get

$$e^{-A(t)}y(t) = y(0) + \int_0^t e^{-A(s)}b(s)\, ds.$$

Hence

$$y(t) = e^{A(t)}y(0) + \int_0^t e^{A(t)-A(s)}b(s)\, ds.$$

The quantity $e^{-A(s)}$ is used to make the left-hand side of $y - a(s)y$ to be a total differential. This quantity is called an integration factor. We shall give thorough discussion in the next section.

Example. Consider

$$y' + \frac{2}{t}y = t - 1.$$

Let

$$A(t) = -\int \frac{2\, dt}{t} = \ln t^{-2}$$

and $e^{-A(t)} = t^2$. By multiplying $e^{-A(t)}$ on both sides, we obtain

$$t^2y' + 2ty = \frac{d}{dt}(t^2y) = t^2(t-1).$$

Integrating in t, we get

$$t^2y = \frac{t^4}{4} - \frac{t^3}{3} + C.$$

Hence,

$$y(t) = \frac{t^2}{4} - \frac{t}{3} + \frac{C}{t^2}.$$

Homeworks Solve the following equations

1. $y' + y = te^{-t}$
2. $y' + 2ty = te^{-t^2}$
3. $t^3y' + 4t^2y = e^{-t}$, $y(-1) = 0$, $y < 0$.
4. Consider the initial value problem

$$y' + y = 1 + \cos 2t,\ y(0) = 0.$$

   Find the solution and discuss the solution behavior for large t.

5. Find the solutions of

$$y' + ay = be^{-\lambda t}$$

   and discuss their large time behaviors.

1.4.2 Separation of variables

We can write the ODE

$$\frac{dy}{dt} = f(t, y)$$

in this differential form:

$$dy - f(t, y)dt = 0.$$

A general differential 1-form looks like

$$M(t, y)dt + N(t, y)dy = 0. \tag{1.15}$$

This is called the Pfaffian equation. Suppose the functions $M(t, y)$ and $N(t, y)$ in (1.15) are separable, that is

$$\begin{aligned} M(t, y) &= f_1(t)f_2(y), \\ N(t, y) &= g_1(t)g_2(y), \end{aligned}$$

Dividing (1.15) by $f_2(y)g_1(t)$, then the Pfaffian equation (1.15) becomes

$$\frac{f_1(t)}{g_1(t)}\,dt + \frac{g_2(y)}{f_2(y)}dy = 0.$$

We can integrate it to obtain an integral ϕ:

$$\psi(t, y) := \int \frac{f_1(t)}{g_1(t)}\,dt + \int \frac{g_2(y)}{f_2(y)}dy.$$

Then $\psi(t, y) = constant$ defines a solution implicitly. In this example, $1/(f_2(y)g_1(t))$ is called an integration factor, which makes the Pfaffian equation become a total differential and thus integrable.

Homework

1. $y' = t/y^2$. Ans: $y^3/3 = t^2/2 + C$, or
$$y(t) = \left(\frac{3t^2}{2} + k\right)^{1/3}.$$
2. $(x^2+1)(y^2-1)\,dx + xy\,dy = 0$. The answer is
$$y^2 = 1 + C\frac{e^{-x^2}}{x^2}.$$
3. $y' = t^2/(1-y^2)$. Ans.: $-t^3 + 3y - y^3 = const.$
4. $y' = (4x - x^3)/(4 + y^3)$. Ans. $y^4 + 16y + x^4 - 8x^2 = const.$
5. $y' = \frac{3x^2+4x+2}{2(y-1)}$. Ans. $y^2 - 2y = x^3 + 2x^2 + 2x + 3$.
6. $y' = \frac{y(-1/2+x/4)}{x(2-y/2)}$. This equation is arisen from predator-prey model.

1.4.3 Other special classes that are solvable

Bernoulli equation Bernoulli equation has the form
$$y' = a(t)y + b(t)y^n \tag{1.16}$$
Divide both sides by y^{-n}, we obtain
$$y^{-n}y' = a(t)y^{-n+1} + b(t).$$
Or
$$\frac{1}{1-n}\left(y^{1-n}\right)' = a(t)y^{1-n} + b(t)$$
This suggests the following change of variable:
$$z = y^{1-n}.$$
Then
$$z' = (1-n)a(t)z + (1-n)b(t) \tag{1.17}$$
which can be solved.

Homeworks (Courant and John, Vol. II, pp. 690) Solve the following equations

1. $xy' + y = y^2 \log x$
2. $xy^2(xy' + y) = a^2$
3. $(1-x^2)y' - xy = axy^2$.

*** Riccati equation** (Courant and John, Vol. II, pp. 690) The Riccati equation reads

$$y' = P(x)y^2 + Q(x)y + R(x) \tag{1.18}$$

It can be transformed into a linear equation if we know a particular solution $y = y_1(x)$. We introduce the new unknown

$$u = \frac{1}{y - y_1}.$$

*** Homeworks** Courant and John, Vol. II, pp. 690, Exercises 4–8. Use the above substitution to solve the following equations.

1. $y' - x^2y^2 + x^4 - 1 = 0$ that possesses a particular solution $y_1 = x$.
2. Show that if two solutions, $y_1(x)$ and $y_2(x)$, of Riccati's equation are known, then the general solution is given by

$$y - y_1 = c(y - y_2)\exp[\int P(y_2 - y_1)\,dx],$$

   where c is an arbitrary constant.
3. Find the general solution of

$$y' - y\tan x = y^2\cos x - \frac{1}{\cos x},$$

   which has solutions of the form $a\cos^n x$.

1.5 Vector Fields and Family of Curves

In this section, I will provide geometric view of first order equation, which is, family of plane curves. They can be generated by vector fields or by first order ODEs.

1.5.1 Vector Fields

A vector field $\mathbf{V}(x, y) = (u(x, y), v(x, y))$ on a domain $\Omega \subset \mathbb{R}^2$ is a mapping $\Omega \to \mathbb{R}^2$. For instance, $\mathbf{V}(x, y) = (-y, x)$ is a vector fields on the plane, while $\mathbf{V}(x, y) = (-y, x)/(x^2 + y^2)$ is a vector field defined on $\mathbb{R}^2 \setminus \{\mathbf{0}\}$. A curve $(x(\tau), y(\tau))$ with parameter $\tau \in (a, b)$ is called an integral curve of the vector field $\mathbf{V} = (u(x, y), v(x, y))$ if its tangent is parallel to the vector field $\mathbf{V}$ along this curve, that is,

$$\begin{bmatrix} dx/d\tau \\ dy/d\tau \end{bmatrix} \parallel \begin{bmatrix} u(x(\tau), y(\tau)) \\ v(x(\tau), y(\tau)) \end{bmatrix} \text{ for all } \tau.$$

From this definition, the integral curves of the two vector fields $\mathbf{V} = (-y, x)$ and $\mathbf{V} = (-y, x)/(x^2 + y^2)$ are identical because they are parallel to each other at every points on $\mathbb{R}^2 \setminus \{\mathbf{0}\}$.

A general single first-order differential equation $y'(x) = f(x, y)$ induces a vector field $\mathbf{V}(x, y) := (1, f(x, y))$ on the plane. Conversely, given vector field $\mathbf{V}(x, y) = (u(x, y), v(x, y))$ on the plane, a curve $\{(x, y(x))|x \in I\}$ is an integral curve of $\mathbf{V}$ if its tangent

$$\begin{bmatrix} 1 \\ dy/dx \end{bmatrix} \parallel \begin{bmatrix} u(x, y) \\ v(x, y) \end{bmatrix}.$$

This is equivalent to

$$\frac{dy}{dx} = \frac{v(x, y)}{u(x, y)}, \tag{1.19}$$

provided $u(x, y) \neq 0$. Thus the integral curves of $\mathbf{V}(x, y) = (u(x, y), v(x, y))$ satisfy the single first-order ODE (1.19).

Integral of a vector field A function $\psi : \Omega \to \mathbb{R}$ is called an integral of the vector field $\mathbf{V}(x, y)$ if its *level sets* $\{(x, y)|\psi(x, y) = C\}$ for all C, are integral curves of $\mathbf{V}(x, y)$. This definition is equivalent to

$$\nabla\psi \cdot \mathbf{V} = 0. \tag{1.20}$$

This is because along any integral curve $(x(\tau), y(\tau))$, we have $\psi(x(\tau), y(\tau)) = C$. We differentiate it in τ and using $(\dot{x}, \dot{y}) \parallel \mathbf{V}$ to obtain (1.20). Conversely, if $(x(\cdot), y(\cdot))$ is an integral curve, then along this curve, we have $(\dot{x}, \dot{y}) = \sigma\mathbf{V}$ for some scalar function σ. Thus,

$$\frac{d}{d\tau}\psi(x(\tau), y(\tau)) = \psi_x\dot{x} + \psi_y\dot{y} = \nabla\psi \cdot (\sigma\mathbf{V}) = 0.$$

Thus, $\psi(x(\cdot), y(\cdot))$ is a constant.

Example Consider the vector $\mathbf{V} = (-y, x)/r^2$, where $r^2 = x^2 + y^2$. Its integral curves satisfy

$$(dx, dy) \parallel (-y, x).$$

Its ODE form is

$$y' = -x/y.$$

Using the technique of separation of variable, we get

$$xdx + ydy = 0,$$

Integrating it, we obtain

$$x^2 + y^2 = C.$$

Thus, the function $\psi(x, y) = x^2 + y^2$ is an integral of the vector field $(-y, x)/r^2$.

Integration factor To find an integral ψ of a vector field $\mathbf{V}(x,y) = (u(x,y), v(x,y))$ in a domain $\Omega \subset \mathbb{R}^2$, from (1.20), we see that this is equivalent to

$$\nabla\psi \parallel (-v, u),$$

or

$$\nabla\psi = \mu(-v, u)$$

for some function $\mu(x,y)$. Such function μ is called an *integration factor*. Let us denote $-\mu v$ by P and μu by Q. With μ, then we can obtain ψ by the line integral

$$\psi(x,y) := \int_{(x_0,y_0)}^{(x,y)} P(x,y)dx + Q(x,y)dy$$

along any path from (x_0, y_0) to (x, y) in Ω. This line integral should be independent of paths, otherwise, ψ is not well-defined. But this is equivalent to that the line integral is 0 along any closed curve. Let us choose the closed curve to be the boundary of an arbitrary simply connected domain $D \subset \Omega$. From Green's theorem

$$\int_{\partial D} P(x,y)dx + Q(x,y)dy = \int_D (Q_x - P_y)\, dx\, dy = 0.$$

Since D is arbitrary, we get

$$Q_x - P_y \equiv 0 \text{ in } \Omega.$$

This is

$$(\mu u)_x + (\mu v)_y = 0 \text{ in } \Omega.$$

This is a partial differential equation (PDE) for the integration factor μ. There are some PDE technique (called method of characteristics) to find μ, at least, *locally*. This means that under very mind condition, any vector field is locally integrable up to an integration factor. However, we shall not go into this topics here. We shall just give some examples instead. Once μ is obtained, we obtain an integral ψ of $\mathbf{V}$.

Examples

1. Consider the linear equation

$$y' = 2y + t. \tag{1.21}$$

We claim that $\mu = e^{-2t}$ is an integration factor. In fact, the equation can be rewritten as

$$dy - 2ydt = tdt.$$

We multiply both sides by $\mu = e^{-2t}$ to get

$$e^{-2t}(dy - 2ydt) = te^{-2t}\, dt \tag{1.22}$$

The left-hand side (LHS) is a total differential:

$$e^{-2t}(dy - 2ydt) = d(e^{-2t}y)$$

The right-hand side (RHS) is also a total differential:

$$te^{-2t}\,dt = d\int te^{-2t}\,dt$$

and

$$\int te^{-2t}\,dt = -\frac{1}{2}\int tde^{-2t} = -\frac{1}{2}te^{-2t} + \frac{1}{2}\int e^{-2t}\,dt = -\frac{1}{2}te^{-2t} - \frac{1}{4}e^{-2t} + C.$$

Hence, (1.22) can be expressed as

$$d\left(e^{-2t}y + \frac{1}{2}te^{-2t} + \frac{1}{4}e^{-2t}\right) = 0.$$

Thus, $\psi := e^{-2t}y + \frac{1}{2}te^{-2t} + \frac{1}{4}e^{-2t}$ an integral of (1.21).

2. In the linear equation (1.14)

$$y' = a(t)y + b(t),$$

we multiply (1.14) by $\mu(t) = e^{-A(t)}$ where $A'(t) = a(t)$, we obtain

$$e^{-A(t)}y' - A'(t)e^{-A(t)}y = e^{-A(t)}b(t)$$

$$\frac{d}{dt}\left(e^{-A(t)}y\right) = e^{-A(t)}b(t).$$

We can then integrate this formula in t to obtain the solution for (1.14). In this method, $\mu = e^{-A(t)}$ is an integration factor and

$$\psi = e^{-A(t)}y - \int e^{-A(t)}b(t)\,dt$$

is an integral.

Notice that the integration factor and the integral are not unique. Suppose ψ is an integral and μ is the corresponding integration factor. Consider a composition function

$$\phi(x, y) := h(\psi(x, y)),$$

where $h(\cdot) : \mathbb{R} \to \mathbb{R}$ is any smooth function with $h' \neq 0$. Then

$$d\phi = h'd\psi = h'\mu\,(-vdx + udy) = 0.$$

Hence, ϕ is another integral with a new integration factor $h'(\psi(x, y))\mu(x, y)$.

Certainly, if both ϕ and ψ are integrals of (1.15), which means that their level sets represent the same integral curves. Thus, there is an one-to-one correspondence between the level sets of ψ and ϕ:

$$\psi(x, y) = C_1 \text{ if and only if } \phi(x, y) = C_2.$$

Two functions ϕ and ψ with this property is called *functional dependent*. If we define a function h which maps:$C_1 \mapsto C_2$, then $\phi(x, y) = h(\psi(x, y))$. Thus, any two integrals of $\mathbf{V}$ are functional dependent.

Stream functions of velocity fields In fluid mechanics, $\mathbf{V}(x, y)$ is the velocity field, while its integral $\psi(x, y)$ represents stream function, and the level sets $\psi(x, y) = C$ are the stream lines.

Find the stream functions of the velocity field

$$\mathbf{V} = (u, v) = \left(\frac{y^2 - x^2}{(x^2 + y^2)^2}, \frac{2xy}{(x^2 + y^2)^2} \right).$$

This is the velocity field of a potential flow around a unit circular cylinder. In this example, the ODE the stream function satisfies

$$\frac{dy}{dx} = \frac{y^2 - x^2}{2xy}.$$

Let us define a homogeneous variable: $\eta = y/x$. We use x and η as the new variables. We have $dy = d(x\eta) = \eta dx + x d\eta$, or $dy/dx = \eta + x d\eta/dx$. Plug this into the equation, we get

$$\eta + x\frac{d\eta}{dx} = \frac{y^2 - x^2}{2xy} = \frac{\frac{y^2}{x^2} - 1}{2\frac{y}{x}} = \frac{\eta^2 - 1}{2\eta}.$$

$$x\frac{d\eta}{dx} = \frac{\eta^2 - 1}{2\eta} - \eta = -\frac{1 + \eta^2}{2\eta}.$$

Separating variables, we get

$$\frac{2\eta}{1 + \eta^2} d\eta + \frac{dx}{x} = 0.$$

Integrating this,

$$\ln(1 + \eta^2) + \ln|x| = C$$

The level sets are

$$|x| + \frac{y^2}{|x|} = C.$$

In this example, we introduces an important technique for ODE of the form:

$$\frac{dy}{dx} = \frac{v(x, y)}{u(x, y)},$$

where u and v are homogeneous functions of degree n. Following Leibnitz's method, we define a homogeneous variable $\eta = y/x$. We use x and η as our new variables. We have $dy = d(x\eta) = x\, d\eta + \eta\, dx$. From homogeneity, we have $u(x, x\eta) = x^n u(1, \eta)$ and $v(x, x\eta) = x^n v(1, \eta)$. The equation becomes

$$(-v(1, \eta) + \eta u(1, \eta))\, dx + x u(1, \eta) d\eta = 0.$$

We can use method of separation of variables:

$$\frac{d\eta}{R(\eta)} + \frac{dx}{x} = 0,$$

where

$$R(\eta) = \eta - \frac{v(1, \eta)}{u(1, \eta)}.$$

The solution is

$$\int \frac{d\eta}{R(\eta)} + \log|x| = C.$$

Remark Vector fields are arisen in natural world, for examples, fluid mechanics and electromagnetism. In two dimensional incompressible flow, the velocity $\mathbf{V}(x,y) = (u(x,y), v(x,y))$ satisfies the incompressibility condition

$$\nabla \cdot \mathbf{V} = 0.$$

If in addition,

$$\nabla \times \mathbf{V} = 0,$$

such flows are called irrotational flows. An incompressible and irrotational flow is called a potential flow. For two dimensional potential flow, the functions

$$\phi(x,y) := -\int^{(x,y)} udx + vdy,\ \psi(x,y) := \int^{(x,y)} -vdx + udy$$

are called its velocity potential and stream function. They are orthogonal to each other.

Homeworks : Solve the equations

1. $y' = \frac{x+y}{x-y}$. Ans. $\arctan v - \frac{1}{2}\log(1+v^2) = \log|x| + C$.
2. $y' = \frac{x^2+xy+y^2}{x^2}$.
3. $y' = -\frac{x^2}{x^2+y^2}$

1.5.2 Family of curves and Orthogonal trajectories

We have seen that general solutions of a first-order ODE form a family of curves on the plane. Conversely, we will show that a family of curves on the plane satisfy a first-order ODE. This is a geometric picture of first-order ODEs. Let us start from the following examples.

1. The family of exponential curves $y = Ce^{-\alpha x}$ satisfies $y' = \alpha y$. In fact, we differentiate them in x and get $y' = -\alpha C e^{-\alpha x}$. Then we eliminate C from these two equations to get the equation $y' = \alpha y$.

2. Consider the family of quadratic curves $y = (x-C)^2$. We differentiate it in x and obtain $y' = 2(x-C)$. We eliminate C from both equations and get $\sqrt{y} = (x-C) = y'/2$. Thus, this family of quadratic curves satisfies $y' = 2\sqrt{y}$.

3. The family $y = K/(1 - C_1 e^{-rt})$ satisfies $y' = ry(1 - y/K)$. This is left for you as an exercise.

4. Consider the concentric circles: $x^2 + y^2 = C$. We differentiate it in x and obtain $2x + 2yy' = 0$. This gives $y' = -x/y$.

5. The family of confocal ellipses can be expressed as

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1 \text{ with } a^2-b^2=c^2 \text{ fixed.}$$

We look for the ODE that these confocal ellipses satisfy and their orthogonal trajectories. Without loss of generality, let us assume the foci are $(\pm 1, 0)$, i.e. $c=1$. The family of these confocal ellipses is

$$\frac{x^2}{a^2}+\frac{y^2}{a^2-1}=1,\ a>1. \tag{1.23}$$

Let us differentiate this equation in x and obtain

$$\frac{x}{a^2}+\frac{yy'}{a^2-1}=0. \tag{1.24}$$

We eliminate a from the above two equations to obtain an ODE as the follows. From (1.24), we get

$$a^2=\frac{x}{yy'+x}.$$

Plug this into (1.23), we obtain

$$\frac{x^2}{\frac{x}{yy'+x}}+\frac{y^2}{\frac{-yy'}{yy'+x}}=1.$$

After rearrangement, we get

$$xy\left(y'-\frac{1}{y'}\right)+x^2-y^2=1.$$

This is the ODE for the confocal ellipses.

We summarize the above procedure below. A family of plane curves can be expressed as

$$\psi(x,y)=C \quad \text{(level set representation),}$$

or

$$\Psi(x,y,C)=0 \text{ (implicit representation).} \tag{1.25}$$

In the former representation, we can differentiate it in x and obtain

$$\psi_x(x,y)+\psi_y y'=0.$$

This is a first order ODE

$$y'=-\psi_y(x,y)/\psi_x(x,y)$$

that this family of curves satisfies. In the latter representation, we differentiate (1.25) in x and obtain

$$\Psi_x(x,y,C)+\Psi_y(x,y,C)y'=0. \tag{1.26}$$

We use (1.25), (1.26) to eliminate C and obtain an equation $F(x,y,y')=0$. This is the ODE for this family of curves.

Remarks A family of curves $\Psi(x, y, C) = 0$ can also be represented as $\psi(x, y) = C$. This is valid locally under the condition $\Psi_C(x_0, y_0, C_0) \neq 0$. This is called the implicit function theorem, which will be used very often in this course.

Theorem 1.2 (Implicit Function Theorem)**.** *Suppose F is a continuous differentiable function and $F(x_0, y_0, z_0) = 0$. If $F_z(x_0, y_0, z_0) \neq 0$, then there exist a neighbor U of (x_0, y_0), a neighborhood V of z_0 and a continuous differentiable function $f : U \to V$ such that*

$$F(x, y, f(x, y)) = 0.$$

The function $z = f(x, y)$ is obtained by solving $F(x, y, z) = 0$ for $z \sim z_0$. A simple example is $F(x, y, z) = x^2+y^2+z^2-1 = 0$. We can solve this equation for z and obtain $z = \pm\sqrt{1 - x^2 - y^2}$. Suppose $F(x_0, y_0, z_0) = 0$. Then we choose the solution $z = \sqrt{1 - x^2 - y^2}$ if $z_0 > 0$, and $z = -\sqrt{1 - x^2 - y^2}$ if $z_0 < 0$. If $z_0 = 0$, $F_z(x_0, y_0, z_0) = 0$, we cannot apply the implicit function theorem for this case.

Orthogonal Trajectories In geometry and physics, we encounter problems of constructing orthogonal curvilinear coordinate systems. For instance, polar coordinate system is an orthogonal system. In this example, suppose we are given one family of concentric circles: $x^2 + y^2 - C = 0$. We ask how to construct another family of curves which are orthogonal to these curves. In general, suppose we are given a family of curves which is represented by $\Psi(x, y, C) = 0$, we look for their orthogonal trajectories. The method to find orthogonal trajectories is very simple. First, we find the ODE: $F(x, y, y') = 0$ that this family of curves satisfies. Next, we replace y' by $-1/y'$ in this ODE (i.e. $F(x, y, -1/y') = 0$). This is the ODE for the orthogonal trajectories. Finally, we solve this ODE to get the orthogonal trajectories.

In the example of concentric circles, the tangent of concentric circles is $y' = -x/y$. So, its normal should satisfy $-1/y' = -x/y$. The general solutions of this equation are $y/x = C$. This is the family of the orthogonal trajectories of the concentric circles.

In the example of confocal ellipses, the corresponding ODE is

$$xy\left(y' - \frac{1}{y'}\right) + x^2 - y^2 = 1.$$

Their orthogonal trajectories satisfy the ODE with y' replaced by $-1/y'$. This leads to the same equation. The general solutions have the same form (1.23). We notice that when $0 < a < 1$, these curves are confocal hyperbolas:

$$\frac{x^2}{a^2} - \frac{y^2}{1 - a^2} = 1, \ 0 < a < 1, \tag{1.27}$$

which are the orthogonal trajectories of the confocal ellipses.

1.5.3 *Envelop.

Given a family of curves $\Psi(x, y, C) = 0$. The degenerate part $\Psi_C(x, y, C) = 0$ and $\Psi(x, y, C) = 0$ forms an envelop of this family of curves. The envelop is tangent to each member of the family and

it is composed of all these tangent points. The envelop appears in geometric optics, called caustics. As an example, consider the following family of straight lines:

$$\frac{x}{C} + \frac{y}{1-C} - 1 = 0$$

One can express it as

$$\Psi(x, y, C) := C^2 + (-x + y - 1)C + x = 0.$$

The envelop is determined by $\Psi(x, y, C) = 0, \Psi_C(x, y, C) = 0$, which are

$$\begin{cases} C^2 + (-x + y - 1)C + x = 0 \\ 2C + (-x + y - 1) = 0. \end{cases}$$

This is a curve. We can also solve (x, y) in terms of C, which give the representation of this envelop in parametric form:

$$\begin{aligned} x &= C^2 \\ y &= (C-1)^2 \end{aligned}$$

We can eliminate C and get

$$y = (\sqrt{x} - 1)^2 = x + 1 - 2\sqrt{x}.$$

This is the explicit form of the envelop. Moving $x + 1$ to the left-hand side, taking square, we get

$$(y - x - 1)^2 = 4x.$$

This is the implicit form of the envelop.

Next, we explain why the envelop of a family of curves $\Psi(x, y, C) = 0$ is given by

$$\begin{cases} \Psi(x, y, C) = 0 \\ \Psi_C(x, y, C) = 0. \end{cases}$$

Since the envelop is composed of all tangent points, it can be parameterized by the parameter C. Thus, we may write the envelop in parametric form as $(x(C), y(C))$. It satisfies

$$\Psi(x(C), y(C), C) = 0,$$

because the tangent point $(x(C), y(C))$ lies on the curve $\Psi(x, y, C) = 0$. Differentiate this equation in C, we obtain

$$\Psi_x(x(C), y(C), C)x'(C) + \Psi_y(x(C), y(C), C)y'(C) + \Psi_C(x(C), y(C), C) = 0,$$

Since the tangent of this envelop, $(x'(C), y'(C))$, is tangent to the curve $\Psi(x, y, C) = 0$ as well, we obtain

$$\Psi_x(x(C), y(C), C)x'(C) + \Psi_y(x(C), y(C), C)y'(C) = 0.$$

Thus, we get

$$\Psi_C(x(C), y(C), C) = 0.$$

Given a family of curves, it may not have envelop at all. Confocal ellipses, confocal hyperbolas have no envelop. Below are some examples that possess envelops.

- The family of curve $y = (x - C)^2$ has envelop $y(x) \equiv 0$.
- Consider the cycles:
$$(x - \cos\theta)^2 + (y - \sin\theta)^2 = R^2.$$
The parameter θ running in $[0, 2\pi)$. You can show that its envelop is again a circle.

Homeworks

1. Find the orthogonal trajectories of family of parabolas with common vertex and common tangent at the common vertex:
$$y = cx^2,\ c \in \mathbb{R}.$$
2. Find the orthogonal trajectories of the family of parabola $y^2 = 4c(x + c)$.
3. *The potential of an electric dipole is
$$V_{dip}(r, \theta) = \frac{p\cos\theta}{4\pi\epsilon_0 r^2}$$
Here, (r, θ, ϕ) is the spherical coordinate system, ϵ_0 is the dielectric coefficient in vacuum, p is the dipole moment. The equipotential forms a family of curves. The electric field line is their orthogonal trajectories. Find these orthogonal trajectories.

1.5.4 *An example from thermodynamics – existence of entropy

Consider a thermodynamic system: a container with fixed amount of gases inside and having one free end (a piston) which allows volume change. The basic thermodynamic variables are the volume V, the pressure p, the internal energy e, and the temperature T. In order to have a thermo system to exchange energy with external world, we will also introduce a thermo variable S, called the entropy, which will be defined below. These five variables V, p, e, T, S are not independent. There are two constitutive relations plus the first law of thermodynamics relate them. The last one is a differential relation. Finally, they are only two independent thermo variables. Below, we introduce the simplest constitutive relations: the ideal gas law and the linear constitutive law.

The ideal gas law is
$$pV = RT,$$
where R is called the universal gas constant. For so-called polytropic gases, the internal energy is linearly proportional to the temperature T, i.e.
$$e = c_v T$$
where c_v is called the specific heat at constant volume. It means that the amount of energy you need to add to the system at constant volume to gain one degree increase of temperature.

In order to have energy exchange with external world, we introduce the notion “entropy” below. First, we can change the volume V of the system by moving the piston. If the process is moved

slowly, we imagine that the system has no energy exchange with external environment except the work that we apply to it through the piston. Such a process is called an *adiabatic process* (no heat exchange with the external world). In such a process, by the conservation of energy,

$$de = -pdV,$$

where $-pdV$ is the work we apply to the system. This is a Pfaffian equation. Using the ideal gas law and the assumption of polytropic gas, we get

$$\frac{c_v}{R}(pdV + Vdp) = -pdV.$$

This gives

$$\left(1 + \frac{c_v}{R}\right) pdV + \frac{c_v}{R} V dp = 0.$$

We divide both sides by c_v/R, we get

$$\gamma pdV + Vdp = 0,$$

where

$$\gamma := \frac{1 + \frac{c_v}{R}}{\frac{c_v}{R}},$$

is called the gas constant. This Pfaffian equation can be integrated by using the technique of separation of variable:

$$\frac{\gamma dV}{V} + \frac{dp}{p} = 0.$$

Thus, we get

$$\ln p + \gamma \ln V = C$$

Hence,

$$pV^\gamma$$

is a constant. This means that each adiabatic process keeps pV^γ invariant (the integral of an adiabatic process). The quantity pV^γ labels a thermo state of the system. It is called an entropy. Notice that any function of pV^γ is also invariant under an adiabatic process. The one which has $1/T$ as an integration factor for the Pfaffian equation $de + pdV = 0$ is called *the physical entropy*. That is

$$TdS = de + pdV.$$

This leads to

$$\begin{aligned} dS &= \frac{1}{T}(de + pdV) \\ &= \frac{R}{pV}\left(\frac{c_v}{R}(pdV + Vdp) + pdV\right) \\ &= c_v\left(\gamma\frac{dV}{V} + \frac{dp}{p}\right) \\ &= c_v d\ln(pV^\gamma) \\ &= dc_v \ln(pV^\gamma) \end{aligned}$$

Thus, the physical entropy

$$S = c_v \ln(pV^\gamma).$$

In conclusion, the first law of thermodynamics is

$$de = TdS - pdV. \tag{1.28}$$

This means that the change of internal energy can be due to the heat TdS exchange with external world and the work $-pdV$ exerted from outside. For ideal polytropic gases, using the ideal gas law and the constitutive relation, plus the first law of thermodynamics, we can choose p, V as independent variables and express

$$T = \frac{pV}{R},\ e = \frac{c_v}{R}pV,\ S = c_v \ln(pV^\gamma).$$

***Homework**

1. Express thermo variables in terms of e, V for ideal polytropic gases..

1.6 Existence and Uniqueness

In this section, we shall state but without proof the existence and uniqueness theorems. We also give examples and counter-examples regarding to the existence and uniqueness. Finally, we give application of these fundamental theorems.

Existence

Theorem 1.3 (Local existence theorem)**.** *Suppose $f(t, y)$ is continuous in a neighborhood of (t_0, y_0). Then the initial value problem*

$$\begin{aligned} y'(t) &= f(t, y), \\ y(t_0) &= y_0 \end{aligned}$$

has a solution $y(\cdot)$ existing on a small interval $(t_0 - \epsilon, t_0 + \epsilon)$ for some small number $\epsilon > 0$.

This theorem states that there exists an interval (may be small) where a solution does exist. The solution may not exist for all t. Let us see the following example.

Examples Consider the initial value problem

$$\begin{aligned} y' &= y^2 \\ y(0) &= y_0 \end{aligned}$$

By the method of separation of variable,

$$\frac{dy}{y^2} = dt$$

$$\int_{y_0}^{y} \frac{dy}{y^2} = t$$
$$-y^{-1} + y_0^{-1} = t$$
$$y(t) = \frac{y_0}{1 - ty_0}.$$

When $y_0 < 0$, the solution does exist in $(1/y_0, \infty)$. But when $y_0 > 0$, the solution can only exist in $(-\infty, 1/y_0)$. The solution blows up when $t \to 1/t_0$:

$$\lim_{t \to 1/y_0} y(t) = \infty.$$

The maximal interval of existence is $(-\infty, 1/y_0)$ when $y_0 > 0$ and is $(1/y_0, \infty)$ when $y_0 < 0$.

In the local existence theorem, it only states that the solution exists in a small region. If the solution does have a limit at the end, say t_1, of this interval, we can solve the equation again to extend this function. One can show that this extended function also satisfies the differential equation at t_1 and beyond. Eventually, we can find the maximal interval of existence. If the solution remains bounded whenever it exists, then we can always find globally exists if $y(\cdot)$ stays bounded whenever it exists. We have the following corollary.

Corollary 1.1. *Consider the ODE: $y' = f(t, y)$. Suppose $f(t, y)$ is continuous on $\mathbb{R} \times \mathbb{R}$ and assume a solution stays bounded as long as it exists, then this solution exists for all time.*

Proof. Suppose the maximal interval of existence is (t_0, t_1). The assumption that $y(t)$ remains bounded in (t_0, t_1) plus $f(t, y)$ is continuous imply that $\lim_{t \to t_1} y(t)$ exists (why?). Then we can extend $y(\cdot)$ beyond t_1 by the local existence theorem. This contradicts to the hypothesis that (t_0, t_1) is the maximal interval of existence. □

Homeworks Find the maximal interval of existence for the problems below.

1. $y' = 1 + y^2,\ y(0) = y_0$
2. $y' = y^3,\ y(0) = y_0$
3. $y' = e^y,\ y(0) = y_0$
4. $y' = y \ln y,\ y(0) = y_0 > 0.$

Uniqueness

Theorem 1.4 (Uniqueness)**.** *Assume that f and $\partial f/\partial y$ are continuous in a small neighborhood of (t_0, y_0). Suppose $y_1(t)$ and $y_2(t)$ are two solutions that solve the initial value problem*

$$y' = f(t, y),\ y(t_0) = y_0$$

on an interval $(t_0 - \epsilon, t_0 + \epsilon)$ for some $\epsilon > 0$. Then

$$y_1(t) = y_2(t), \text{ for all } t \in (t_0 - \epsilon, t_0 + \epsilon).$$

In other word, no two solutions can pass through the same point in the t-y plane.

Application 1. Reduce high order equation to first-order system The above existence and uniqueness theorems also hold for general first-order ODE system:

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$$

where $\mathbf{f} : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ is a Lipschitz function. This means that: *given initial data* $(t_0, \mathbf{y}_0) \in \mathbb{R} \times \mathbb{R}^n$*, there exists a unique solution* $\mathbf{y} : (t_0 - \epsilon, t_0 + \epsilon) \to \mathbb{R}^n$ *with* $\mathbf{y}(t_0) = \mathbf{y}_0$. This theorem can be applied to high-order equations too. Indeed, any high-order equation can be transformed to an equivalent first-order system. Namely, the general n-th order equation

$$y^{(n)} = f(t, y, y', \cdots, y^{(n-1)}) \tag{1.29}$$

is equivalent to the following system

$$\begin{cases} y^{1\prime} &= y^2 \\ y^{2\prime} &= y^3 \\ &\vdots \\ y^{n\prime} &= f(t, y^1, y^2, \cdots, y^n) \end{cases} \tag{1.30}$$

We need n conditions to determine a unique solution for the first-order system (1.30). Likely, we need n conditions to determine a unique solution for the nth-order differential equations (1.29).

Application 2 Let us apply the existence and uniqueness to the qualitative study of the autonomous system in one dimension. For instance, let consider a smooth $f(y)$ which has the property (i) the only zeros of f are 0 and K, (ii) $f(y) > 0$ for $0 < y < K$. The logistic model: $y' = f(y) := ry(1 - y/K)$, is one such example. The constant states 0 and K naturally partition the domain $\mathbb{R}$ into three regions: $I_1 = (-\infty, 0)$, $I_2 = (0, K)$ and $I_3 = (K, \infty)$. By the uniqueness theorem, no solution can cross these two constant states. Thus, starting $y(0) \in (0, K)$, the trajectory $y(t)$ stays in $(0, K)$ for all t because it cannot cross these two constant solutions. So, the solution stays bounded and thus exists for all time. The limit $\lim_{t\to\infty} y(t)$ must exist because the function $y(\cdot)$ monotonically increases and stays bounded above. Let us call $\lim_{t\to\infty} y(t) = \bar{y} \in [0, K]$. Then

$$\lim_{t\to\infty} y'(t) = \lim_{t\to\infty} f(y(t)) = f(\bar{y}).$$

We claim that $f(\bar{y}) = 0$. Suppose not, then we have $f(\bar{y}) > 0$ because $f(y) > 0$ for $y \in (0, K)$. We choose $\epsilon > 0$ so that $f(\bar{y}) - \epsilon > 0$. With this ϵ, there exists $M > 0$ such that $f(y(t)) > f(\bar{y}) - \epsilon$ for all $t > M$. Thus,

$$y(t) - y(M) = \int_M^t f(y(s))\, ds > (f(\bar{y}) - \epsilon)(t - M) \to \infty \text{ as } t \to \infty.$$

This is a contradiction. Thus, we get $f(\bar{y}) = 0$. But the only constant states are 0 and K. It has to be K because $0 < y(0) < y(t)$ for all $t > 0$. This shows that when $y(0) \in (0, K)$, we have $y(t) \to K$ as $t \to \infty$. This is asymptotic stability result. We will see more applications of the uniqueness theorem in the subsequent chapters.

Remarks.

1. The initial value problem may not have a unique solution. Let us see the following problem:

$$y' = 2y^{1/2},\ y(0) = 0$$

By the method of separation of variable,

$$\frac{dy}{2\sqrt{y}} = dt,$$

$$\sqrt{y} = t - C$$

With the initial condition $y(0) = 0$, we get $C = 0$. Hence

$$y(t) = t^2$$

is a solution. On the other hand, we know $y(t) \equiv 0$ is also a solution. We should be careful here. The portion $y(t) = t^2$ for $t < 0$ is not a solution because $y' < 0$ for $t < 0$. This portion does not satisfy the equation $y' = 2\sqrt{y} > 0$. Therefore, one solution is

$$y(t) = \begin{cases} 0 & \text{for } -\infty < t < 0 \\ t^2 & \text{for } t \geq 0. \end{cases}$$

We have known that $y(t) \equiv 0$ is another solution. In fact, there are infinite many solutions passing through $(0, 0)$:

$$y(t) = \begin{cases} 0 & \text{for } -\infty < t < C \\ (t - C)^2 & \text{for } t \geq C, \end{cases}$$

with parameter $C \geq 0$ being arbitrary.

It is important to notice two things. (1) The ODE associated with the family of parabolas $y = (t - C)^2$ is $y'^2 = 4y$, see the subsection 1.5.2. It contains two branches: $y' = \pm 2\sqrt{y}$. The solutions also contain two branches. The branch $y = (t-C)^2, t \geq C$ satisfies $y' = 2\sqrt{y}$, while the branch $y = (t - C)^2, t \leq C$ satisfies $y' = -2\sqrt{y}$. (2) The curve $y(t) \equiv 0$ is the envelop of both families of parabolas.

2. You can find non-uniqueness examples easily from the envelop of a family of curves. In fact, suppose the family of curve $\Psi(x, y, C) = 0$ is the solution of some ODE: $F(x, y, y') = 0$. Suppose $((x(C), y(C))$ is the envelop of this family of curves. Then at C, both $\Psi(x, y, C) = 0$ and the envelop $(x(C), y(C))$ are the solution curves of the ODE: $F(x, y, y') = 0$ at $(x(C), y(C))$.

3. For vector field $\mathbf{V}(x, y) = (u(x, y), v(x, y))$, its integral curves do not intersect if $\mathbf{V} \neq 0$. More precisely, if $(u(x_0, y_0), v(x_0, y_0)) \neq (0, 0)$, then the integral curve through (x_0, y_0) is unique. To show this, if $u(x_0, y_0) \neq 0$, then the integral curve of (u, v) satisfies

$$\frac{dy}{dx} = \frac{v(x, y)}{u(x, y)} = f(x, y).$$

The function $f(x,y)$ is well-defined in a neighborhood of (x_0,y_0) because $u(x_0,y_0)\neq 0$. By the uniqueness theorem, the solution $y(x)$ of $y'=f(x,y)$ with $y(x_0)=y_0$ is unique. Therefore, the integral curve is unique there. If on the other hand, $u(x_0,y_0)=0$, then $v(x_0,y_0)\neq 0$, we solve

$$\frac{dx}{dy}=\frac{u(x,y)}{v(x,y)}$$

instead.

However, the integral curves can "intersect" at those critical points where $\mathbf{V}(x,y)=(0,0)$. For instance, the integral curves of $\mathbf{V}(x,y)=(-x,-y)$ point to $(0,0)$. The integral curve of $\mathbf{V}=(-x,y)$ are $xy=C$. As $C=0$, the corresponding integral curve is $x=0$ or $y=0$. They intersect at $(0,0)$.

4. In the example of application 2, we cannot obtain the rate of convergence for $y(t)\to K$ as $t\to\infty$. However, if we know that $f'(K)\neq 0$ (in fact, $f'(K)<0$), then we can get that $y(t)\to K$ at exponential rate. This means that

$$|y(t)-K|\leq Ce^{f'(K)t}$$

as $t\to\infty$. A concrete is the logistic model, where $f(y)=ry(1-y/K)$ and $f'(K)=-r$. For $y\sim K$, the Taylor expansion of f gives $f(y)\sim -r(y-K)$. The equation $y'=f(y)\sim -r(y-K)$ leads to $y(t)\sim O(1)e^{-rt}$.

On the other hand, if $f'(K)=0$, then $f(y)$ is of high order near $y=K$. In this case, we can not have exponential convergence, as you can see from this simple example: $y'=(y-K)^2$ with $y(0)<K$.

1.7 *Numerical Methods: First Order Difference Equations

1.7.1 Euler method

Consider the first order equation

$$y'=f(t,y).$$

If the solution is smooth (this is what we would expect), we may approximate the derivative $y'(t)$ by a finite difference

$$y'(t)\sim\frac{y(t+h)-y(t)}{h}.$$

Thus, we choose a time step size h. Let us denote $t^0+nh=t^n$ and t^0 is the initial time. We shall approximate $y(t^n)$ by y^n. For $t^n<t<t^{n+1}$, $y(t)$ is approximated by a linear function. Thus, we approximate $y'=f(t,y)$ by

$$\frac{y^{n+1}-y^n}{h}=f(t^n,y^n). \tag{1.31}$$

This is called the Euler method. It approximates the solution by piecewise linear function. The approximate solution y^{n+1} can be computed from y^n. If we refine the mesh size h, we would

expect the solution get closer to the true solution. To be more precise, let us fix any time t. Let us divide $[0, t]$ into n subintervals evenly. Let $h = t/n$ be the step size. We use Euler method to construct y^n. The convergence at t means that $y^n \to y(t)$ as $n \to \infty$ (with $nh = t$ fixed, hence $h \to 0$).

Homework

1. Use Euler method to compute the solution for the differential equation

$$y' = ay$$

where a is a constant. Find the condition on h such that the sequence y^n so constructed converges as $n \to \infty$ and $nh = t$ is fixed.

1.7.2 First-order difference equation

This subsection is a computer project to study the discrete logistic map:

$$y_{n+1} = \rho y_n \left(1 - \frac{y_n}{k}\right). \tag{1.32}$$

It is derived from the Euler method for the logistic equation.

$$\frac{y_{n+1} - y_n}{h} = r y_n \left(1 - \frac{y_n}{K}\right),$$

with $\rho = 1 + rh$ and $k = K(1 + rh)/rh$. We use the following normalization: $x_n = y_n/k$ to get

$$x_{n+1} = \rho x_n (1 - x_n) := F(x_n). \tag{1.33}$$

This mapping ($F : x_n \mapsto x_{n+1}$) is called the logistic map. The project is to study the behaviors of this logistic map by computer simulations.

Iterative map In general, we consider a function $F : \mathbb{R} \to \mathbb{R}$. The mapping

$$x_{n+1} = F(x_n), n = 0, 1, 2, \cdots,$$

is called an iterative map. We denote the composition $F \circ F$ by F^2.

A point x^* is called a fixed point (or an equilibrium) of the iterative map F if it satisfies

$$F(x^*) = x^*$$

A fixed point x^* is called stable if we start the iterative map from any x_0 close to x^*, the sequence $\{F^n(x_0)\}$ converges to x^*. A fixed point x^* is called unstable if we start the iterative map from any x_0 arbitrarily close to x^*, the sequence $\{F^n(x_0)\}$ cannot converge to x^*. The goal here is to study the behavior (stable, unstable) of a fixed point as we vary the parameter ρ.

1. Find the condition on ρ such that the logistic map F maps $[0, 1]$ into $[0, 1]$.
2. For $\rho = 0.5, 1.5, 2.5$ find the region of x_0 in $[0, 1]$ so that $\lim_{n\to\infty} F^n x_0$ converges. Such region is called the region of contraction. It means that any point inside will be contract to a fixed point.
3. A point x is called a period n point if $F^n(x) = x$ but $F^m(x) \neq x$ for all $0 < m < n$. Find the set of period 2 points.

1.8 Historical Note

You can find the figures below from Wikipedia.

Data, modeling

- Tycho Brahe (1546-1601)
- Galileo Galilei (1564-1642)
- Johannes Kepler (1571-1630)

Calculus and Numerical Method

- Isaac Newton (1642-1726)
- Leonhard Euler (1707-1783)

Population model

- Thomas Malthus (1766-1834)
- Pierre Verhulst (1804-1849)

Chapter 2

Second Order Linear Equations

In this chapter, we study linear second-order equations of the form:

$$ay'' + by' + cy = f(t), \tag{2.1}$$

with constant coefficients and $a \neq 0$. We shall investigate the model of linear oscillator in great detail. It is a fundamental model in physics. We shall use spring-mass system and the electric circuit system as examples.

2.1 Models for linear oscillators

2.1.1 The spring-mass system

Consider a mass attached to a spring in one dimension. Let y be its location, and let $y = 0$ be its position at rest. The motion of the mass is governed by Newton's force law:

$$my'' = F.$$

There are three kinds of forces the mass may be exerted.

- **Restoration force.** As the mass moves to y, it is exerted a restoration force by the spring. The simplest form of the restoration force is based on Hook's law: this restoration force is linearly proportional to y with reverse direction. That is,

  $$F_r = -ky$$

  where k is called the spring constant. The minus sign indicates that the force is opposite to the direction of the mass motion.

- **Friction force.** The friction force is proportional to the velocity with opposite direction. That is

  $$F_f = -\gamma y',$$

  where γ is called the damping (or friction) coefficient.

- **External force.** The mass may be exerted by the gravitational force, or some other external force modeled by $f(t)$.

The Newton's law of motion then gives

$$my'' = -\gamma y' - ky + f(t). \tag{2.2}$$

2.1.2 Electric circuit system

Consider a circuit which consists of an inductor, a resistor, a capacitor and a battery. Suppose the wire is uniform in width. In this case, we may assume the current between each component is uniform (i.e. it is independent of the position). This is a good approximation when the electromagnetic wave length is much larger than the wire length. Through each component, the rate of change of charge (i.e. current $I(t) := dQ(t)/dt$) should be the same. This is the law of conservation of charge, one of the Kirkohoff law in the circuit theory. When the electric current passing through these components, there is a potential difference on the two ends of each components. Namely, the potential difference through each component is

- resistor $\Delta V_r = RI$:
  A resister is a dielectric material. It is used to reduce current flow, to lower voltage levels within circuits. The potential difference between the two ends of a resistance induces an electric field E. It drives electrons in the resistance move at current I. The Ohm law states that I is proportional to E and hence $\Delta V_r = Ed = \alpha Id$, where d is the length of the resistance.

- capacitor $\Delta V_c = Q/C$:
  A typical capacitor is a pair of parallel plates with equal charges and opposite signature. There is an electric field E induced by the charges on the two plates. By Gauss law, $\epsilon EA = Q$, where A is the area of the plate and ϵ is the dielectric coefficient. It is clear that the more charges on the plates, the higher the electric field. The potential difference on the two plates is $\Delta V = Ed$, where d is the distance between the two plates. Hence,

  $$\Delta V_c = Ed = \frac{d}{\epsilon A} Q = \frac{Q}{C}.$$

  Capacitor is used to store charges or energy within circuits.

- inductor $\Delta V_i = L\frac{dI}{dt}$.
  An inductance is a solenoid. By the Amperè law, the current on a circular wire induces a magnetic field mainly through the disk that the circle surrounds. The time-varying current (i.e. dI/dt) induces a time-varying magnetic field. By the Farady law, this time-varying magnetic field induces an electric field E which can move (electromotive force) the charges in the wire, called induced current. The magnitude of E is proportional to dI/dt. Thus, there is a linear relation between the potential drop ΔV (which is Ed, d is the length of the inductance) and dI/dt.

The constants R, C, L are called the resistance, conductance and inductance, respectively. From another Kirkohoff law (conservation of energy), we have

$$L\frac{d^2Q}{dt^2} + R\frac{dQ}{dt} + \frac{1}{C}Q = f(t) \tag{2.3}$$

where $f(t)$ is the external potential from the battery.

We notice there is an analogy between mechanical oscillators and electrical oscillators.

2.2 Methods to solve homogeneous equations

We rewrite the above linear oscillator equation in an abstract form:

$$ay'' + by' + cy = f(t), \tag{2.4}$$

where $a \neq 0, b, c$ are constants. We are allowed $a, b, c \in \mathbb{C}$. We should prescribe initial data:

$$y(0) = y_0,\ y'(0) = y_1 \tag{2.5}$$

for physical consideration. We may express (2.4) in an operator form:

$$L(D)y = f, \tag{2.6}$$

where

$$L(D) = aD^2 + bD + c,\ D = \frac{d}{dt}.$$

The term f is called the source term.

2.2.1 Homogeneous equations (complex case)

Equation (2.4) without source term is called a homogeneous equation:

$$L(D)y := ay'' + by' + cy = 0. \tag{2.7}$$

We try a solution of the form $y(t) = e^{\lambda t}$ (called an *ansatz*) for the homogeneous equation. Plug this *ansatz* into the homogeneous equation. We obtain

$$L(D)\left(e^{\lambda t}\right) = L(\lambda)e^{\lambda t} = \left(a\lambda^2 + b\lambda + c\right)e^{\lambda t} = 0.$$

This leads to

$$a\lambda^2 + b\lambda + c = 0.$$

This polynomial equation is called the *characteristic equation* of (2.4). Let λ_1, λ_2 be its two roots (possible complex roots). There are two cases:

- Case 1: $\lambda_1 \neq \lambda_2$. In this case, we have found two solutions $y_1(t) = e^{\lambda_1 t}$ and $y_2(t) = e^{\lambda_2 t}$.

- Case 2: $\lambda_1 = \lambda_2$. In this case, we can check $y_1(t) = e^{\lambda_1 t}$ and $y_2(t) = te^{\lambda_1 t}$ are two solutions. Let me explain why $te^{\lambda_1 t}$ is a solution. Indeed, from λ_1 being the double root of $L(\lambda) = 0$, we have $L(\lambda_1) = 0$, and $L'(\lambda_1) = 0$. By plugging $te^{\lambda_1 t}$ into the equation (2.7), we obtain

$$L\left(\frac{d}{dt}\right)\left(te^{\lambda_1 t}\right) = L(\lambda_1)\left(te^{\lambda_1 t}\right) + L'(\lambda_1)\left(e^{\lambda_1 t}\right) = 0.$$

  Another way to understand the appearance of the solution $te^{\lambda_1 t}$ is by taking the limiting process: $\lambda_2 \to \lambda_1$. The double root of λ_1 can be viewed as the limit process of two distinguishing roots λ_2 and λ_1 with $\lambda_2 \to \lambda_1$. When $\lambda_2 \neq \lambda_1$, the function $(e^{\lambda_2 t} - e^{\lambda_1 t})/(\lambda_2 - \lambda_1)$ is a solution. As $\lambda_2 \to \lambda_1$, the limit

$$\lim_{\lambda_2 \to \lambda_1} \frac{e^{\lambda_2 t} - e^{\lambda_1 t}}{\lambda_2 - \lambda_1} = te^{\lambda_1 t}$$

  is also a solution of $L(D)y = 0$, with λ_1 becoming a double root of $L(\lambda) = 0$.

Fundamental solutions (complex case) The solutions

$$\begin{aligned} y_1(t) &= e^{\lambda_1 t} \\ y_2(t) &= \begin{cases} e^{\lambda_2 t} & \text{if } \lambda_2 \neq \lambda_1 \\ te^{\lambda_1 t} & \text{if } \lambda_2 = \lambda_1 \end{cases} \end{aligned} \tag{2.8}$$

are called the fundamental solutions of (2.7). We claim that they have the following properties:

1. linearity: any of their linear combination $C_1 y_1 + C_2 y_2$ is a solution;
2. independence: y_1 and y_2 are linear independent;
3. any complex-valued solution of (2.7) can be expressed as $C_1 y_1 + C_2 y_2$ for some coefficients $C_1, C_2 \in \mathbb{C}$.

We call the solution set

$$\mathcal{S}_0 := \{C_1 y_1(\cdot) + C_2 y_2(\cdot) | C_1, C_2 \in \mathbb{C}\}$$

the *solution space* of (2.7). It is a two dimensional vector space over the complex field $\mathbb{C}$.
Let us prove these claims below.

1. **Linearity** In fact,

$$L\left(\frac{d}{dt}\right)(C_1 y_1 + C_2 y_2) = C_1 L\left(\frac{d}{dt}\right) y_1 + C_2 L\left(\frac{d}{dt}\right) y_2 = 0,$$

   because the operator $L(D)$ is linear.

2. **Independence and Wronskian** We shall show that if $y(t) := C_1y_1(t) + C_2y_2(t) = 0$ for $t \in \mathbb{R}$, then $C_1 = C_2 = 0$.
In fact, if $y(t) \equiv 0$, then $y(t_0) = 0$ and $y'(t_0) = 0$ at some point t_0.

   Next, we write the second order equation (2.7) as a 2×2 system of first order equations:

$$\begin{bmatrix} y \\ y' \end{bmatrix}' = \begin{bmatrix} 0 & 1 \\ -\frac{c}{a} & -\frac{b}{a} \end{bmatrix} \begin{bmatrix} y \\ y' \end{bmatrix} \tag{2.9}$$

   From the existence and uniqueness theorem, any solution of (2.9) is uniquely determined by $(y(t_0), y'(t_0))$. We have known that $y(\cdot) = C_1y_1(\cdot) + C_2y_2(\cdot)$ is a solution. Plug it into the initial condition, we obtain two equations for C_1 and C_2:

$$\begin{aligned} y_1(t_0)C_1 + y_2(t_0)C_2 &= y(t_0) = 0 \\ y_1'(t_0)C_1 + y_2'(t_0)C_2 &= y'(t_0) = 0 \end{aligned}$$

   The two coefficients C_1 and C_2 can be uniquely determined by $y(t_0)$ and $y'(t_0)$ if and only if the determinant

$$W(y_1, y_2)(t_0) := \begin{vmatrix} y_1(t_0) & y_2(t_0) \\ y_1'(t_0) & y_2'(t_0) \end{vmatrix} \neq 0 \tag{2.10}$$

   We call this determinant the *Wronskian* of y_1 and y_2 at t_0. Plug (2.8) into (2.10), we get

$$W(y_1, y_2)(t_0) = \begin{cases} (\lambda_2 - \lambda_1)e^{(\lambda_1+\lambda_2)t_0} & \text{when } \lambda_1 \neq \lambda_2 \\ e^{2\lambda_1 t_0} & \text{when } \lambda_1 = \lambda_2 \end{cases}$$

   We see this Wronskian is never zero for any t_0. Hence $C_1 = 0$ and $C_2 = 0$. We conclude that $y_1(\cdot)$ and $y_2(\cdot)$ are linearly independent.

3. **Solution representation** Suppose $\tilde{y}$ is a solution, we claim that there are constants C_1 and C_2 such that $\tilde{y} = C_1y_1 + c_2y_2$. In fact, the initial data $(\tilde{y}(0), \tilde{y}'(0))$ determines a unique pair of C_1 and C_2 such that

$$\begin{aligned} y_1(0)C_1 + y_2(0)C_2 &= \tilde{y}(0) \\ y_1'(0)C_1 + y_2'(0)C_2 &= \tilde{y}'(0). \end{aligned}$$

   This is due to the fact that the Wronskian $W(y_1, y_2)(0) \neq 0$. With these C_1, C_2, the solutions $y(t) = C_1y_1(t) + C_2y_2(t)$ and $\tilde{y}(t)$ have identical data at $t = 0$. By the uniqueness theorem, $\tilde{y}(\cdot) = y(\cdot) = C_1y_1(\cdot) + C_2y_2(\cdot)$.

2.2.2 Homogeneous equation (real case)

In many applications, the equation

$$L(D)y = ay'' + by' + c = 0 \tag{2.11}$$

has *real* coefficients, i.e. $a \neq 0, b, c \in \mathbb{R}$. Here, D denotes for d/dt. The above complex-value theory is still applicable with small modification to produce real-valued solutions. Let us list two basic facts.

1. If $y(\cdot)$ is a complex-valued solution of (2.11), so are its real part $Re(y(\cdot))$ and imaginary part $Im(y(\cdot))$. This is due to the linearity of $L(D)y$ and the fact that $a, b, c \in \mathbb{R}$.

2. The roots of the characteristic equation $L(\lambda) = 0$ are complex conjugate.

Below, we shall construct two independent real-valued solutions and show that the solution space is a two-dimensional vector space over $\mathbb{R}$. To show these, there are three cases.

- **Case 1.** $\lambda_1 \neq \lambda_2$ **and real**. A general solution for the homogeneous equation has the form

  $$y(t) = C_1 y_1(t) + C_2 y_2(t),$$

  where

  $$y_1(t) := e^{\lambda_1 t},\ y_2(t) := e^{\lambda_2 t}.$$

  The constants C_1 and C_2 are determined by the initial condition (2.5):

  $$\begin{aligned} C_1 + C_2 &= y(0) \\ \lambda_1 C_1 + \lambda_2 C_2 &= y'(0). \end{aligned}$$

  From $\lambda_1 \neq \lambda_2$, we see that C_1 and C_2 can be solved uniquely.

- **Case 2.** $\lambda_1 \neq \lambda_2$ **and complex**. In this case, they are conjugate to each other. Let us denote $\lambda_1 = \alpha + i\omega$ and $\lambda_2 = \alpha - i\omega$. We have found two solutions

  $$\begin{aligned} y_1(t) &= Re(e^{\lambda_1 t}) = e^{\alpha t}\cos\omega t \\ y_2(t) &= Im(e^{\lambda_1 t}) = e^{\alpha t}\sin\omega t \end{aligned}$$

  A general solution of the form

  $$y(t) = C_1 y_1(t) + C_2 y_2(t),$$

  satisfying the initial condition (2.5) leads to

  $$\begin{aligned} y(0) &= C_1 \\ y'(0) &= C_1\alpha + C_2\omega. \end{aligned}$$

  The constants C_1 and C_2 can be solved uniquely because we have $\omega \neq 0$ in this case.

- **Case 3.** $\lambda_1 = \lambda_2 \in \mathbb{R}$**.** In this case,

  $$y_1(t) := e^{\lambda_1 t} \text{ and } y_2(t) := te^{\lambda_1 t}$$

  are two independent solutions. So, general solution has the form $C_1 y_1(t) + C_2 y_2(t)$. The constants C_1 and C_2 are determined by the initial data: to

  $$\begin{aligned} C_1 &= y(0) \\ \lambda_1 C_1 + C_2 &= y'(0). \end{aligned}$$

In the above three cases, the functions $\{y_1(\cdot), y_2(\cdot)\}$ form a basis of the solution space.

Homeworks.

1. Consider the ODE: $ay'' + by' + cy = 0$, with a, b, c being real. Suppose $y(t) = y_1(t) + iy_2(t)$ be a complex solution.
   (a) Show that both its real part y_1 and imaginary part y_2 are solutions too.
   (b) Show any linear combination of y_1 and y_2 is also a solution.
2. Let $\lambda = \alpha + i\omega$. Find the Wronskians $W(e^{\lambda t}, e^{\bar{\lambda} t})$, $W(e^{\alpha t}\cos\omega t, e^{\alpha t}\sin\omega t)$ and $W(e^{\lambda t}, te^{\lambda t})$.
3. Solve the initial value problem $y'' - y' - 2y = 0$, $y(0) = \alpha$, $y'(0) = 2$. Then find α so that the solution approaches zero as $t \to \infty$.
4. Consider the ODE
$$y'' - (2\alpha - 1)y' + \alpha(\alpha - 1)y = 0.$$
   (a) Determine the values of α for which all solutions tend to zero as $t \to \infty$.
   (b) Determine the values of α for which all solutions become unbounded as $t \to \infty$.

2.3 Methods to solve Inhomogeneous equations

Now, we study the inhomogeneous equation with general forcing term f:

$$ay'' + by' + cy = f(t).$$

We may abbreviate it by an operator notation:

$$L(D)[y] = f,$$

where $L(s) = as^2 + bs + c$. From the theory for homogeneous equations, we know that we can find two independent solutions. Let $y_1(\cdot)$ and $y_2(\cdot)$ be a pair of such fundamental solutions:

$$L(D)[y_i] = 0, i = 1, 2.$$

Suppose $y_p(\cdot)$ is a special solution of (2.4). This means that

$$L(D)[y_p] = f.$$

In this case, $y_p + C_1y_1 + C_2y_2$ is also a special solution for any constants C_1 and C_2. This is because the linearity of the equation. Namely,

$$L[y_p + C_1y_1 + C_2y_2] = L[y_p] + C_1L[y_1] + L[y_2] = f + 0 + 0.$$

From the existence and uniqueness of ODEs, we know that the solution set depends on two parameters. We can conclude that the solution set $\mathcal{S}$ to (2.4) is $\mathcal{S} = y_p + \mathcal{S}_0$, where $\mathcal{S}_0$ is the solution space corresponding to the homogeneous equation. In other words, the solution set of (2.4) is an *affine space*. The choice of the special solution y_p is not unique. If y_q is another special solution, then any solution represented by $y = y_p + z$ with $z \in \mathcal{S}_0$ can also be represented as $y = y_q + w$ with $w = y_p - y_q + z \in \mathcal{S}_0$. Thus, it suffices to find just one special solution.

We introduce two methods to find a special solution. In later chapter, we will further introduce the method of Laplace transform to find special solutions.

2.3.1 Method of underdetermined coefficients

Suppose λ_1 and λ_2 are the two roots of the characteristic equation $L(\lambda) = 0$. Suppose the source term is of the form:

$$t^k e^{\lambda t}.$$

Then we can use the following *method of underdetermined coefficient* to find a special solution. We use the following examples to explain.

- Case: $\lambda \neq \lambda_1$ and $\lambda \neq \lambda_2$. We try a special solution of the form

$$y_p(t) = (a_k t^k + a_{k-1} t^{k-1} + \cdots + a_0) e^{\lambda t}.$$

- Case: $\lambda = \lambda_1$. We try a special solution of the form

$$y_p(t) = t(a_k t^k + a_{k-1} t^{k-1} + \cdots + a_0) e^{\lambda t}.$$

Plugging this special form into equation, we obtain a polynomial equations. Equating both sides and we obtain $k+1$ linear equations for $k+1$ coefficients $a_k, ..., a_0$.

Examples

1. Let $f(t) = t$. We try $y_p = a_1 t + a_0$. Plug it into the equation, that is, $(aD^2 + bD + c)y_p = t$. We get

$$a \cdot 0 + b \cdot (a_1) + c \cdot (a_1 t + a_0) = t.$$

   This yields

$$\begin{aligned} ca_1 &= 1 \\ ba_1 + ca_0 &= 0. \end{aligned}$$

   We get that $y_p = t/c - b/c^2$ is a special solution.

2. Find a special solution for $y'' - y = te^{2t}$. We choose $y_p(t) = (at + b)e^{2t}$. Plug this into the equation, we get

$$4(at + b)e^{2t} + 4ae^{2t} - (at + b)e^{2t} = te^{2t}$$

   This yields

$$\begin{aligned} 3a &= 1 \\ 4b + 4a - b &= 0. \end{aligned}$$

   Hence, $a = 1/3$ and $b = -4/9$.

3. Let us consider $y'' - y = e^t$ as an example. We try $y_p = ate^t$. We have

$$\begin{aligned} y_p' &= ae^t + (at)e^t \\ y_p'' &= 2ae^t + (at)e^t \end{aligned}$$

The equation $y'' - y = e^t$ yields

$$(at)e^t + 2ae^t - (at)e^t = e^t.$$

This gives

$$\begin{aligned} a - a &= 0 \\ 2a &= 1 \end{aligned}$$

Hence, $y_p = \frac{1}{2}te^t$ is a special solution.

4. $f(t) = t^k e^{\alpha t}\cos(\Omega t)$, or $t^k e^{\alpha t}\sin(\Omega t)$. In this case, we introduce a complex forcing term

$$f(t) = t^k e^{\lambda t},\ \lambda := \alpha + i\Omega.$$

The real part of a solution to this complex forcing term is a special solution to the forcing term $t^k e^{\alpha t}\cos(\omega t)$. For this complex forcing term, it can be reduced to the previous case.

Homework. Find a special solution for the following equations.

1. $y'' - y = te^t$.
2. $y'' - 2y' + y = e^t$.
3. $y'' - 2y' + y = te^t$.
4. $y'' + 4y = te^{it}$.
5. $y'' + y = te^{it}$.
6. $y'' + 2y' + 2y = \sin t$.
7. $y'' + 2y' + 2y = e^{-t}\sin t$.

2.3.2 Method of Variation of Constants

We use variation of constants to solve the inhomogeneous equation (2.4). For simplicity, we may assume the coefficient a of (2.4) is 1. Suppose $y_1(\cdot)$ and $y_2(\cdot)$ are two independent solutions of the homogeneous equation (2.7). First, we write (2.4) in vector form

$$\mathbf{y}'(t) = \mathbf{A}\mathbf{y}(t) + \mathbf{f}, \tag{2.12}$$

where

$$\mathbf{y}(t) = \begin{pmatrix} y(t) \\ y'(t) \end{pmatrix}, \mathbf{A} = \begin{pmatrix} 0 & 1 \\ -c & -b \end{pmatrix}, \mathbf{f}(t) = \begin{pmatrix} 0 \\ f(t) \end{pmatrix}.$$

We assume a special solution of (2.4) has the form

$$\mathbf{y}(t) = C_1(t)\mathbf{y}_1(t) + C_2(t)\mathbf{y}_2(t), \tag{2.13}$$

where $C_i(t) \in \mathbb{C}$ are the coefficients to be determined, and $\mathbf{y}_i$ satisfies the homogeneous equation:

$$\mathbf{y}_i' = \mathbf{A}\mathbf{y}_i, i = 1, 2,$$

We can express $\mathbf{y}$ in matrix form

$$\mathbf{y}(t) = \Phi(t)\mathbf{C}(t), \quad \mathbf{C}(t) = \begin{pmatrix} C_1(t) \\ C_2(t) \end{pmatrix}$$

where

$$\Phi(t) := \begin{pmatrix} y_1(t) & y_2(t) \\ y_1'(t) & y_2'(t) \end{pmatrix}. \tag{2.14}$$

is the fundamental solution, which satisfies

$$\Phi'(t) = \mathbf{A}\Phi(t).$$

We plug this ansatz to (2.12):

$$\Phi'(t)\mathbf{C}(t) + \Phi(t)\mathbf{C}'(t) = \mathbf{A}\Phi(t)\mathbf{C}(t) + \mathbf{f},$$

we get

$$\Phi(t)\mathbf{C}'(t) = \mathbf{f}. \tag{2.15}$$

This gives us a first-order differential equation for $\mathbf{C}(t)$:

$$\mathbf{C}'(t) = \Phi^{-1}(t)\mathbf{f}(t) = \Phi^{-1}(t) \begin{pmatrix} 0 \\ f(t) \end{pmatrix}. \tag{2.16}$$

By integrating (2.16), taking $\mathbf{C}(0) = 0$, we obtain

$$\begin{aligned} \mathbf{C}(t) &= \int_0^t \Phi(s)^{-1} \begin{pmatrix} 0 \\ f(s) \end{pmatrix} ds \\ &= \int_0^t \frac{1}{W(y_1, y_2)(s)} \begin{pmatrix} y_2'(s) & -y_2(s) \\ -y_1'(s) & y_1(s) \end{pmatrix} \begin{pmatrix} 0 \\ f(s) \end{pmatrix} ds \\ &= \begin{pmatrix} C_1(0) \\ C_2(0) \end{pmatrix} + \int_0^t \frac{1}{W(y_1, y_2)(s)} \begin{pmatrix} -y_2(s)f(s) \\ y_1(s)f(s) \end{pmatrix} ds \end{aligned}$$

This gives

$$\mathbf{y}_p(t) = \Phi(t)\mathbf{C}(t).$$

The $y_p(t)$ component reads

$$y_p(t) = -y_1(t)\int_0^t \frac{y_2(s)f(s)}{W(y_1,y_2)(s)}\,ds + y_2(t)\int_0^t \frac{y_1(s)f(s)}{W(y_1,y_2)(s)}\,ds. \tag{2.17}$$

The general solution $y(t)$ has the form

$$y_p(t) + c_1y_1(t) + c_2y_2(t),$$

where c_1, c_2 are constants.

Examples

1. Solve the equation
$$y'' - y = f(t)$$
with initial data
$$y(0) = 0,\ y'(0) = 0.$$
Answer. The homogeneous equation $y'' - y = 0$ has fundamental solutions $y_1(t) = e^{-t}$ and $y_2(t) = e^t$. The corresponding Wronskian
$$W(y_1,y_2)(t) = \begin{vmatrix} y_1(t) & y_2(t) \\ y_1'(t) & y_2'(t) \end{vmatrix} = \begin{vmatrix} e^{-t} & e^t \\ -e^{-t} & e^t \end{vmatrix} = 2.$$
Thus, the special solution
$$\begin{aligned} y_p(t) &= -e^{-t}\int_0^t \frac{e^s f(s)}{2}\,ds + e^t\int_0^t \frac{e^{-s}f(s)}{2}\,ds \\ &= \int_0^t \sinh(t-s)f(s)\,ds \end{aligned}$$
You may check this special solution satisfies the initial conditions $y(0) = y'(0) = 0$.

2. Find a particular solution of
$$y'' + y = \csc t$$
for t near $\pi/2$.
Answer. The fundamental solutions corresponding to the homogeneous equation is
$$y_1(t) = \cos t,\ y_2(t) = \sin t.$$
The Wronskian $W(y_1,y_2)(t) = 1$. A special solution is given by
$$\begin{aligned} y_p(t) &= -y_1(t)\int_{\pi/2}^t \frac{y_2(s)f(s)}{W(y_1,y_2)(s)}\,ds + y_2(t)\int_{\pi/2}^t \frac{y_1(s)f(s)}{W(y_1,y_2)(s)}\,ds \\ &= -\cos t\int_{\pi/2}^t \sin(s)\csc(s)\,ds + \sin t\int_{\pi/2}^t \cos(s)\csc(s)\,ds \\ &= -(t-\pi/2)\cos t + \sin t\cdot\ln\sin t. \end{aligned}$$

Homeworks.

1. Find the solution of $y'' - y = f$ with $y(0) = y'(0) = 0$.
2. Find the solution of $y'' - y = f$ with $y(0) = y(\pi) = 0$.
3. Find the solution of $y'' + 2y' + y = f$ with $y(0) = y'(0) = 0$.
4. Find the solution of $y'' - 2\alpha y' + \alpha^2 y + \omega^2 y = f$ with $y(0) = y'(0) = 0$.

2.4 Linear oscillators

2.4.1 Harmonic oscillators

To understand the physical meaning of the solutions of the linear oscillation systems, let us first consider the case when there is no damping term (i.e. friction or resistance). That is

$$L(D)\, y = a\frac{d^2y}{dt^2} + cy = 0. \tag{2.18}$$

We call such system a harmonic oscillator or free oscillator. The corresponding characteristic equation $a\lambda^2 + c = 0$ has two characteristic roots

$$\lambda_1 = -i\sqrt{\frac{c}{a}},\ \lambda_2 = i\sqrt{\frac{c}{a}},$$

which are pure imaginary due to both $a, c > 0$ in a harmonic oscillator. Let us denote

$$\omega_0 = \sqrt{\frac{c}{a}} \tag{2.19}$$

Then the general solution for (2.18) is

$$C_1 e^{-i\omega_0 t} + C_2 e^{i\omega_0 t}.$$

Its real part forms the real solution of (2.18). It has the form

$$y(t) = B_1 \cos\omega_0 t + B_2 \sin\omega_0 t,$$

where B_i are real. We may further simplify it as

$$y(t) = A\cos(\omega_0 t + \theta_0) \tag{2.20}$$

where

$$A = \sqrt{B_1^2 + B_2^2},\ \cos(\theta_0) = B_1/A, \sin(\theta_0) = -B_2/A,$$

A is called the amplitude and θ_0 is the initial phase. They are related to the initial data y_0 and y_1 by

$$y_0 = A\cos(\theta_0),\ y_1 = \omega_0 A\cos(\theta_0).$$

This motion is called harmonic oscillation or free oscillation. It is important to note that through a transformation:

$$y = \cos\theta$$

the ODE (2.18) is converted to a linear motion with constant speed:

$$\frac{d^2\theta}{dt^2} = 0, \ \frac{d\theta}{dt} = \omega_0 \tag{2.21}$$

Its solution solution is given by $\theta(t) = \theta_0 + \omega_0 t$. So it can be viewed as a circular motion with constant angular speed.

2.4.2 Damping

In this section, we consider (2.4) with damping term:

$$ay'' + by' + cy = 0.$$

The coefficient $b > 0$. We recall that the homogeneous equation has two independent solutions $e^{\lambda_1 t}$ and $e^{\lambda_2 t}$, where

$$\lambda_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}, \ \lambda_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a},$$

are the two roots of the characteristic equation $a\lambda^2 + b\lambda + c = 0$. We have the following cases: $\Delta = b^2 - 4ac < 0, = 0$ or > 0.

Case 1. damped free oscillation When $b^2 - 4ac < 0$, we rewrite

$$\lambda_1 = -\alpha + i\omega, \ \lambda_2 = -\alpha - i\omega,$$

where $\alpha = b/2a > 0$, $\omega = \sqrt{4ac - b^2}/2a > 0$. Then two independent solutions are

$$y_1(t) = e^{-\alpha t}\cos(\omega t), \ y_2(t) = e^{-\alpha t}\sin(\omega t).$$

So, the general solutions for the homogeneous equation oscillate (the damper is not so strong and the oscillation is still maintained), but their amplitudes damp to zero exponentially fast at rate $b/2a$. The relaxation time is $\tau := 2a/b$. Thus, the smaller b is (weeker damper), the longer the relaxation time is. But, as long as $b > 0$, the solution decays to zero eventually.

In the spring-mass system, $a = m$, $b = \gamma$, $c = k$. The free oscillation frequency is $\omega_0^2 = k/m$. The effective oscillation $y_i = e^{-(\gamma/2m)t}e^{i\omega t}$ has frequency $\omega = \sqrt{4mk - \gamma^2}/2m < \sqrt{k/m} = \omega_0$. Thus, the damping slows down the oscillation frequency. The frequency ω is called the *quasifrequency*.

Case 2. Critical damping When $b^2 - 4ac = 0$, the eigenvalue $\lambda_1 = -b/2a$ is a double root. In additional to the solution $y_1(t) = e^{\lambda_1 t}$, we can check

$$y_2(t) = te^{\lambda_1 t}$$

is another solution. You may check that this solution still decays to zero as $t \to \infty$. Certainly it is slower than $y_1(t)$. A concrete example is $y'' + 2y' + y = 0$.

Case 3. Overdamping When $b^2 - 4ac \geq 0$, λ_i are real and negative. The two independent solutions

$$y_i(t) = e^{\lambda_i t} \to 0, \text{ as } t \to \infty, \ i = 1, 2.$$

We call this is overdamping. It means that the damper is too strong so that the solution has no oscillation at all and decays to 0 exponentially fast. The decay rate is $O(e^{-\alpha t})$, where $\alpha = b/2a$. The quantity $1/\alpha$ is called the relaxation time. As a concrete example, consider $y'' + 3y' + y = 0$. One eigenvalue is $\lambda_1 = -3/2 + \sqrt{5}/2$. The other is $\lambda_2 = -3/2 - \sqrt{5}/2$. We see the solution $y_1(t) = e^{\lambda t}$ decays slower than $y_2(t) := e^{\lambda_2 t}$.

Homeworks.

1. Consider the ODE $my'' + \gamma y' + ky = 0$ with $\gamma > 0$. Show that the energy defined by

$$E(t) := \frac{m}{2} y'(t)^2 + \frac{1}{2} k y(t)^2$$

   satisfies $E'(t) \leq 0$.

2. Consider the ODE $my'' + \gamma y' + ky = 0$ with $y(0) = y_0$, $y'(0) = v$, $\gamma > 0$. Express the solution in the form $y(t) = R \exp(-\gamma t/2m) \cos(\omega_0 t - \delta)$ and determine R in terms of m, γ, k, y_0 and v explicitly.

3. Consider the ODE $y'' + \alpha y' + \omega_0^2 y = 0$ with $\alpha, \omega > 0$. In the critical case ($\alpha = 2\omega_0$), there is a solution $y^*(t) = te^{-\omega_0 t}$. When $\alpha < 2\omega_0$, construct a solution y_α such that $y_\alpha \to y^*$ as $\alpha \to 2\omega_0$.

4. There are many interesting resonance phenomena in nature, search into websites with key word "resonance".

2.4.3 Forcing and Resonance

In this section, we study forced vibrations. We will study two cases: free vibration with periodic forcing and damped vibration with periodic forcing.

Free vibration with periodic forcing Let us consider the free vibration with a periodic forcing

$$y'' + \omega_0^2 y = F_0 \cos(\Omega t).$$

We have two subcases.

Case 1. $\Omega \neq \omega_0$.

It is reasonable to guess that there is a special solution which is synchronized with the periodic external forcing. Thus, we try a special solution of the form $C\cos(\Omega t)$. By plugging into the equation, we can find the coefficient $C = F_0/(a(\Omega^2 - \omega_0^2))$. Thus, the function

$$y_p(t) = \frac{F_0}{a(\Omega^2 - \omega_0^2)}\cos(\Omega t)$$

is a special solution. Let us still abbreviate $F_0/(\Omega^2 - \omega_0^2)$ by C. The general solution can be expressed as

$$\begin{aligned}
y(t) &= C\cos(\Omega t) + A\cos(\omega_0 t) + B\sin(\omega_0 t) \\
&= C\cos((\omega_l - \omega_h)t) + A\cos((\omega_l + \omega_h)t) + B\sin((\omega_l + \omega_h)t) \\
&= C\left(\cos(\omega_l t)\cos(\omega_h t) + \sin(\omega_l t)\sin(\omega_h t)\right) \\
&\quad + A\left(\cos(\omega_l t)\cos(\omega_h t) - \sin(\omega_l t)\sin(\omega_h t)\right) \\
&\quad + B\left(\sin(\omega_l t)\cos(\omega_h t) + \cos(\omega_l t)\sin(\omega_h t)\right) \\
&= \left[(C + A)\cos(\omega_l t) + B\sin(\omega_l t)\right]\cos(\omega_h t) \\
&\quad + \left[B\cos(\omega_l t) + (C - A)\sin(\omega_l t)\right]\sin(\omega_h t) \\
&= \tilde{A}\cos(\omega_l t - \Omega_1)\cos(\omega_h t) + \tilde{B}\cos(\omega_l t - \Omega_2)\sin(\omega_h t),
\end{aligned}$$

where

$$\omega_h = \frac{\omega_0 + \Omega}{2},\ \omega_l = \frac{\omega_0 - \Omega}{2}$$

indicate low and high frequencies, respectively; and

$$(C + A, B) = \tilde{A}(\cos(\Omega_1), \sin(\Omega_1)),\ (C - A, B) = \tilde{B}(\cos(\Omega_2), \sin(\Omega_2)).$$

Let us take the case when $\Omega \sim \omega_0$. In this case,

$$C = \frac{F_0}{a(\Omega^2 - \omega_0^2)}$$

is very large, and hence $\tilde{A}$ is very large. We concentrate on the solution $y(t) = \tilde{A}\cos(\omega_l t - \Omega_1)\cos(\omega_h t)$. In this solution, we may view $\tilde{A}\cos(\omega_l t - \Omega_1)$ as the amplitude of the high frequency wave $\cos(\omega_h t)$. This amplitude itself is a low frequency wave, which is the *envelope* of the solution $y(t)$. We call it the *modulation wave*. This phenomenon occurs in acoustics when two tuning forks of nearly equal frequency are sound simultaneously.

Case 2. $\Omega = \omega_0$.

In this case, we try a special solution of this form:

$$y_p = Ct\cos(\omega_0 t) + Dt\sin(\omega_0 t).$$

By plugging into the equation, we find a special solution

$$y_p = Rt\sin(\omega_0 t),\ R := \frac{F_0}{2a\omega_0}$$

The general solution is

$$y(t) = R\,t\sin(\omega_0 t) + A\cos(\omega_0 t + \theta_0) \tag{2.22}$$

The amplitude of this solution increases linearly in time. Such a phenomenon is called *resonance*.

Damped vibrations with periodic forcing We consider a damped vibration system with periodic forcing:

$$y'' + by' + cy = F_0\cos(\Omega t).$$

To find a special solution for the inhomogeneous equation, we try

$$y_p = C\cos(\Omega t) + D\sin(\Omega t).$$

By plugging into the equation, we find

$$\begin{aligned} &-\Omega^2(C\cos(\Omega t) + D\sin(\Omega t)) + b\Omega(-C\sin(\Omega t) + D\cos(\Omega t)) \\ &+ c(C\cos(\Omega t) + D\sin(\Omega t)) = F_0\cos(\Omega t). \end{aligned}$$

This yields

$$\begin{aligned} -\Omega^2 C + b\Omega D + cC &= F_0 \\ -\Omega^2 D - b\Omega C + cD &= 0 \end{aligned}$$

This solves C and D:

$$C = (c-\Omega^2)F_0/\Delta, \quad D = b\Omega F_0/\Delta,$$

where

$$\Delta = (c-\Omega^2)^2 + b^2\Omega^2.$$

Notice that $\Delta \neq 0$ whenever there is a damping. Let

$$A := \sqrt{C^2+D^2} = \frac{F_0}{\Delta},\ \Omega_0 = \arctan\left(\frac{-b\Omega}{c-\Omega^2}\right).$$

Then

$$\begin{aligned} y_p &= C\cos(\Omega t) + D\sin(\Omega t) \\ &= A\cos(\Omega_0)\cos(\Omega t) - A\sin(\Omega_0)\sin(\Omega t) \\ &= A\cos(\Omega t + \Omega_0) \end{aligned}$$

Thus, a special solution is again a cosine function with amplitude A and initial phase Ω_0. The general solution is

$$y(t) = A\cos(\Omega t + \Omega_0) + C_1 y_1(t) + C_2 y_2(t).$$

Notice that $y(t) \to A\cos(\Omega t + \Omega_0)$ as $t \to \infty$ because both $y_1(t)$ and $y_2(t)$ tend to 0 as $t \to \infty$. We call the solution $A\cos(\Omega t + \Omega_0)$ the steady-state solution or the forced response. This solution synchronized with the external periodic forcing.

Remarks.

- We notice that the amplitude A has maximum when $\Omega = \omega_0 := \sqrt{c}$, that is, the external forcing has the same period as the internal period ω_0.
- We also notice that $A \to \infty$ only when $b = 0$ (no damping) and $c = \Omega^2$. This is the resonance case. Otherwise, there is no resonance. In other word, general solutions approach the forced responsed solution, even in the case of resonance with damping.

Homework.

Find a special solution for the following equations

1. Compute the general solution of the given equation.
   (a) $y'' + 4y = 3\cos 2t$.
   (b) $y'' + 9y = \sin t + \sin 2t + \sin 3t$.
   (c) $y'' + 4y = \cos^2 t$.
2. Solve the initial value problem $y'' + 4y = 3\cos 2t + \cos t, y(0) = 2, y'(0) = 1$.
3. Consider the ODE $y'' + \omega_0^2 y = \cos \omega t$ with $\omega \sim \omega_0$, say $\omega = \omega_0 + \Delta\omega$. For each $\Delta\omega$, find a particular solution of this equation so that its limit approaches the resonant solution as $\Delta\omega \to 0$.

2.5 2×2 linear systems

Second-order ODE can be reduced to a 2×2 first-order system A general high-order ODE can be reduced to a system of first-order equations by introducing high derivatives as new unknowns. For example, the linear second-order ODE

$$ay'' + by' + cy = f \tag{2.23}$$

can be rewritten as

$$\begin{cases} y' &= v \\ av' &= -bv - cy + f \end{cases} \tag{2.24}$$

If (y, v) is a solution of this first-order system (2.24), then from $av' = -bv - cy + f$, we get v' is continuous. From first equation, we have $y'' = v'$. Hence, $y \in C^2$. Combining the two equations of (2.24), we conclude that y satisfies $ay'' + by' + cy = f$. Conversely, if y satisfies (2.23), then y is twice differentiable. Let us name $y' = v$. Then $v' = y''$. From (2.23), $av' + bv + cy = f$. Hence, these two equations are equivalent.

In the LRC circuit system, we have Kirchhoff voltage law which states the sums of potential drops along a closed cycle is zero. If the loop consists of an inductor ($V_i = LI'$), a resistor ($V_r = IR$) and a capacitor ($V_c = Q/C$). The equation is

$$LI' + IR + Q/C = f.$$

Here, f is the potential drop from battery. If we introduce a new variable V for the potential drop across the capacitor, call it V, we have $V = Q/C$. Let us differentiate it in t and get $CV' = I$. Then, we obtain two equations for I and V:

$$\begin{cases} LI' + RI + V = f \\ CV' - I = 0. \end{cases}$$

This is equivalent to the second order equation

$$LQ'' + RQ' + Q/C = f.$$

Solution space In this subsection, we shall study general solutions of 2×2 linear homogeneous equations

$$\mathbf{y}' = \mathbf{A}\mathbf{y} \tag{2.25}$$

where

$$\mathbf{y} = \begin{pmatrix} y^1 \\ y^2 \end{pmatrix}, \ \mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

and the stability of its critical state $\mathbf{0}$. First we notice that the solution space is linear space because if $\mathbf{y}(\cdot)$ and $\mathbf{z}(\cdot)$ are solutions, so is their linear combination $a\mathbf{y}(\cdot) + b\mathbf{z}(\cdot)$. To determine the dimension of the solution space, we notice that all solutions are uniquely determined by their initial data (the *existence and uniqueness theorem*),

$$\mathbf{y}(0) = \mathbf{y}_0 \in \mathbb{C}^2.$$

Thus, there is a 1-1 correspondence between $\mathbb{C}^2$ and the solution space $\mathcal{S}_0$ by the mapping: $\mathbf{y}(0) \mapsto \mathbf{y}(\cdot)$.

Theorem 2.1. *The solution space $\mathcal{S}_0$ for equation (2.25) is a two-dimensional vector space.*

2.5.1 Independence and Wronskian

In the solution space $\mathcal{S}_0$, two solutions $\mathbf{y}_1$ and $\mathbf{y}_2$ are called independent if $C_1\mathbf{y}_1(t) + C_2\mathbf{y}_2(t) = 0$ implies $C_1 = C_2 = 0$. This definition is for all t, but based on the uniqueness theorem, we only need to check this condition at just one point. We have the following theorem.

Theorem 2.2. *Suppose $\mathbf{y}_1$ and $\mathbf{y}_2$ are solutions of (2.25). If $\mathbf{y}_1(t_0)$ and $\mathbf{y}_2(t_0)$ are independent in $\mathbb{R}^2$ ($\mathbb{C}^2$), then $\mathbf{y}_1(t)$ and $\mathbf{y}_2(t)$ are independent in $\mathbb{R}^2$ ($\mathbb{C}^2$) for all t.*

Proof. Let t_1 be a point lying in the maximal interval of existence containing t_0. Suppose $\mathbf{y}_1(t_1)$ and $\mathbf{y}_2(t_1)$ are linearly dependent, then there exist constants C_1 and C_2 such that

$$C_1\mathbf{y}_1(t_1) + C_2\mathbf{y}_2(t_1) = 0.$$

Let $\mathbf{y} = C_1\mathbf{y}_1 + C_2\mathbf{y}_2$. Notice that both $\mathbf{y}$ and the zero constant solution have the same value at t_1. By the uniqueness theorem, $\mathbf{y} \equiv 0$ on the maximal interval of existence containing t_1, hence, containg t_0. This contradicts to $\mathbf{y}_1(t_0)$ and $\mathbf{y}_2(t_0)$ being independent. □

Definition 2.1. *Given any two solutions* $\mathbf{y}_1$ *and* $\mathbf{y}_2$*, we the Wronskian*

$$W(\mathbf{y}_1, \mathbf{y}_2)(t) = det(\mathbf{y}_1(t), \mathbf{y}_2(t)) = \begin{vmatrix} y_{1,1} & y_{2,1} \\ y_{1,2} & y_{2,2} \end{vmatrix} \tag{2.26}$$

The Wronskian is used to test the independence of $\mathbf{y}_1$ and $\mathbf{y}_2$.

Theorem 2.3. *Let* $\mathbf{y}_1$ *and* $\mathbf{y}_2$ *be two solutions of (2.25). Let us abbreviate the Wronskian* $W(\mathbf{y}_1, \mathbf{y}_2)(t)$ *by* $W(t)$*. We have*

(i)

$$\frac{dW}{dt} = (tr\mathbf{A})W$$

(ii) $W(t_0) \neq 0$ *for some* t_0 *if and only if* $W(t) \neq 0$ *for all* t*.*

Proof. Let $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2)$. Then we have

$$\mathbf{Y}' = \mathbf{AY}.$$

The Wronskian $W(t)$ is $det\mathbf{Y}(t)$. We differentiate W in t, We get

$$\begin{aligned} W' &= y'_{1,1}y_{2,2} - y'_{1,2}y_{2,1} - y'_{2,1}y_{1,2} + y'_{2,2}y_{1,1} \\ &= \sum_k (a_{1,k}y_{k,1}y_{2,2} - a_{1,k}y_{k,2}y_{2,1} - a_{2,k}y_{k,1}y_{1,2} + a_{2,k}y_{k,2}y_{1,1}) \\ &= (a_{1,1} + a_{2,2})(y_{1,1}y_{2,2} - y_{1,2}y_{2,1}) \\ &= tr(A)W \end{aligned}$$

Since $W(t) = W(t_0)\exp(tr(A)(t - t_0))$, we see that $W(t_0) \neq 0$ if and only if $W(t) \neq 0$. □

Remark This theorem is also true for $n \times n$ system. Namely, if $\mathbf{Y}'(t) = \mathbf{AY}(t)$, and $W(t) = det\mathbf{Y}(t)$, then

$$W'(t) = (tr\mathbf{A})W(t).$$

You can try to prove this theorem by using the determinant formula

$$det\mathbf{Y} = \sum_i y_{ij}C_{ij}, \quad \sum_k y_{kj}C_{ij} = 0 \text{ if } i \neq k,$$

where C_{ij} is called the cofactor of $\mathbf{Y}$, which is $(-1)^{i+j}det(\mathbf{Y}_{ij})$, and $\mathbf{Y}_{ij}$ is the $(n-1) \times (n-1)$ matrix obtained by eliminating the ith row and jth column from $\mathbf{Y}$.

2.5.2 Finding the fundamental solutions and Phase Portrait

In this subsection, we look for find two independent solutions for the homogeneous equation

$$\mathbf{y}'(t) = \mathbf{A}\mathbf{y}(t).$$

We try a solution of the form $\mathbf{y}(t) = e^{\lambda t}\mathbf{v}$, where $\mathbf{v} \in \mathbb{C}^2$ is a constant vector. Plugging into (2.25), we get

$$\lambda \mathbf{v} e^{\lambda t} = \mathbf{A}\mathbf{v}e^{\lambda t}.$$

We find that $\mathbf{y}(t) = e^{\lambda t}\mathbf{v}$ is a solution of (2.25) if and only if

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}. \tag{2.27}$$

That is, λ is the eigenvalue and $\mathbf{v}$ is the corresponding eigenvector. The eigenvalue λ satisfies the following characteristic equation

$$\det(\lambda \mathbf{I} - \mathbf{A}) = 0.$$

In two dimensions, this is

$$\lambda^2 - T\lambda + D = 0,$$

where

$$T = a + d,\ D = ad - bc$$

are the trace and determinant of $\mathbf{A}$, respectively. The eigenvalues are

$$\lambda_1 = \frac{T + \sqrt{T^2 - 4D}}{2},\ \lambda_2 = \frac{T - \sqrt{T^2 - 4D}}{2}.$$

There are three cases for the eigenvalues:

- Case 1: $T^2 - 4D > 0$. Then $\lambda_1 \neq \lambda_2$ and are real.
- Case 2: $T^2 - 4D < 0$. Then λ_1, λ_2 and are complex conjugate.
- Case 3: $T^2 - 4D = 0$. Then λ_1 is a double root.

Case 1. Both λ_1 and λ_2 are real

Suppose the two corresponding real eigenvectors are $\mathbf{v}_1$ and $\mathbf{v}_2$.

1. **Finding fundamental solutions** The corresponding two independent solutions are

   $$\mathbf{y}_1 = e^{\lambda_1 t}\mathbf{v}_1,\ \mathbf{y}_2 = e^{\lambda_2 t}\mathbf{v}_2.$$

   A general solution has the form

   $$\mathbf{y}(t) = C_1\mathbf{y}_1(t) + C_2\mathbf{y}_2(t)$$

Let us denote the fundamental solutions by a matrix

$$\Phi(t) = [\mathbf{y}_1(t), \mathbf{y}_2(t)],$$

and the solution $\mathbf{y}(\cdot)$ can be expressed as

$$\mathbf{y}(t) = \Phi(t)\mathbf{C}, \quad \mathbf{C} = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}.$$

If the initial data is $\mathbf{y}_0$, then

$$\mathbf{y}(0) = \Phi(0)\mathbf{C} = \mathbf{y}_0.$$

We get

$$\mathbf{C} = \Phi^{-1}(0)\mathbf{y}_0.$$

Thus, general solution is

$$\mathbf{y}(t) = \Phi(t)\Phi^{-1}(0)\mathbf{y}_0 = C_1 e^{\lambda_1 t}\mathbf{v}_1 + C_2 e^{\lambda_2 t}\mathbf{v}_2. \tag{2.28}$$

2. **Phase Portrait** In the solution expression (2.28), we may call

$$\mathbf{y}(t) = \eta_1(t)\mathbf{v}_1 + \eta_2(t)\mathbf{v}_2,$$

where

$$\eta_1(t) = C_1 e^{\lambda_1 t}, \quad \eta_2(t) = C_2 e^{\lambda_2 t}.$$

By taking $\ln$, we can eliminate t from η_1 and η_2 to get an implicit expression for this trajectories:

$$\frac{1}{\lambda_1}\ln|\eta_1| = \frac{1}{\lambda_2}\ln|\eta_2| + C,$$

where C is a constant depending on C_1, C_2. We can also express them as

$$|\eta_2|^{1/\lambda_2} = C\,|\eta_1|^{1/\lambda_1}$$

where C is another constant. From this, we can draw a family of solution trajectories. For instant, if $\lambda_1 = -1$, $\lambda_2 = 1$, then the solution curves are

$$|\eta_2| = C|\eta_1|^{-1}.$$

These are hyperbolae. We will see more examples below.

3. **Stability of the 0 state** The $\mathbf{0}$ state is an equilibrium (i.e. a constant state solution). Its behavior is determined by the sign of the eigenvalues λ_1, λ_2:

   - $\lambda_1, \lambda_2 < 0$: all solutions tend to 0 as $t \to \infty$. We call 0 state a sink. It is a *stable* equilibrium.
   - $\lambda_1, \lambda_2 > 0$: all solutions tend to infinity as $t \to \infty$. In fact, all solutions tend to the 0 state as $t \to -\infty$. We call 0 state a source. It is an *unstable* equilibrium.

- $\lambda_1 \cdot \lambda_2 < 0$. Let us take $\lambda_1 < 0$ and $\lambda_2 > 0$ as an example for explanation. A general solution has the form
$$\mathbf{y}(t) = C_1 e^{\lambda_1 t}\mathbf{v}_1 + C_2 e^{\lambda_2 t}\mathbf{v}_2.$$
We have three cases:
  - If $\mathbf{y}(0) \in \mathcal{M}_s := \{\gamma \mathbf{v}_1, \gamma \in \mathbb{R}\}$, then the corresponding $C_2 = 0$, and $\mathbf{y}(t) \to 0$ as $t \to \infty$. We call the line $\mathcal{M}_s$ a stable manifold.
  - If $\mathbf{y}(0) \in \mathcal{M}_u := \{\gamma \mathbf{v}_2, \gamma \in \mathbb{R}\}$, then the corresponding $C_1 = 0$ and $\mathbf{y}(t) \to 0$ as $t \to -\infty$. We call the line $\mathcal{M}_u$ an unstable manifold.
  - For any other $\mathbf{y}_0$, the corresponding $\mathbf{y}(t)$ has the following asymptotics:
$$\mathbf{y}(t) \to \mathbf{v}_1\text{-axis, as } t \to -\infty,$$
$$\mathbf{y}(t) \to \mathbf{v}_2\text{-axis, as } t \to +\infty.$$
That is, all solutions approach the stable manifold as $t \to \infty$ and the unstable manifold as $t \to -\infty$.

  The $\mathbf{0}$ state is the intersection of the stable and unstable manifolds. It is called a *saddle point*.
- $\lambda_1 = 0$ and $\lambda_2 \neq 0$. In this case, a general solution has the form:
$$\mathbf{y}(t) = C_1 \mathbf{v}_1 + C_2 e^{\lambda_2 t}\mathbf{v}_2.$$
The equilibrium $\{\bar{\mathbf{y}} | A\bar{\mathbf{y}} = 0\}$ is a line: $\{C_1\mathbf{v}_1 | C_1 \in \mathbb{R}\}$.
  - If $\lambda_2 < 0$, then all solutions approach $C_1\mathbf{v}_1$. This means that the line $C_1\mathbf{v}_1$ is a stable line.
  - If $\lambda_2 > 0$, then all solutions leave $C_1\mathbf{v}_1$. This means that the line $C_1\mathbf{v}_1$ is an unstable line.

Examples

1. Consider
$$\mathbf{y}' = \mathbf{A}\mathbf{y}, \quad \mathbf{A} = \begin{pmatrix} 1 & 1 \\ 4 & 1 \end{pmatrix}.$$
The corresponding characteristic equation is
$$\det(\lambda\mathbf{I} - \mathbf{A}) = (\lambda - 1)^2 - 4 = 0.$$
Hence, the two eigenvalues are
$$\lambda_1 = 3, \ \lambda_2 = -1.$$
The eigenvector $\mathbf{v}_1$ corresponding to $\lambda_1 = 3$ satisfies
$$(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{v}_1 = 0.$$

This gives

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

Similarly, the eigenvector corresponding to $\lambda_2 = -1$ is

$$\mathbf{v}_2 = \begin{pmatrix} 1 \\ -2 \end{pmatrix}.$$

A general solution has the form

$$\mathbf{y}(t) = C_1 e^{3t}\mathbf{v}_1 + C_2 e^{-t}\mathbf{v}_2.$$

2. Consider

$$\mathbf{y}' = \mathbf{A}\mathbf{y}, \;\; \mathbf{A} = \begin{pmatrix} 8 & -11 \\ 6 & -9 \end{pmatrix}.$$

The eigenvalues of $\mathbf{A}$ are roots of the characteristic equation $\det(\lambda\mathbf{I} - \mathbf{A}) = 0$. This yields two eigenvalues $\lambda_1 = -3$ and $\lambda_2 = 2$. The corresponding eigenvectors satisfy $(\mathbf{A} - \lambda_i)\mathbf{v}_i = 0$. For $\mathbf{v}_1$, we have

$$\begin{pmatrix} 8+3 & -11 \\ 6 & -9+3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

This yields

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Similarly, we obtain

$$\mathbf{v}_2 = \begin{pmatrix} 11 \\ 6 \end{pmatrix}.$$

The general solution is

$$\mathbf{y}(t) = C_1 e^{-3t}\mathbf{v}_1 + C_2 e^{2t}\mathbf{v}_2.$$

The line in the direction of $\mathbf{v}_1$ is a stable manifold, whereas the line in $\mathbf{v}_2$ direction is a unstable manifold. The origin is a saddle point.

3. Consider

$$\mathbf{y}' = \mathbf{A}\mathbf{y}, \;\; \mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}.$$

The eigenvalues of $\mathbf{A}$ are $\lambda_1 = 0$ and $\lambda_2 = -5$. The corresponding eigenvectors are

$$\mathbf{v}_1 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

The general solutions are $\mathbf{y}(t) = C_1\mathbf{v}_1 + C_2 e^{-5t}\mathbf{v}_2$. All solutions approach the line $C_1\mathbf{v}_1$.

Case 2. λ_i **are complex conjugate.**

$$\lambda_1 = \alpha + i\omega,\ \lambda_2 = \alpha - i\omega.$$

Since $\mathbf{A}$ is real-valued, the corresponding eigenvectors are also complex conjugate:

$$\mathbf{w}_1 = \mathbf{u} + i\mathbf{v},\ \mathbf{w}_2 = \mathbf{u} - i\mathbf{v}.$$

We have two independent complex-valued solutions: $\mathbf{z}_1 = e^{\lambda_1 t}\mathbf{w}_1$ and $\mathbf{z}_2 = e^{\lambda_2 t}\mathbf{w}_2$.

1. **Finding real fundamental solutions** Since our equation (2.25) has real coefficients, its real-valued solution can be obtained by taking the real part (or pure imaginary part) of the complex solution. In fact, suppose $\mathbf{z}(t) = \mathbf{x}(t) + i\mathbf{y}(t)$ is a complex solution of the real-value ODE (2.25). Then
$$\frac{d}{dt}\left(\mathbf{x}(t) + i\mathbf{y}(t)\right) = \mathbf{A}\left(\mathbf{x}(t) + i\mathbf{y}(t)\right).$$
By taking the real part and the imaginary part, using the fact that $\mathbf{A}$ is real, we obtain
$$\frac{d\mathbf{x}}{dt} = \mathbf{A}\mathbf{x}(t),\ \frac{d\mathbf{y}}{dt} = \mathbf{A}\mathbf{y}(t)$$
Hence, both the real part and the imaginary part of $\mathbf{z}(t)$ satisfy the equation.

   Now, let us take the real part and the imaginary part of one of the above solution:
$$\mathbf{z}_1(t) = \left(e^{\alpha t}(\cos\omega t + i\sin\omega t)\right)\ (\mathbf{u} + i\mathbf{v})$$
Its real part and imaginary part are respectively
$$\begin{aligned}\mathbf{y}_1(t) &= e^{\alpha t}\left(\cos\omega t\mathbf{u} - \sin\omega t\mathbf{v}\right)\\ \mathbf{y}_2(t) &= e^{\alpha t}\left(\sin\omega t\mathbf{u} + \cos\omega t\mathbf{v}\right).\end{aligned}$$
The other solution $\mathbf{z}_2$ is the complex conjugate of $\mathbf{z}_1$. We will get the same real solutions from taking the real and imaginary parts of $\mathbf{z}_2$.

   You may wonder now whether $\mathbf{u}$ and $\mathbf{v}$ are independent. Indeed, if $\mathbf{v} = c\mathbf{u}$ for some $c \in \mathbb{R}$, then
$$\mathbf{A}(\mathbf{u} + i\mathbf{v}) = \lambda_1(\mathbf{u} + i\mathbf{v})$$
gives
$$\mathbf{A}(1 + ic)\mathbf{u} = \lambda_1(1 + ic)\mathbf{u}$$
$$\mathbf{A}\mathbf{u} = \lambda_1\mathbf{u} = (\alpha + i\omega)\mathbf{u}$$
This yields
$$\mathbf{A}\mathbf{u} = \alpha\mathbf{u},\ \text{and}\ \omega\mathbf{u} = 0,$$
because $\mathbf{A}$ is real. This implies $\omega = 0$ if $\mathbf{u} \neq 0$. This contradicts to that the eigenvalue λ_1 has nontrivial imaginary part. This shows that $\mathbf{u}$ and $\mathbf{v}$ are independent.

   From the independence of $\mathbf{u}$ and $\mathbf{v}$, we conclude that $\mathbf{y}_1$ and $\mathbf{y}_2$ are also independent, and they constitute a basis in the solution space $\mathcal{S}_0$.

2. **Phase portrait** A general solution is given by

$$\begin{aligned}\mathbf{y}(t) &= C_1\mathbf{y}_1(t) + C_2\mathbf{y}_2(t)\\ &= C_1e^{\alpha t}\left(\cos\omega t\mathbf{u} - \sin\omega t\mathbf{v}\right) + C_2e^{\alpha t}\left(\sin\omega t\mathbf{u} + \cos\omega t\mathbf{v}\right)\\ &= e^{\alpha t}\left((C_1\cos\omega t + C_2\sin\omega t)\mathbf{u} + (C_2\cos\omega t - C_1\sin\omega t)\mathbf{v}\right)\\ &= Ae^{\alpha t}\left(\cos(\omega t - \omega_0)\mathbf{u} + \sin(\omega t - \omega_0)\mathbf{v}\right),\end{aligned}$$

where $(C_1, C_2) = A(\cos\omega_0, \sin\omega_0)$.

- When $\alpha = 0$, these are circles (ellipses);
- When $\alpha \neq 0$, the trajectories are spirals.

3. **Stability of the 0 state.** There are three cases for the structure of the solutions.

- $\alpha = 0$: The eigenvalues are pure imaginary. All solutions are ellipses.
- $\alpha < 0$: The solution are spirals and tend to 0 as $t \to \infty$. The 0 state is a *spiral sink*.
- $\alpha > 0$: The solution are spirals and tend to 0 as $t \to -\infty$. The 0 state is a *spiral source*.

Example

1. Consider the matrix

$$\mathbf{A} = \begin{pmatrix} 2 & 1\\ -4 & -1\end{pmatrix},$$

The characteristic equation is $\det(\lambda\mathbf{I} - \mathbf{A}) = \lambda^2 - \lambda - 2 = 0$. The roots are $\lambda_1 = (1 + i\sqrt{7})/2$ and $\lambda_2 = (1 - i\sqrt{7})/2$. The corresponding eigenvectors are

$$\mathbf{v}_1 = \begin{pmatrix} -2\\ 3 - i\sqrt{7}\end{pmatrix} := \mathbf{u} + i\mathbf{w},\ \mathbf{v}_2 = \begin{pmatrix} -2\\ 3 + i\sqrt{7}\end{pmatrix} := \mathbf{u} - i\mathbf{w}.$$

$$\mathbf{u} = \begin{pmatrix} -2\\ 3\end{pmatrix},\quad \mathbf{w} = \begin{pmatrix} 0\\ -\sqrt{7}\end{pmatrix}.$$

We get two complex-valued solutions $\mathbf{z}_1 = e^{\lambda_1 t}\mathbf{v}_1$ and $\mathbf{z}_2 = e^{\lambda_2 t}\mathbf{v}_2$. The real solutions are their real parts and imaginary parts. They are

$$\begin{aligned}\mathbf{y}_1 &= e^{t/2}\left(\cos(\omega t)\mathbf{u} - \sin(\omega t)\mathbf{w}\right),\\ \mathbf{y}_2 &= e^{t/2}\left(\sin(\omega t)\mathbf{u} + \cos(\omega t)\mathbf{w}\right),\end{aligned}$$

where $\omega = \sqrt{7}/2$. The general solutions are spirals leaving from $\mathbf{0}$. The $\mathbf{0}$ is an unstable state.

Case 3. $\lambda_1 = \lambda_2$ are real and there is only one eigenvector.

Let us see some examples first to get some intuition how to find fundamental solutions.

Examples

1. Consider the ODE
$$\mathbf{y}' = \mathbf{A}\mathbf{y}, \quad \mathbf{A} = \begin{pmatrix} r & 1 \\ 0 & r \end{pmatrix},$$
where r is a constant. The eigenvalue of $\mathbf{A}$ is r and the corresponding eigenvector is
$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$
The y_2 component satisfies the single equation
$$y_2' = ry_2.$$
We obtain $y_2(t) = C_2e^{rt}$. By plugging this into the first equation
$$y_1' = ry_1 + C_2e^{rt},$$
we find $y_1(t) = C_2te^{rt}$ is a special solution. The general solution of y_1 is
$$y_1(t) = C_2te^{rt} + C_1e^{rt}.$$
We can express these general solutions in vector form:
$$\begin{aligned} \mathbf{y}(t) &= C_1e^{rt}\begin{pmatrix} 1 \\ 0 \end{pmatrix} + C_2\left[e^{rt}\begin{pmatrix} 0 \\ 1 \end{pmatrix} + te^{rt}\begin{pmatrix} 1 \\ 0 \end{pmatrix}\right] \\ &= C_1\mathbf{y}_1(t) + C_2\mathbf{y}_2(t), \end{aligned}$$
where
$$\mathbf{y}_1(t) = e^{rt}\mathbf{e}_1, \quad \mathbf{y}_2(t) = te^{rt}\mathbf{e}_1 + e^{rt}\mathbf{e}_2.$$

2. Consider the ODE
$$\mathbf{y}' = \mathbf{A}\mathbf{y}, \quad \mathbf{A} = \begin{pmatrix} 1 & -1 \\ 1 & 3 \end{pmatrix}.$$
The characteristic equation
$$0 = \det(\lambda\mathbf{I} - \mathbf{A}) = (\lambda - 1)(\lambda - 3) + 1 = (\lambda - 2)^2.$$
has a double root $\lambda = 2$. The corresponding eigenvector satisfies
$$(\mathbf{A} - 2\mathbf{I})\mathbf{v} = 0$$
$$\begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} v^1 \\ v^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

This yields a solution, called $\mathbf{v}_1$:

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

This is the only eigenvector. The solution $e^{2t}\mathbf{v}_1$ is a solution of the ODE. To find the other independent solution, we expect that there is a *resonant* solution te^{2t} in the direction of $\mathbf{v}_1$. Unfortunately, $te^{2t}\mathbf{v}_1$ is not a solution unless $\mathbf{v}_1 = 0$. Therefore, let us try another kind of solution

$$\mathbf{y}(t) = te^{2t}\mathbf{v}_1 + e^{\mu t}\mathbf{v}_2,$$

for some unknown vector $\mathbf{v}_2$. We plug it into the equation $\mathbf{y}' = \mathbf{A}\mathbf{y}$ to find $\mathbf{v}_2$:

$$\mathbf{y}' = (e^{2t} + 2te^{2t})\mathbf{v}_1 + \mu e^{\mu t}\mathbf{v}_2,$$

we obtain

$$2\mathbf{v}_1 te^{2t} + \mathbf{v}_1 e^{2t} + \mu e^{\mu t}\mathbf{v}_2 = \mathbf{A}(\mathbf{v}_1 te^{2t} + \mathbf{v}_2 e^{\mu t})$$

Using $\mathbf{A}\mathbf{v}_1 = 2\mathbf{v}_1$, we get

$$\mathbf{v}_1 e^{2t} + \mu e^{\mu t}\mathbf{v}_2 = \mathbf{A}\mathbf{v}_2 e^{\mu t}$$

This should be valid for all t. Hence, we get $\mu = 2$ and

$$(\mathbf{A} - 2\mathbf{I})\mathbf{v}_2 = \mathbf{v}_1.$$

That is

$$\begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} v^1 \\ v^2 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

This gives $v^1 + v^2 = -1$. So,

$$\mathbf{v}_2 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}.$$

is a solution.

Now, we find two solutions

$$\begin{aligned} \mathbf{y}_1 &= e^{2t}\mathbf{v}_1 \\ \mathbf{y}_2 &= te^{2t}\mathbf{v}_1 + e^{2t}\mathbf{v}_2. \end{aligned}$$

Now, let us explain general theory.

1. **Finding fundamental solutions** The double root case can be thought as a limiting case of two distinguished roots λ_1 and λ_2 with $\lambda_2 \to \lambda_1$. In this limiting process,

$$\frac{1}{\lambda_2 - \lambda_1}\left(e^{\lambda_2 t}\mathbf{v}_2(\lambda_2) - e^{\lambda_1 t}\mathbf{v}_1\right)$$

is a solution for all λ_1 and λ_2. We fix λ_1 and let $\lambda_2 \to \lambda_1$. The eigenvector $\mathbf{v}_2$ depends on λ_2. This limiting process is equivalent to differentiate $e^{\lambda t}\mathbf{v}(\lambda)$ in λ at λ_1, where $\mathbf{v}(\lambda)$ is the eigenvector corresponding to λ. This derivative is

$$\frac{d}{d\lambda}\left(e^{\lambda t}\mathbf{v}(\lambda)\right) = te^{\lambda_1 t}\mathbf{v}_1 + e^{\lambda_1 t}\frac{\partial \mathbf{v}}{\partial \lambda}.$$

The new vector $\frac{\partial \mathbf{v}}{\partial \lambda}$ is denoted by $\mathbf{v}_2$. By plugging $te^{\lambda_1 t}\mathbf{v}_1 + e^{\lambda_1 t}\mathbf{v}_2$ into the equation, we conclude that $\mathbf{v}_2$ should satisfies

$$(\mathbf{A} - \lambda_1 \mathbf{I})\mathbf{v}_2 = \mathbf{v}_1.$$

The solvability of $\mathbf{v}_2$ comes from the follows. Let $\mathcal{N}_k$ be the kernel (null space) of $(\mathbf{A}-\lambda_1\mathbf{I})^k$, $k = 1, 2$. From the definition of $\mathcal{N}_k$, we have the following mapping chain

$$\mathcal{N}_2 \stackrel{\mathbf{A}-\lambda_1\mathbf{I}}{\longrightarrow} \mathcal{N}_1 \stackrel{\mathbf{A}-\lambda_1\mathbf{I}}{\longrightarrow} \{0\}.$$

Since $\mathbf{v}_1$ is the only eigenvecto, we thus have $\mathcal{N}_1 =< \mathbf{v}_1 >$, the span of $\mathbf{v}_1$. In the map $\mathbf{A} - \lambda_1\mathbf{I} : \mathcal{N}_2 \to \mathcal{N}_1$, the domain space is $\mathcal{N}_2 = \mathbb{R}^2$ from Caley-Hamilton theorem. [1] We have seen that the kernel is $\mathcal{N}_1$, which has dimension 1. Therefore the range space has dimension 1. Here, we use a theorem of linear map: the sum of the dimensions of range and kernel spaces equals the dimension of the domain space. We conclude that the range $(\mathbf{A} - \lambda_1\mathbf{I})\mathcal{N}_2$ has to be $\mathcal{N}_1$. Therefore, there exists a $\mathbf{v}_2 \in \mathcal{N}_2$ such that

$$(\mathbf{A} - \lambda_1 \mathbf{I})\mathbf{v}_2 = \mathbf{v}_1.$$

The matrix $\mathbf{A}$, as represented in the basis $\mathbf{v}_1$ and $\mathbf{v}_2$, has the form

$$\mathbf{A}[\mathbf{v}_1, \mathbf{v}_2] = [\mathbf{v}_1, \mathbf{v}_2]\begin{pmatrix} \lambda_1 & 1 \\ 0 & \lambda_1 \end{pmatrix}$$

This is called the Jordan canonical form of $\mathbf{A}$. We can find two solutions from this form:

$$\begin{aligned} \mathbf{y}_1(t) &= e^{\lambda_1 t}\mathbf{v}_1, \\ \mathbf{y}_2(t) &= te^{\lambda_1 t}\mathbf{v}_1 + e^{\lambda_1 t}\mathbf{v}_2 \end{aligned}$$

[1] The Caley-Hamilton theorem states that $\mathbf{A}$ satisfies the matrix equation:

$$p(\mathbf{A}) = 0.$$

This can be seen from the following argument. Let $\mathbf{Q}(\lambda)$ be the adjugate matrix of $\mathbf{A} - \lambda\mathbf{I}$, i.e.

$$\mathbf{Q}(\lambda) = \begin{pmatrix} d-\lambda & -b \\ -c & a-\lambda \end{pmatrix} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} - \lambda\mathbf{I}.$$

This adjugate matrix commutes with $\mathbf{A}$ (check by yourself). Further,

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{Q}(\lambda) = \mathbf{Q}(\lambda)(\mathbf{A} - \lambda\mathbf{I}) = p(\lambda)\mathbf{I}.$$

This is a polynomial in λ with matrix coefficients. The coefficients commute with $\mathbf{A}$. When we plug $\lambda = \mathbf{A}$, we immediately get $p(\mathbf{A}) = 0$.

You can check the Wronskian $W[\mathbf{y}_1, \mathbf{y}_2](t) \neq 0$. Thus, $\mathbf{y}_1$ and $\mathbf{y}_2$ form a fundamental solution. The general solution has the form

$$\mathbf{y}(t) = C_1\mathbf{y}_1(t) + C_2\mathbf{y}_2(t).$$

2. **Stability of 0 state** The stability of the $\mathbf{0}$ state (called the critical state) relies on the sign of λ_1. We have

   - $\lambda_1 < 0$: the $\mathbf{0}$ state is a stable equilibrium.
   - $\lambda_1 > 0$: the $\mathbf{0}$ state is an unstable equilibrium.
   - $\lambda_1 = 0$: the general solution reads

     $$\mathbf{y}(t) = C_2 t\mathbf{v}_2 + C_1\mathbf{v}_1,$$

     which tends to ∞ as $t \to \infty$. Therefore, the $\mathbf{0}$ state is "unstable."

Summary of Equilibria We can plot a stability diagram on the plane of the two parameters T and D, the trace and the determinant of $\mathbf{A}$:

$$T = a + d, \quad D = ad - bc.$$

The eigenvalues of $\mathbf{A}$ are

$$\lambda_1 = \frac{T + \sqrt{T^2 - 4D}}{2}, \ \lambda_2 = \frac{T - \sqrt{T^2 - 4D}}{2}.$$

Let $\Delta := T^2 - 4D$. On the T-D plane, the parabola $\Delta = 0$, the line $D = 0$ and the line $T = 0$ partition the plane into the following regions. The status of the origin is as the follows.

- $\Delta > 0$, $D < 0$, the origin is a saddle point.
- $\Delta > 0$, $D > 0$, $T > 0$, the origin is an unstable node (source).
- $\Delta > 0$, $D > 0$, $T < 0$, the origin is an stable node (sink).
- $\Delta < 0$, $T < 0$, the origin is a stable spiral point.
- $\Delta < 0$, $T > 0$, the origin is an unstable spiral point.
- $\Delta < 0$, $T = 0$, the origin is an stable center point.
- $\Delta = 0$, $T < 0$, the origin is a stable node.
- $\Delta = 0$, $T > 0$, the origin is an unstable node.
- $\Delta = 0$, $T = 0$, the origin is an unstable node.

Bifurcations

- The transition from $D < 0$ to $D > 0$, the eigenvalues change from opposite sign change to same sign sign, the origin changes from a saddle to a node. Such transition is called a saddle-node bifurcation.
- The transition from $T < 0$ to $T > 0$ in the region $D > 0$, the origin changes from stable spiral to unstable spiral. At $T = 0$, the origin is a center. Such transition is called an Andronov-Hopf bifurcation.

For detailed figure, see `http://www.scholarpedia.org/article/Equilibrium#Non-hyperbolic_Equilibria`

Homeworks.

1. Consider $\mathbf{A} = \begin{pmatrix} 1 & -2 \\ 3 & -4 \end{pmatrix}$. Find the exact solution of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ and analyze the stability of $\mathbf{0}$.
2. Consider $\mathbf{A} = \begin{pmatrix} 3 & 6 \\ -1 & -2 \end{pmatrix}$. Find the exact solution of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ and analyze the stability of $\mathbf{0}$.
3. Consider $\mathbf{A} = \begin{pmatrix} 1 & i \\ -i & 1 \end{pmatrix}$. Find the exact solution of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ and analyze the stability of $\mathbf{0}$.
4. Solve the circuit system
$$\begin{pmatrix} I \\ V \end{pmatrix}' = \begin{pmatrix} -\frac{R_1}{L} & -\frac{1}{L} \\ \frac{1}{C} & -\frac{1}{CR_2} \end{pmatrix} \begin{pmatrix} I \\ V \end{pmatrix}$$
and analyze the stability of the $\mathbf{0}$ state.
5. B-D, pp. 411: 25, 26, pp. 493: 19.

Chapter 3

Nonlinear systems in two dimensions

3.1 Three kinds of physical models

We shall introduce three kinds of physical models which are 2×2 nonlinear dynamical systems.

- Lotka-Velterra system
- Conservative mechanical system
- Dissipative mechanical system

3.1.1 Lotka-Volterra system

Predator-prey model

The populations of a predator and prey exhibit interesting periodic phenomenon. A simple example is the fox-rabbit system. Let $R(t)$ be the population of rabbit and $F(t)$ the population of fox. The model proposed by Lotka-Volterra reads

$$\begin{aligned} \dot{R} &= \alpha R - \beta RF \\ \dot{F} &= -\gamma F + \delta RF. \end{aligned}$$

Here,

- α the growth rate of rabbits,
- γ death rate of foxes,
- RF the interaction rate of rabbits and foxes
- βRF the amount of rabbits being eaten
- δRF the amount of foxes increase from eating rabbits

Examples of numerical values of the parameters are: $\alpha = 2$, $\beta = 1.2$, $\gamma = 1$, $\delta = 0.9$.

If we take the environmental constraint into account, the model for the rabbits should be changed to

$$\dot{R} = \alpha R\left(1 - \frac{R}{K}\right) - \beta RF.$$

An epidemic model

Consider the spread of a viral epidemic through an isolated population. Let $x(t)$ denote the number of susceptible people at time t, $y(t)$ the number of infected people. The epidemic model reads

$$\begin{aligned} \dot{x} &= 0.0003x - 0.005xy \\ \dot{y} &= -0.1y + 0.005xy \end{aligned}$$

The first equation means that the birth rate of susceptible people is 0.0003. Susceptible people are infected through interaction and the infected rate is proportional to xy. The second equation means that the death rate of infected people is 0.1. The infected rate is the same as that in the first equation.

Competitive Lotka-Volterra equation

This is a model for population dynamics of two species that competing same resource. Let x_1 and x_2 are the populations of two species that compete same resources. The model for each species follows the logistic equation. The competing model reads

$$\begin{aligned} \dot{x_1} &= r_1 x_1\left(1 - \frac{x_1}{K_1}\right) - \alpha_1 x_1 x_2 \\ \dot{x_2} &= r_2 x_2\left(1 - \frac{x_2}{K_2}\right) - \alpha_2 x_1 x_2 \end{aligned}$$

The quantity x_1x_2 is the interaction rate. It causes decreasing of population of each species due to competition. These decreasing rates are $\alpha_1 x_1 x_2$ and $\alpha_2 x_1 x_2$, respectively. Here $\alpha_1 > 0$, $\alpha_2 > 0$. As an example, we see two types of snails, the left-curling and the right-curling, compete the same resource. Because they are the same kind of snail, they have the same growth rate and carrying constant. Let us take $r_1 = r_2 = 1$ and $K_1 = K_2 = 1$. We take $\alpha_1 = \alpha_2 = a$. We will see later that the structure of the solutions is very different between $a < 1$ and $a > 1$. For further study, see
`https://en.wikipedia.org/wiki/Competitive_Lotka-Volterra_equations`
`http://www.scholarpedia.org/article/Predator-prey_model`

3.1.2 Conservative mechanical system

The Newton's mechanics read

$$m\ddot{x} = F(x).$$

When F has the form

$$F = -\nabla V(x).$$

The mechanical system is called conservative. The function V iOS called the potential. Typical examples are

- Spring-mass system: $V(x) = \frac{1}{2}kx^2$;
- Gravitational system: $V(x) = -\frac{G}{|x|}$;
- Simple pendulum: the unknown is θ, the equation is

  $$ml\ddot{\theta} = -mg\sin\theta.$$

  The potential V is $V(\theta) = -\frac{g}{l}\cos\theta$.
- Duffing oscillator: the potential is

  $$V(x) = -\frac{\delta}{2}x^2 + \frac{1}{4}x^4.$$

  The dynamics is described by

  $$\ddot{x} = -\nabla V(x). \tag{3.1}$$
- Cubic potential: we consider the same equation (3.1) with

  $$V(x) = \frac{1}{2}(-x^2 + x^3).$$

These conservative systems conserve energy. We will give more examples in later chapter.

3.1.3 Dissipative systems

In real world, many conservative mechanical systems are to ideal. Physically, there are some friction which dissipates energy. Here are some examples

- spring-mass system with damping:

  $$\ddot{x} + \gamma\dot{x} + \omega^2 x = f$$

  where $\gamma > 0$ is the damping coefficient.
- van der pol oscillator:
  This is a nonlinear LRC circuit, where the resistance is replaced by a vacuum tube (or a semiconductor). The I-V relation of a vacuum tube is

  $$V = \alpha\left(\frac{I^3}{3} - I\right).$$

  Thus, the circuit equation is modified to

  $$L\frac{dI}{dt} + \frac{Q}{C} + \alpha\left(\frac{I^3}{3} - I\right) = V(t).$$

In terms of I, we get

$$L\ddot{I} + \alpha(I^2 - 1)\dot{I} + \frac{I}{C} = f(t).$$

The term $\alpha(I^2 - 1)$ is a friction term when $|I| < 1$ and becomes an amplification term when $|I| > 1$. http://www.scholarpedia.org/article/Van_der_Pol_oscillator

- Duffing oscillator with damping:

$$\ddot{x} + \gamma\dot{-}\delta x + x^3 = f(t).$$

https://en.wikipedia.org/wiki/Duffing_equation. http://www.scholarpedia.org/article/Duffing_oscillator

3.2 Autonomous systems

We consider general system of the form

$$\begin{cases} \dot{x} &= f(x, y) \\ \dot{y} &= g(x, y) \end{cases} \tag{3.2}$$

We shall study the initial value problem for this system with initial data $(x(t_0), y(t_0)) = (x_0, y_0)$, where t_0 is the starting time. We may write this problem in vector form

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) \tag{3.3}$$

$$\mathbf{y}(t_0) = \mathbf{y}_0. \tag{3.4}$$

First, we have the standard existence and uniqueness theorems.

Theorem 3.1. *If* $\mathbf{f}$ *is continuously differentiable, then the initial value problem (3.3) and (3.4) has a unique solution for* t *in some small interval* $(t_0 - \delta, t_0 + \delta)$.

Notice that the vector field $\mathbf{f}(\mathbf{y})$ we consider here is independent of t explicitly. Such systems are called *autonomous systems.* For autonomous systems, we notice the following things.

- It is enough to study the initial value problems with $t_0 = 0$. For if $\mathbf{y}(t)$ is the solution with $\mathbf{y}(t_0) = \mathbf{y}_0$, then $\mathbf{z}(t) := \mathbf{y}(t - t_0)$ is the solution with $\mathbf{z}(0) = \mathbf{y}_0$, and $\mathbf{y}(\cdot)$ and $\mathbf{z}(\cdot)$ trace the same trajectory on the plane. We call such trajectories the orbits, the $\mathbf{y}$-plane the phase plane.
- Two orbits cannot intersect on the phase plane. This follows from the uniqueness theorem.
- An orbit cannot end in finite region unless its maximum interval of existence goes to infinity. This means that it is not possible to find a finite time T such that (i) $\mathbf{y}(\cdot)$ exists in $[0, T)$, (ii) $\mathbf{y}(\cdot)$ can not be extended beyond T, and $\{\mathbf{y}(t)|t \in [0, T)\}$ stays in finite region. For the limit $\lim_{t\to T-} \mathbf{y}(t)$ must exist and the existence theorem allows us to extend the solution beyond T. Therefore, we can only have either $\lim_{t\to T-} |\mathbf{y}(t)| = \infty$ or $T = \infty$.

Our goal is to characterize the orbital structure on the phase plane. There are some special orbits which play important roles in the characterization of the whole orbital structure. They are

(i) equilibria,

(ii) periodic orbits,

(iii) equilibria-connecting orbits: heteroclinic orbits, homoclinic orbits.

3.3 Equilibria and linearization

Definition 3.1. *A state $\bar{\mathbf{y}}$ is called an equilibrium of (3.3) if $\mathbf{f}(\bar{\mathbf{y}}) = 0$.*

The constant function $\mathbf{y}(t) \equiv \bar{\mathbf{y}}$ is a solution. We want to study the behaviors of solutions of (3.3) which take values near $\bar{\mathbf{y}}$. It is natural to take Taylor expansion of $\mathbf{y}$ about $\bar{\mathbf{y}}$. We have

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) = \mathbf{f}(\bar{\mathbf{y}}) + \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(\bar{\mathbf{y}})(\mathbf{y} - \bar{\mathbf{y}}) + O(|\mathbf{y} - \bar{\mathbf{y}}|^2).$$

Let $\mathbf{u} = \mathbf{y} - \bar{\mathbf{y}}$. Then $\mathbf{u}(t)$ satisfies

$$\dot{\mathbf{u}} = \mathbf{A}\mathbf{u} + \mathbf{g}(\mathbf{u}), \tag{3.5}$$

where

$$\mathbf{A} := \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(\bar{\mathbf{y}}), \ \mathbf{g}(\mathbf{u}) := \mathbf{f}(\bar{\mathbf{y}} + \mathbf{u}) - \mathbf{f}(\bar{\mathbf{y}}) - \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(\bar{\mathbf{y}})\,\mathbf{u} = O(|\mathbf{u}|^2).$$

System (3.5) is called the linearized equation (or the perturbed equation) of (3.3) about $\bar{\mathbf{y}}$. We have already known the structure of the linear equation

$$\dot{\mathbf{v}} = \mathbf{A}\mathbf{v}. \tag{3.6}$$

Do the orbits of (3.5) and (3.6) look "similar"?

3.3.1 Hyperbolic equilibria

Before answering the question in the last part of the above subsection, let us first study the following two examples to get feeling about perturbation.

Example 1: linear perturbation problem. We consider the following system

$$\dot{\mathbf{v}} = \mathbf{A}\mathbf{v}, \quad \mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}. \tag{3.7}$$

and its perturbation system

$$\dot{\mathbf{v}}_1 = \mathbf{A}_1\mathbf{v}_1, \tag{3.8}$$

with $\mathbf{A}_1 \sim \mathbf{A}$. We ask when do the solutions of (3.8) and (3.7) look similar? The quantitative behaviors of solutions of (3.7) are determined by the eigenvalues of $\mathbf{A}$. Namely,

$$\lambda_1 = \frac{1}{2}\left(T + \sqrt{T^2 - 4D}\right), \lambda_2 = \frac{1}{2}\left(T - \sqrt{T^2 - 4D}\right).$$

where $T = a + d$ and $D = ad - bc$. It is clear that λ_i are continuous in T and D, hence in a, b, c, d, or hence in $\mathbf{A}$. Thus, if we vary $\mathbf{A}$ slightly, then the change of λ_i is also small on the complex plane.

Now suppose

$$Re\lambda_i(\mathbf{A}) \neq 0, i = 1, 2. \tag{3.9}$$

Then this property is still satisfied for those $\mathbf{A}_1$ sufficiently close to $\mathbf{A}$.[1] The property (3.9) corresponds to that the zero state is a (spiral) source, a (spiral) sink, or a saddle. We conclude that *sink, source and saddle are persistent under small linear perturbation.*

Homework.

1. Suppose $Re\lambda_i(\mathbf{A}) \neq 0, i = 1, 2$. Let

$$\mathbf{A}_1 = \begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix}.$$

be a perturbation of $\mathbf{A}$. Find the condition on $\mathbf{A}_1$ so that

$$Re\lambda_i(\mathbf{A}_1) \neq 0, i = 1, 2.$$

Second order perturbation The above structure of trajectories near **0** is still valid for nonlinear perturbation. Let us consider

$$\begin{cases} \dot{x} &= r_1 x \\ \dot{y} &= r_2 y + \beta x^2. \end{cases} \tag{3.10}$$

The solution for $x(t)$ is

$$x(t) = x_0 e^{r_1 t}. \tag{3.11}$$

Plug this into the second equation, we get

$$\dot{y} = r_2 y + \beta x_0^2 e^{2r_1 t}.$$

Let us assume $r_2 \neq 2r_1$ to avoid the resonance situation. Then the general solution for $y(t)$ is

$$y(t) = Ae^{r_2 t} + Be^{2r_1 t}.$$

We plug this into the y-equation and obtain

$$Ar_2 e^{r_2 t} + 2r_1 Be^{2r_1 t} = r_2(Ae^{r_2 t} + Be^{2r_1 t}) + \beta x_0^2 e^{2r_1 t}.$$

[1] The eigenvalues λ_i are continuous function in T and D, or equivalently, a, b, c, d. If $Re\lambda_i \neq 0$ for a specific a, b, c, d, then a small perturbation of a, b, c, d still have nonzero $Re\lambda_i$.

This yields

$$2r_1 B = r_2 B + \beta x_0^2.$$

Thus, general solutions of $y(t)$ reads

$$y(t) = Ae^{r_2 t} + \frac{\beta x_0^2}{2r_1 - r_2} e^{2r_1 t}. \tag{3.12}$$

We see that the asymptotic behavior of $(x(t), y(t))$ is

- When $r_1 < 0$ and $r_2 < 0$, then $(x(t), y(t)) \to (0, 0)$ as $t \to \infty$. We call $(0, 0)$ a *sink*.
- When $r_1 > 0$ and $r_2 > 0$, then $(x(t), y(t)) \to (0, 0)$ as $t \to -\infty$. We call $(0, 0)$ a *source*.
- When $r_1 > 0$ and $r_2 < 0$, we have two subcases:
  - when $x_0 = 0$, then $(x(t), y(t)) \to (0, 0)$ as $t \to \infty$,
  - when $A = 0$, then $(x(t), y(t)) \to (0, 0)$ as $t \to -\infty$,

  The orbit with $x_0 = 0$ is called a *stable* manifold passing $(0, 0)$, while the orbit with $A = 0$ a *unstable* manifold. We denote the former one by M_s and the latter one by M_u. We call $(0, 0)$ a saddle point. By eliminate t from (3.11) and (3.12), we can obtain the equations of M_s and M_u as the follows.

$$\begin{aligned} M_s &: x = 0, \\ M_u &: y = \frac{\beta}{2r_1 - r_2} x^2. \end{aligned}$$

- When $r_1 < 0$ and $r_2 > 0$, $(0, 0)$ is a saddle point. The stable and unstable manifolds are

$$\begin{aligned} M_u &: x = 0, \\ M_s &: y = \frac{\beta}{2r_1 - r_2} x^2. \end{aligned}$$

General Theory for Hyperbolic Critical Points Let us go back to the general formulation (3.3). We have the following definitions.

Definition 3.2. *An equilibrium $\bar{\mathbf{y}}$ of (3.3) is called hyperbolic if all eigenvalues of the variation matrix $\mathbf{A} := \partial \mathbf{f} / \partial \mathbf{y}(\bar{\mathbf{y}})$ have only nonzero real parts.*

Definition 3.3. *An equilibrium $\bar{\mathbf{y}}$ of (3.3) is called*

- *a sink if $\mathbf{y}(t) \to \bar{\mathbf{y}}$ as $t \to \infty$,*
- *a source if $\mathbf{y}(t) \to \bar{\mathbf{y}}$ as $t \to -\infty$,*

where $\mathbf{y}(t)$ is any solution of (3.3) with $\mathbf{y}(0) \sim \bar{\mathbf{y}}$.

Definition 3.4. *1. A curve $M_s(\bar{\mathbf{y}})$ is called a stable manifold passing through the equilibrium $\bar{\mathbf{y}}$ if $\mathbf{y}(t) \to \bar{\mathbf{y}}$ as $t \to \infty$ for any solution $\mathbf{y}(t)$ with $\mathbf{y}(0) \in M_s(\bar{\mathbf{y}})$.*

2. A curve $M_u(\bar{\mathbf{y}})$ is called a unstable manifold passing through the equilibrium $\bar{\mathbf{y}}$ if $\mathbf{y}(t) \to \bar{\mathbf{y}}$ as $t \to -\infty$ for any solution $\mathbf{y}(t)$ with $\mathbf{y}(0) \in M_u(\bar{\mathbf{y}})$.

3. An equilibrium $\bar{\mathbf{y}}$ which is the intersection of a stable manifold and a unstable manifold is called a saddle point.

Theorem 3.2. *Consider the autonomous system (3.3) and its linearization (3.7) about an equilibrium. Suppose $\bar{\mathbf{y}}$ is hyperbolic. Then*

$$\bar{\mathbf{y}} \text{ is a } \left\{ \begin{array}{l} \text{source} \\ \text{sink} \\ \text{saddle} \end{array} \right\} \text{ of the nonlinear equation}$$

if and only if

$$\mathbf{0} \text{ is a } \left\{ \begin{array}{l} \text{source} \\ \text{sink} \\ \text{saddle} \end{array} \right\} \text{ of the linearized equation}$$

In other word, hyperbolicity is persistent under small perturbation.

Remark. The proof of this theorem is beyond the scope of this note, you may read Arnold's book for the proof. But if you have learned the existence theorem, then it is a modification of the existence theorem.

Remark. If an equilibrium $\bar{\mathbf{y}}$ is not hyperbolic, then the perturbation can break the local orbital structure. Let us see the following example. Consider

$$\left\{ \begin{array}{l} \dot{x} = y + \gamma \frac{(x^2+y^2)}{2} x \\ \dot{y} = -x + \gamma \frac{(x^2+y^2)}{2} y \end{array} \right.$$

When $\gamma = 0$, the orbits are circles with center at the origin. To see the effect of perturbation, we multiply the first equation by x and the second equation by y then add them together. We obtain

$$\dot{\rho} = \gamma \rho^2$$

where $\rho = x^2 + y^2$. The solution of $\rho(t)$ is

$$\rho(t) = \frac{1}{\rho(0)^{-1} - \gamma t}.$$

When $\gamma < 0$, the solution tends to 0 as $t \to \infty$. When $\gamma > 0$, the solution tends to zero as $t \to -\infty$. Moreover, the solution $\rho(t) \to \infty$ as $t \to \rho(0)^{-1}/\gamma$. Thus, the center becomes a sink if $\gamma < 0$ and a source when $\gamma > 0$.

In fact, we can solve this equation in polar coordinate. Let $\rho = x^2 + y^2$, $\tan\theta = y/x$. We have found the equation for ρ. The equation for θ is

$$\dot{\theta} = \frac{d}{dt}\tan^{-1}\left(\frac{y}{x}\right) = \frac{x\dot{y} - y\dot{x}}{x^2 + y^2}.$$

Plug the equation $\dot{x}, \dot{y}$ into θ equation, we get

$$\dot{\theta} = -1.$$

The solutions are spirals.

3.3.2 The equilibria in the competition model

Competition model The two-species competition model reads

$$\dot{x_1} = r_1 x_1\left(1 - \frac{x_1}{K_1}\right) - \alpha_1 x_1 x_2 = f_1(x_1, x_2)$$
$$\dot{x_2} = r_2 x_2\left(1 - \frac{x_2}{K_2}\right) - \alpha_2 x_1 x_2 = f_2(x_1, x_2).$$

Equilibria We will study the stability of its equilibria, which are the zeros of

$$f_1(x_1, x_2) = 0, \quad f_2(x_1, x_2) = 0.$$

The null line of the vector field in the x-direction are

$$r_1 x_1\left(1 - \frac{x_1}{K_1} - \frac{x_2}{L_1}\right) = 0,$$

where

$$L_1 = \frac{r_1}{\alpha_1}.$$

This yields

$$x_1 = 0,\ 1 - \frac{x_1}{K_1} - \frac{x_2}{L_1} = 0.$$

They are called the x-nullclines. Similarly, the y-nullclines are

$$x_2 = 0,\ 1 - \frac{x_2}{K_2} - \frac{x_1}{L_2} = 0,$$

where $L_2 = \frac{r_2}{\alpha_2}$.

Some biological relevant parameters The quantity $L_1 = r_1/\alpha_1$ measures the "competitive capacity" of species 1. The quantity L_1 is large means that r_1 is large (species 1 has large growth rate) or α_1 is small (it is less sensitive to competition from species 2). Let us define

$$s_1 = \frac{L_1}{K_2}, s_2 = \frac{L_2}{K_1}.$$

The quantity s_1 measures the competitive ratio of species 1 relative to the maximal population of species 2. $s_1 > 1$ means that species 1 is more competitive relative to the maximal population of species 2.

The intersection of a x-nullcline and a y-nullcline is an equilibrium. We are only interested in those equilibria in the first quadrant because x_i is the population of the i species which is non-negative. There are four cases.

- Case 1: $s_1 > 1$ and $s_2 < 1$ (species 1 is more competitive)
- Case 2: $s_1 < 1$ and $s_2 > 1$ (species 2 is more competitive)
- Case 3: $s_1 < 1$ and $s_2 < 1$ (both species are not competitive)
- Case 4: $s_1 > 1$ and $s_2 > 1$ (both species are competitive)

In the first two cases, there are three equilibria in the first quadrant:

$$E_0 = (0,0), \quad E_1 = (K_1, 0), \quad E_2 = (0, K_2).$$

In the last two cases, there are four equilibria:

$$E_0 = (0,0), \quad E_1 = (K_1, 0), \quad E_2 = (0, K_2) \text{ and } E^* = (x_1^*, x_2^*),$$

$$\begin{aligned} x_1^* &= \frac{\frac{1}{K_2} - \frac{1}{L_1}}{\frac{1}{K_1 K_2} - \frac{1}{L_1 L_2}} = \frac{L_2(s_1 - 1)}{s_1 s_2 - 1} \\ x_2^* &= \frac{\frac{1}{K_1} - \frac{1}{L_2}}{\frac{1}{K_1 K_2} - \frac{1}{L_1 L_2}} = \frac{L_1(s_2 - 1)}{s_1 s_2 - 1}. \end{aligned}$$

Stability The variation matrix $\frac{\partial \mathbf{f}}{\partial \mathbf{x}}$ at (x_1, x_2) reads

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(x_1, x_2) = \begin{pmatrix} r_1\left(1 - \frac{2x_1}{K_1} - \frac{x_2}{L_1}\right) & -\frac{r_1 x_1}{L_1} \\ -\frac{r_2 x_2}{L_2} & r_2\left(1 - \frac{2x_2}{K_2} - \frac{x_1}{L_2}\right) \end{pmatrix}.$$

We get

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(0,0) = \begin{pmatrix} r_1 & 0 \\ 0 & r_2 \end{pmatrix}, \frac{\partial \mathbf{f}}{\partial \mathbf{x}}(K_1, 0) = \begin{pmatrix} -r_1 & -\frac{K_1}{K_2}\frac{r_1}{s_1} \\ 0 & r_2(1 - \frac{1}{s_2}) \end{pmatrix},$$

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(0, K_2) = \begin{pmatrix} r_1(1 - \frac{1}{s_1}) & 0 \\ -\frac{r_2 K_2}{L_2} & -r_2 \end{pmatrix},$$

In all cases, E_0 is a unstable node.

After some computation, we can draw the following conclusion.

Theorem 3.3. *In the two-species competition model, the equilibria and their stability are the follows.*

- *Case 1:* $s_1 > 1$ *and* $s_2 < 1$*:* E_1 *is a stable sink.* E_2 *is unstable saddle.*
- *Case 2:* $s_1 < 1$ *and* $s_2 > 1$*:* E_2 *is a stable sink.* E_1 *is unstable saddle.*
- *Case 3:* $s_1 < 1$ *and* $s_2 < 1$*:* E_1 *and* E_2 *are stable sinks and* E^* *is a saddle.*
- *Case 4:* $s_1 > 1$ *and* $s_2 > 1$*: both* E_1 *and* E_2 *are saddles and* E^* *is a stable node.*

Ecologically, this theorem says that *co-existence of two competing species can occur only when both are competitive.*

In the case of the competitive model for the left curling snails and right curling snails, both have the same parameters r, K and α. Thus, both have the same competitive ratio:

$$s = \frac{r}{\alpha K}.$$

If $s > 1$, both would be competitive and they would co-exist. But this is not the case we have found. Instead, we find only one kind exists now in nature. To give an explanation, we notice that the term $-r/Kx_1^2$ represents the self competition, while the term $-\alpha x_1 x_2$ the cross competition. We should expect that these two competition terms are about the same magnitude. That is, $r/K \sim \alpha$. In this case, $s \sim 1$. If the cross competition is slightly stronger than the self competition, we would have $s < 1$. This would yield that only one species can survive in long time.
Ref. Clifford Henry Taubes, Modeling Differential Equations in Biology, pp. 23, pp. 73, pp. 81.

Homeworks.

1. Compute the eigenvalues of the variation matrix at E_1 and E_2.
2. Compute the variation matrix at (x^*, y^*) and its eigenvalues.
3. Justify the statements of this theorem.

3.4 Phase plane analysis

In this section, we shall use Maple to plot the vector field and to find orbits which connect nodes.

Include packages we type

```
> with(DEtools):
> with(plots):
```

Define the vector field (f,g) for the competition model

```
> f := r[1]*x(t)*(1-x(t)/K[1])-alpha[1]*x(t)*y(t);
> g := r[2]*y(t)*(1-y(t)/K[2])-alpha[2]*x(t)*y(t);
```

$$f := r_1\,\mathrm{x}(t)\,(1-\frac{\mathrm{x}(t)}{K_1})-\alpha_1\,\mathrm{x}(t)\,\mathrm{y}(t)$$

$$g := r_2\,\mathrm{y}(t)\,(1-\frac{\mathrm{y}(t)}{K_2})-\alpha_2\,\mathrm{x}(t)\,\mathrm{y}(t)$$

Define the following quantities.

```
> L[1] := r[1]/alpha[1]:
> L[2] := r[2]/alpha[2]:
> s[1] := L[1]/K[2]:
> s[2] := L[2]/K[1]:
```

The equilibria are those states where $(f,g) = (0,0)$. They are

E_0 = (0,0), E_1 = (K_1,0), E_2 = (0,K_2), E* = (xs,ys), where (xs,ys) are given by

```
> xs := L[2]*(s[1]-1)/(s[1]*s[2]-1):
> ys := L[1]*(s[2]-1)/(s[1]*s[2]-1):
```

We have four cases: Case 1: s_1 $>$ 1, s_2 $<$ 1:

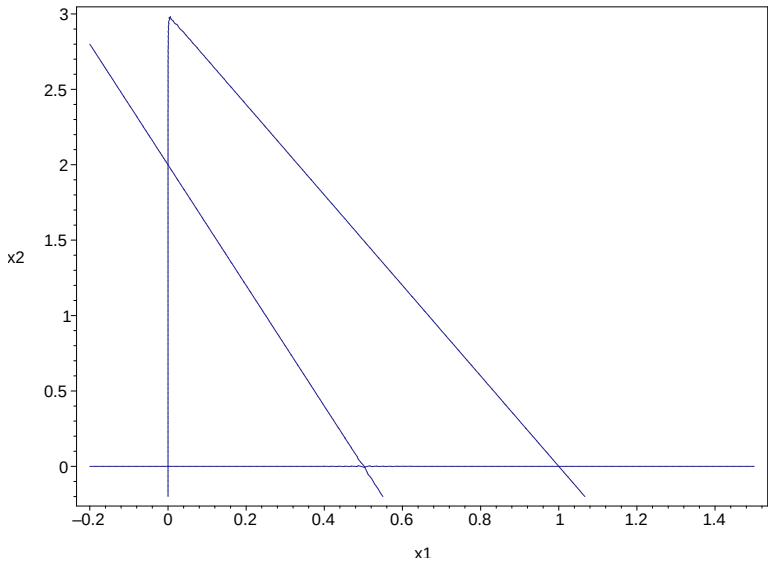
```
> Case1 := {
> r[1] = 3, K[1] = 1, alpha[1] = 1,
> r[2] = 2, K[2] = 2, alpha[2] = 4};
> evalf(subs(Case1,[s[1],s[2]]),3);
```

$$\mathit{Case1} := \{r_1 = 3,\ K_1 = 1,\ \alpha_1 = 1,\ r_2 = 2,\ K_2 = 2,\ \alpha_2 = 4\}$$

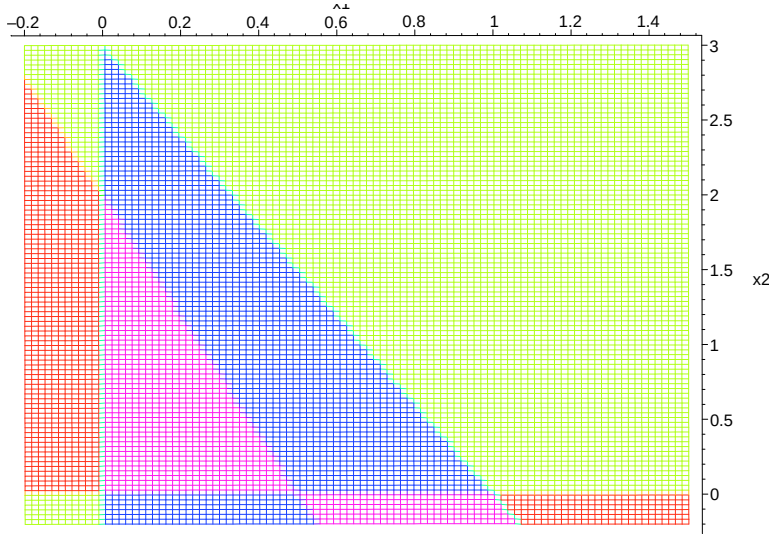
$$[1.50,\ 0.500]$$

Plot the the curves where $(f,g) = (0,0)$:

```
> fig1 :=
> implicitplot( {
> subs(Case1,x(t)=x1,y(t)=x2,f=0),
> subs(Case1,x(t)=x1,y(t)=x2,g=0) },
> x1=-0.2..1.5,x2=-0.2..3,
> grid=[100,100],color=navy):
> display(fig1,axes=boxed);
```

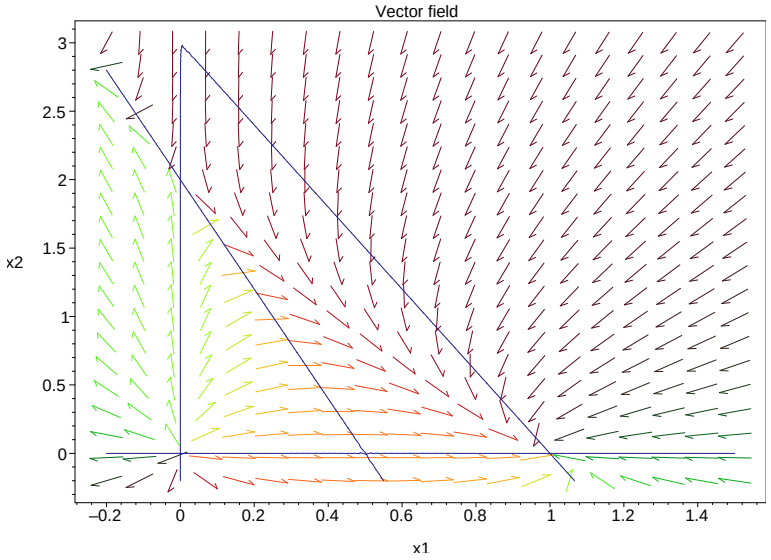



```
> f1    := subs(Case1,x(t)=x1,y(t)=x2,f):
> g1    := subs(Case1,x(t)=x1,y(t)=x2,g):
> vsign := piecewise(
> 2, f1 > 0 and g1 > 0,
> -2, f1 > 0 and g1 < 0,
> -1, f1 < 0 and g1 < 0,
> 1, f1 < 0 and g1 > 0);
> plot3d(vsign,x1=-0.2..1.5,x2=-0.2..3,axes=frame,grid=[100,100],
> orientation =[-90,0],style=HIDDEN,shading=ZHUE);
```



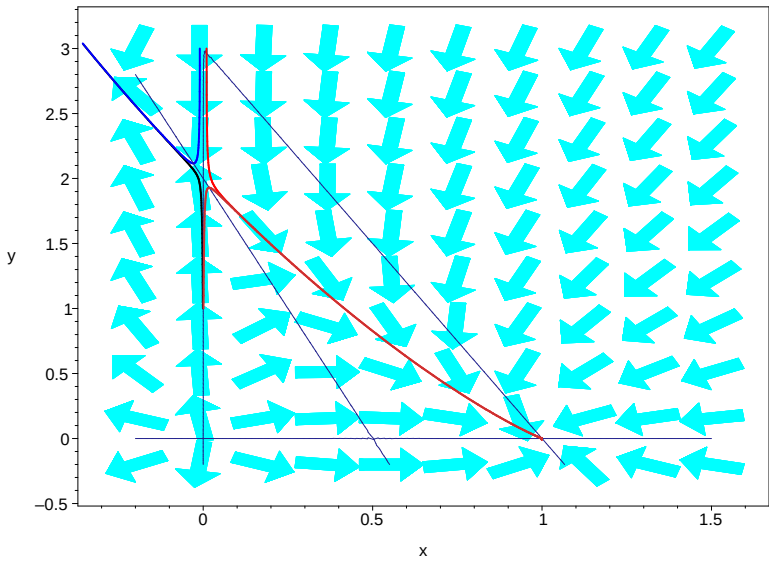
Plot the vector field (f,g) for case 1:

```
> fig2 := DEplot( subs(Case1,
> [diff(x(t),t)=f,diff(y(t),t)=g]),[x(t),y(t)], t=0..20,
> x=-0.2..1.5,y=-0.2..3,
> arrows=small,title=`Vector field`,
> color=subs(Case1,[f/sqrt(f^2+g^2),g/sqrt(f^2+g^2),0.1])):
> display({fig1,fig2},axes=boxed);
```



Find the separametrices. You need to try to find a proper initial data such that it generates a separametrix.

```
> fig3 := DEplot( subs(Case1,
> [diff(x(t),t)=f,diff(y(t),t)=g]),[x(t),y(t)],t=0..20,
> [[x(0)=0.01,y(0)=3]],stepsize=0.05,x=-0.2..1.5,y=- 0.2..3,
> color=cyan,arrows=LARGE,dirgrid=[10,10],linecolor=red):
> fig4 := DEplot( subs(Case1,
> [diff(x(t),t)=f,diff(y(t),t)=g]),[x(t),y(t)], t=0..20,
> [[x(0)=-0.01,y(0)=3]],stepsize=0.05,x=-0.2..1.5,y= -0.2..3,
> color=cyan,arrows=LARGE,dirgrid=[10,10],linecolor=blue):
> fig5 := DEplot( subs(Case1,
> [diff(x(t),t)=f,diff(y(t),t)=g]),[x(t),y(t)], t=0..20,
> [[x(0)=0.001,y(0)=1]],stepsize=0.05,x=-0.2..1.5,y=-0.2..3,
> color=cyan,arrows=LARGE,dirgrid=[10,10],linecolor=orange):
> fig6 := DEplot(subs(Case1,
> [diff(x(t),t)=f,diff(y(t),t)=g]),[x(t),y(t)], t=0..20,
> [[x(0)=-0.001,y(0)=1]],stepsize=0.05,x=-0.2..1.5,y=-0.2..3,
> color=cyan,arrows=LARGE,dirgrid=[10,10],linecolor=black):
> display({fig1,fig3,fig4,fig5,fig6},axes=boxed);
```



3.5 Periodic solutions

3.5.1 Predator-Prey system

Let x be the population of rabit (prey) and y the population of fox (predator). The equation for this predator-prey system is

$$\begin{aligned} \dot{x} &= ax - \alpha xy := f(x,y) \\ \dot{y} &= -by + \beta xy := g(x,y), \end{aligned}$$

where the coefficients $a, b, \alpha, \beta > 0$. The equilibria are those points such that $f(x,y) = 0$ and $g(x,y) = 0$. There are two equilibrate:

$$E_0 = (0,0) \quad \text{and} \quad E_* = (b/\beta, a/\alpha).$$

At E_0, the linearized equation is

$$\dot{\delta \mathbf{y}} = \frac{\partial \mathbf{F}}{\partial \mathbf{y}}(0)\delta \mathbf{y}.$$

The corresponding variation matrix is

$$\frac{\partial \mathbf{F}}{\partial \mathbf{y}}(0) = \begin{pmatrix} a & 0 \\ 0 & -b \end{pmatrix}.$$

We get E_0 is a saddle point, because one eigenvalue is positive and the other is negative. At E_*, the linearized matrix is

$$\frac{\partial \mathbf{F}}{\partial \mathbf{y}}(E_*) = \begin{pmatrix} 0 & -\alpha b/\beta \\ \alpha b/\beta & 0 \end{pmatrix}$$

The eigenvalues are pure imaginary. So E_* is an elliptic equilibrium. Near E_*, the solution is expected to be a closed trajectories (a periodic solution). In fact, we can integrate the predator-prey system as the follows. We notice that

$$\frac{dy}{dx} = \frac{y(-b+\beta x)}{x(a-\alpha y)}$$

is separable. It has the solution:

$$a\ln y - \alpha y + b\ln x - \beta x = C.$$

When C is the integration constant. The trajectories are closed curves surrounding E_*. Thus, the solutions are periodic solutions. For further study, see `http://www.scholarpedia.org/article/Predator-prey_model`

Homeworks.

1. * How does the period T depend on the coefficients?

3.5.2 van der Pol oscillator

In electric circuit theory, van der Pol proposed a model for electric circuit with vacuum tube, where $I = \phi(V)$ is a cubic function. Consider a circuit system with the resistor replaced by a device which obeys a nonlinear Ohm's law: the potential drop across this device is

$$\Delta V = \alpha\left(\frac{I^3}{3} - I\right),\ \alpha > 0.$$

Such a device does appear in vacuum tubes or semiconductors. The corresponding L-C-R circuit equation becomes

$$L\frac{dI}{dt} + \frac{Q}{C} + \alpha\left(\frac{I^3}{3} - I\right) = V(t). \tag{3.13}$$

Differentiate in t, we obtain the Van der Pole equation:

$$L\frac{d^2I}{dt^2} + \alpha(I^2-1)\frac{dI}{dt} + \frac{I}{C} = f(t). \tag{3.14}$$

where $f(t) = \dot{V}(t)$ is the applied electric field. The system is dissipative (damping) when $I^2 > 1$ and self current increasing when $I^2 < 1$.

Let x be the current and let us consider a normalized system:

$$\ddot{x} + \epsilon(x^2-1)\dot{x} + x = 0.$$

Through a Liénard transform:

$$y = x - \frac{x^3}{3} - \frac{\dot{x}}{\epsilon}$$

the van der Pol equation can be expressed as

$$\begin{aligned} \dot{x} &= \epsilon(x - \frac{x^3}{3} - y) \\ \dot{y} &= \frac{x}{\epsilon} \end{aligned}$$

We can draw the nullclines: $f = 0$ and $g = 0$. From the direction field of (f, g), we see that the field points inwards for large (x, y) and outward for (x, y) near $(0, 0)$. This means that there will be a limiting circle in between.

As $\epsilon >> 1$, we can observe that the time scale on x variable is fast whereas it is slow on the y-variable. That is,

$$\dot{x}(t) = O(\epsilon),\ \dot{y}(t) = O(1/\epsilon).$$

On the $x - y$ plane, consider the curve

$$y = x - \frac{x^3}{3}.$$

The solution moves fast to the curve $y = x - \frac{x^3}{3}$. Once it is closed to this curve, it move slowly along it until it moves to the critical points $(\pm 1, \pm\frac{2}{3})$. At which it moves away from the curve fast and move to the other side of the curve. The solution then periodically moves in this way.

Reference. You may google website on the Van der Pol oscillator on the web site of scholarpedia `http://www.scholarpedia.org/article/Van_der_Pol_oscillator` for more details.

Homeworks.

1. B-D: pp. 525, 8, 9
2. B-D: pp. 527, 17
3. Plot phase portraits for the four cases in the competitive model in the last subsection.

3.6 Heteroclinic and Homoclinic and orbits

Definition 3.5. *One the phase plane, an orbit which connects two equilibrium points is called a heteroclinic orbit. If the starting and end equilibrium points are the same, the orbit is called a homoclinic point.*

Below, we shall find the homoclinic orbit for the conservative mechanics with cubic potential

$$\ddot{x} = -\nabla V(x), \quad V(x) = \frac{1}{2}\left(-x^2 + x^3\right). \tag{3.15}$$

By multiply this equation by $\dot{x}$, we can get

$$\ddot{x}\dot{x} + \nabla V(x)\dot{x} = 0.$$

Using Libniz rule and chain rule, we obtain

$$\frac{d}{dt}\frac{1}{2}\dot{x}^2 + \frac{d}{dt}V(x) = 0.$$

This means that the quantity

$$H(x, \dot{x}) := \frac{1}{2}\dot{x}^2 + V(x)$$

is unchanged along a given orbit. Physically, this is called conservation of energy. The energy is composed of the kinetic energy $\frac{1}{2}\dot{x}^2$ and the potential energy $V(x)$. Along an orbit, the energy is determined by its initial state $(x(0), \dot{x}(0))$:

$$E_0 = \frac{1}{2}\dot{x}^2(0) + V(x(0)).$$

The phase plane is the $(x, \dot{x})$ plane. The orbit on the phase plane with this energy E_0 is

$$\frac{1}{2}\dot{x}^2 + \frac{1}{2}\left(-x^2 + x^3\right) = E_0.$$

Since $(0, 0)$ is a saddle, this homoclinic orbit $(x(t), \dot{x}(t))$ of (3.15) satisfies

$$x(\pm\infty) = 0, \quad \dot{x}(\pm\infty) = 0.$$

This leads to $E_0 = 0$. Using separation of variable, we have

$$\dot{x} = \pm\sqrt{x^2 - x^3}$$

$$\int \frac{dx}{x\sqrt{1-x}} = \pm(t + C)$$

Since the system is autonomous, we may normalize $C = 0$. For plus sign, we use the substitution $u = \sqrt{1-x}$, for minus, we use $u = -\sqrt{1-x}$. We get

$$\int \frac{2u\,du}{(1-u^2)u} = t$$

$$\int \left(\frac{1}{1+u} + \frac{1}{1-u}\right) = t.$$

$$\ln\left|\frac{1+u}{1-u}\right| = t.$$

$$\left|\frac{1+u}{1-u}\right| = e^t.$$

When $(1+u)/(1-u) \geq 0$, we obtain

$$u = \frac{e^t - 1}{e^t + 1} = \tanh\left(\frac{t}{2}\right).$$

This yields

$$x(t) = 1 - u^2 = \text{sech}^2\left(\frac{t}{2}\right).$$

When $(1+u)/(1-u) < 0$, we have

$$u = \frac{e^t + 1}{e^t - 1} = \coth\left(\frac{t}{2}\right).$$

This yields

$$x(t) = 1 - u^2 = -\text{csch}^2\left(\frac{t}{2}\right).$$

This should be the solution on the left-half plane in the phase plane. From

$$\dot{x}(t) = \sinh^{-3}\left(\frac{t}{2}\right)\cosh\left(\frac{t}{2}\right)\begin{cases} > & \text{for } t > 0 \\ < & \text{for } t < 0 \end{cases}$$

Hence, the branch on the upper plane is the one with $t \in (0, \infty)$, while the lower branch, $t \in (-\infty, 0)$.

Homeworks.

1. Use the same method to find the homoclinic orbits for the Duffing equation connecting $(0,0)$ to $(0,0)$. The Duffing equation is

$$\ddot{x} = -\nabla V(x), \quad V(x) = -\frac{\delta}{2}x^2 + \frac{1}{4}x^4.$$

Below, we use Maple to plot the orbits of Duffing equation in the phase plane.

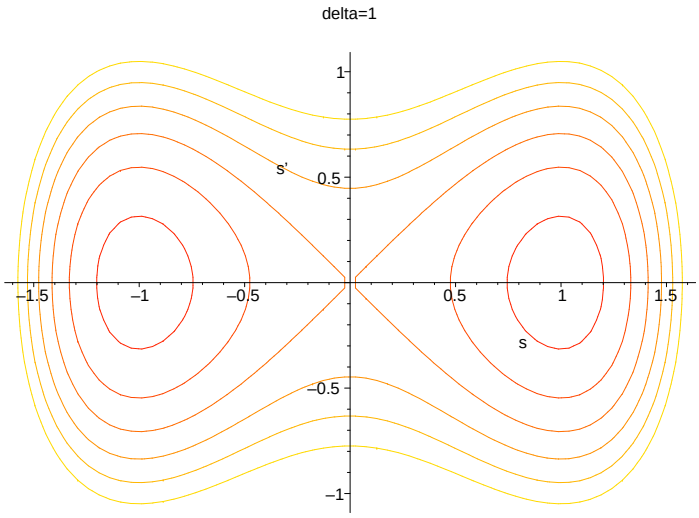
```
> with(DEtools):
with(plots):
```

```
> E := y^2/2+x^4/4-delta*x^2/2;
```

$$E := \frac{1}{2}y^2 + \frac{1}{4}x^4 - \frac{1}{2}\delta x^2$$

Plot the level set for the energy. Due to conservation of energy, these level sets are the orbits.

```
> contourplot(subs(delta=1,E),x=-2..2,y=-2..2,grid=[80,80],contours
=[-0.3,-0.2,-0.1,0,0.1,0.2,0.3],scaling=CONSTRAINED,labels=['s','s'],
title='delta=1');
```



Homework

1. Soliton appears in many physical systems such as water wave, nonlinear optics, etc. In shallow water, the soliton is a traveling of the so-called Korteweg and de Vries (KdV) equation

$$u_t + 6uu_x + u_{xxx} = 0.$$

A traveling wave of this equation is a solution of the form $\phi(x - ct)$. You can plug this ϕ into the KdV equation and obtain

$$-c\phi' + 6\phi\phi' + \phi''' = 0.$$

We can integrate it once and obtain

$$-c\phi + 3\phi^2 + \phi'' = C$$

The constant C depends on the boundary conditions at $\pm\infty$. For (bright) soliton, which are those traveling wave solutions decay fast at $x = \pm\infty$ and the corresponding constant $C = 0$. Thus, we arrive the ODE

$$-c\phi + 3\phi^2 + \phi'' = 0.$$

with $\phi(-\infty) = \phi(\infty) = 0$. The soliton is a homoclinic orbit on the phase plane $(\phi, \dot{\phi})$. Find close form of ϕ.

Heteroclinic Orbits We consider the pendulum equation

$$\ddot{x} = -\sin x.$$

The equation can be written in the form

$$\ddot{x} = -V'(x), \quad V(x) = -\cos x.$$

The system has conservation of energy

$$\frac{d}{dt}\left(\frac{1}{2}\dot{x}^2 - \cos x\right) = 0.$$

Thus,

$$\frac{1}{2}\dot{x}^2 - \cos x = E$$

The system can also be written

$$\dot{x} = p, \quad \dot{p} = -\sin x.$$

On the phase plane $(x, \dot{x})$ or equivalently, (x, p), the critical points are $(n\pi, 0)$. The orbit connecting $(-\pi, 0)$ to $(\pi, 0)$ is a heteroclinic orbit. This orbit has energy

$$E = \frac{1}{2}\dot{x}^2 - \cos x = 0 - \cos(\pi) = 1.$$

Thus, this heteroclinic orbit on the phase plane $(x, \dot{x})$ is determined by

$$\frac{1}{2}\dot{x}^2 - \cos x = 1.$$

We can integrate this equation

$$\dot{x}^2 = 2(1 + \cos) = 4\cos^2\left(\frac{x}{2}\right).$$

$$\dot{x} = \pm 2\cos\left(\frac{x}{2}\right).$$

There are two branches, one is on the upper half plane: $\dot{x} > 0$, the other is on the lower half plane:$\dot{x} < 0$. They are symmetric. We only find the upper one. Using separation of variable, we get

$$\frac{dx}{2\cos\left(\frac{x}{2}\right)} = dt$$

Call $y = x/2$, we get

$$\frac{dy}{\cos y} = dt.$$

Using polar stereographic projection:

$$u = \tan\left(\frac{y}{2}\right), \quad \frac{dy}{du} = \frac{2}{1+u^2},$$

$$\cos y = \frac{1-u^2}{1+u^2}, \quad \sin y = \frac{2u}{1+u^2}$$

We obtain

$$dt = \frac{dy}{\cos y} = \frac{2du}{1+u^2}\frac{1+u^2}{1-u^2} = \frac{2du}{1-u^2}.$$

Integrate this, we get

$$t = \ln\left|\frac{1-u}{1+u}\right|.$$

We obtain the same formula as we did for the cubic potential case. You can fill in the rest of the solution.

Homework The heteroclinic orbit also appears commonly on so-called interface shape function. An interface shape function is which connect two states a and b by a shape function ϕ. It satisfies

$$\phi'' - F'(\phi) = 0.$$

The function $F(a) = F(b) = 0$ and has no zero between them. The shape function $\phi(-\infty) = a$ and $\phi(\infty) = b$, and $\phi'(\pm\infty) = 0$. What kind of condition F should satisfies in order to have a heteroclinic orbit connecting a to b? Can you integrate it for polynomial type of F with degree than 4?

Chapter 4

Linear Systems with Constant Coefficients

4.1 Initial value problems for $n \times n$ linear systems

A general $n \times n$ linear system of differential equation is of the form

$$\mathbf{y}'(t) = \mathbf{A}\mathbf{y}(t) + \mathbf{f}(t), \tag{4.1}$$

where

$$\mathbf{y} = \begin{pmatrix} y^1 \\ y^2 \\ \vdots \\ y^n \end{pmatrix}, \ \mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}, \ \mathbf{f} = \begin{pmatrix} f^1 \\ f^2 \\ \vdots \\ f^n \end{pmatrix},$$

Its initial value problem is to study (4.1) with initial condition:

$$\mathbf{y}(0) = \mathbf{y}_0. \tag{4.2}$$

4.2 Physical Models

4.2.1 Coupled spring-mass systems

Consider a spring-mass system which consists of n masses placed vertically between two walls. The n masses and the two end walls are connected by $n+1$ springs. If all masses are zeros, the springs are "at rest" states. When the masses are greater than zeros, the springs are elongated due to the gravitation force. The mass m_i moves down y_i distance, called the displacement. The goal is to determine the displacements y_i of the masses m_i, $i = 1, ..., n$.

Let us call the spring connecting m_{i-1} and m_i by spring i, $i = 1, ..., n+1$. Suppose the spring i has spring constant k_i. Let us call the downward direction the positive direction.

1. Let me start from the simplest case: $n = 1$ and no bottom wall. The mass m_1 elongates the spring 1 by a displacement y_1. The elongated spring has a *restoration force* $-k_1 y_1$ acting on m_1.[1] Thus, we have
$$m_1 \ddot{y}_1 = -k_1 y_1 + f_1,$$
where $f_1 = m_1 g$, the gravitation force on m_1, and g is the gravitation constant.

2. Next, let us consider the case where there is a bottom wall. In this case, both springs 1 and 2 exert forces upward to m_1. The balance law becomes
$$m_1 \ddot{y}_1 = -k_1 y_1 - k_2 y_1 + f_1.$$

3. Let us jump to a slightly more complicated case, say $n = 3$.

   - The displacements
   $$y_0 = 0, \quad y_4 = 0,$$
   due to the walls are fixed.
   - The displacements y_1, y_2, y_3 cause elongations of the springs:
   $$e_i = y_i - y_{i-1}, i = 1, 2, 3, 4.$$
   The restoration force of spring i is
   $$w_i = k_i e_i.$$
   - The force exerted to m_i by spring i is $-w_i = -k_i e_i$. In fact, when $e_i < 0$, the spring is shortened and it pushes downward to mass m_i (the sign is positive), hence the force is $-k_i e_i > 0$. On the other hand, when $e_i > 0$, the spring is elongated and it pull m_i upward. We still get the force $-w_i = -k_i e_i < 0$.
   - Similarly, the force exerted to m_i by spring $i + 1$ is $w_{i+1} = k_{i+1} e_{i+1}$. When $e_{i+1} > 0$, the spring $i + 1$ is elongated and it pulls m_i downward, the force is $w_{i+1} = k_{i+1} e_{i+1} > 0$. When $k_{i+1} < 0$, it pushes m_i upward, and the force $w_{i+1} = k_{i+1} e_{i+1} < 0$. In both cases, the force exerted to m_i by spring $i + 1$ is w_{i+1}.
   - Thus, the equation of motion of m_i is
   $$m_i \ddot{y}_i = w_{i+1} - w_i + f_i = k_{i+1}(y_{i+1} - y_i) - k_i(y_i - y_{i-1}) + f_i, i = 1, 2, 3.$$

   Let us express the above equations in matrix form. For $n = 4$, we get
   $$\mathbf{M}\ddot{\mathbf{y}} + \mathbf{K}\mathbf{y} = \mathbf{f}.$$

[1] The minus sign is due to the direction of force is upward.

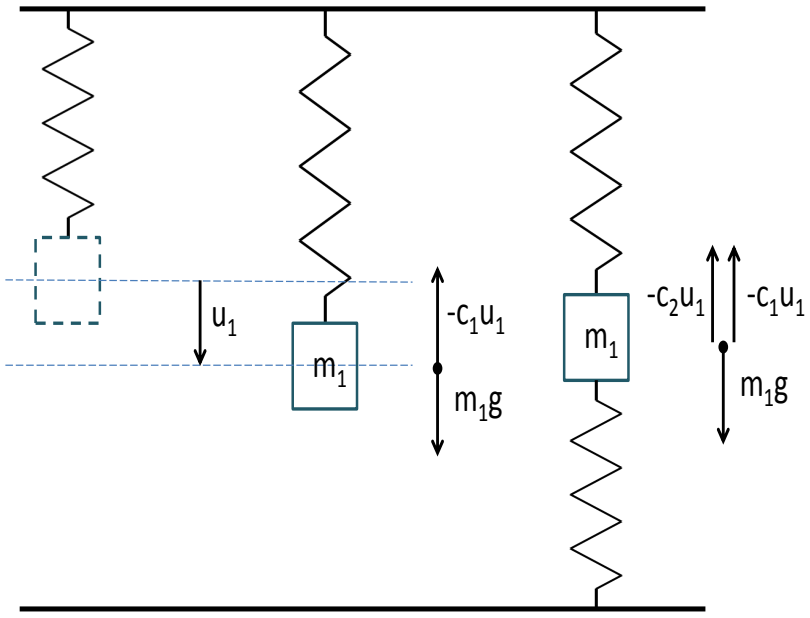


Figure 4.1: The left one is a spring without any mass. The middle one is a spring hanging a mass m_1 freely. The right one is a mass m_1 with two springs fixed on the ceiling and floor.

where

$$\mathbf{M} = \begin{pmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix},$$

$$\mathbf{K} := \begin{pmatrix} k_1 + k_2 & -k_2 & 0 \\ -k_2 & k_2 + k_3 & -k_3 \\ 0 & -k_3 & k_3 + k_4 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} m_1 g \\ m_2 g \\ m_3 g \end{pmatrix}$$

4.2.2 Coupled Circuit Systems

A circuit system consists of inductors, capacitors and resistors connected by wires. It can be modeled by a graph $G = (V, E)$, which consists of vertices $V = \{1, 2, ..., n\}$ and edges $E =$

$\{(1,2),(1,3),...\}$, say for example. An edge $(i,j) \in E$ means that there is an edge connecting vertices i and j. In the circuit model, the edges are the wires. On each wire, there is an electric component. The vertices (or called nodes) are those wire junctions. The circuit theory assumes the current is uniform on each edge, that is, independent of position. Thus, a current I_e is associated with each edge e. On each edge, there is a potential drop across each component. The potential drops of inductor, capacitor and resistor are respectively

- inductor: $V_i = LdI/dt$,
- capacitor: $V_c = Q/C$, or $V_c' = I/C$,
- resistor: $V_r = IR$,

where L, C and R are the inductance, capacitance and resistance. The circuit equations are derived based on the Kirchhoff laws.

- *Kirchhoff's first law: at each junction, the sum of the currents flowing in is equal to the sum of currents flowing out.*
- *Kirchhoff's second law: the potential differences around any closed loop is zero.*

Example Suppose $V = \{1,2,3,4\}$, $E = \{(1,2),(2,3),(3,4),(4,1),(2,4)\}$. The component on each edges are: $R_{12}, C_{34}, L_{41}, R_{24}$. There are 5 edges, thus, 5 unknowns. We choose the unknowns to be $I_{12}, I_{23}, I_{34}, I_{41}, I_{24}$.

The Kirchhoff's first (charge) laws gives

$$\begin{aligned}
&\text{at node 1 }: I_{41} = I_{12},\\
&\text{at node 2 }: I_{12} - I_{23} - I_{24} = 0,\\
&\text{at node 3 }: I_{23} = I_{34},\\
&\text{at node 4 }: I_{34} + I_{24} - I_{41} = 0.
\end{aligned}$$

We eliminate I_{23} and I_{41} right away from the charge laws at node1 and 3. There is one redundant equation because of the cyclic property of this graph. Thus, the only independent condition we need is the charge law at node 4. Now, we have 3 unknowns: I_{12}, I_{24} and I_{34} and one charge law at node 4, which is

$$I_{34} + I_{24} - I_{12} = 0. \tag{4.3}$$

The Kirchhoff's second (voltage) law gives

$$\begin{aligned}
&\text{cycle } (1,2,4) : V_{12} + V_{24} + V_{41} = 0,\\
&\text{cycle } (2,3,4) : V_{23} + V_{34} + V_{42} = 0.
\end{aligned}$$

The first one gives

$$R_{12}I_{12} + R_{24}I_{24} + L_{41}I_{12}' = 0. \tag{4.4}$$

The second one gives

$$Q_{34}/C_{34} - I_{24}R_{24} = 0.$$

We differentiate this equation and get

$$I_{34}/C_{34} - I'_{24}R_{24} = 0. \tag{4.5}$$

We can eliminate I_{34} and get the following two equations for I_{12} and I_{24}.

$$\begin{cases} R_{12}I_{12} + R_{24}I_{24} + L_{41}I'_{12} = 0 \\ I_{12} - I_{24} - I'_{24}R_{24}C_{34} = 0. \end{cases} \tag{4.6}$$

Alternatively, we can choose V_{34} and I_{12} as our unknowns. In cycle $(2, 3, 4)$, we have

$$V_{34} - R_{24}I_{24} = 0.$$

This can represent I_{24} in terms of V_{34}. We use $C_{34}V'_{34} = I_{34}$ to eliminate I_{34}. Finally, at node 4 and the cycle $(1, 2, 4)$, we have

$$\begin{cases} C_{34}V'_{34} + V_{34}/R_{24} - I_{12} = 0 \\ L_{41}I'_{12} + R_{12}I_{12} + V_{34} = 0. \end{cases} \tag{4.7}$$

Homeworks

1. In the spring-mass system, suppose we have only two masses m_1 and m_2. The mass m_1 is hung on the ceiling, the mass m_2 is connected to m_1 on one end, but is free on the other end. Write a mathematical model for this case.
2. Write down the two differential equations for the above circuit system.
3. A parallel *LRC* circuit connects L, R and C in parallel way. Write a graph model and ODE model for it.
4. B-D pp. 411: 25
5. B-D pp. 411: 26.

4.3 Linearity and solution space

We shall first study the homogeneous equation

$$\mathbf{y}' = \mathbf{A}\mathbf{y}. \tag{4.8}$$

Since the equation is linear in $\mathbf{y}$, we can see the following linear property of the solutions. Namely, if $\mathbf{y}_1$ and $\mathbf{y}_2$ are solutions of (4.8), so does their linear combination: $\alpha_1\mathbf{y}_1 + \alpha_2\mathbf{y}_2$, where α_1, α_2 are any two scalar numbers. Therefore, if $\mathcal{S}_0$ denotes the set of all solutions of (4.8), then $\mathcal{S}_0$ is a vector space.

In the case of inhomogeneous equation (4.1), suppose we have already known a particular solution $\mathbf{y}_p$, then so is $\mathbf{y}_p + \mathbf{y}$ for any $\mathbf{y} \in \mathcal{S}_0$. On the other hand, suppose $\mathbf{z}$ is a solution of the inhomogeneous equation:

$$\mathbf{z}' = \mathbf{A}\mathbf{z} + \mathbf{f}$$

then $\mathbf{z} - \mathbf{y}_p$ satisfies the homogeneous equation (4.8). Hence, $\mathbf{z} - \mathbf{y}_p = \mathbf{y}$ for some $\mathbf{y} \in \mathcal{S}$. We conclude that the set of all solutions of the inhomogeneous equation (4.1) is the affine space

$$\mathcal{S} = \mathbf{y}_p + \mathcal{S}_0.$$

To determine the dimension of the solution, we notice that all solutions are uniquely determined by their initial data (the *existence and uniqueness theorems*),

$$\mathbf{y}(0) = \mathbf{y}_0 \in \mathbb{C}^n.$$

Hence, $\mathcal{S}_0$ is n dimensional. We conclude this argument by the following theorem.

Theorem 4.1. *The solution space $\mathcal{S}_0$ for equation (4.8) is an n-dimensional vector space. The solution space for equation (4.1) is the affine space $\mathbf{y}_p + \mathcal{S}_0$, where $\mathbf{y}_p$ is a particular solution of (4.1).*

Fundamental solutions Our goal in this section is to construct a basis $\{\mathbf{y}_1, ..., \mathbf{y}_n\}$ in $\mathcal{S}_0$. A general solution in S_0 can be represented as

$$\mathbf{y}(t) = \sum_{i=1}^{n} C_i \mathbf{y}_i(t).$$

For an initial value problem with $\mathbf{y}(t_0) = \mathbf{y}_0$, the coefficients C_i are determined by the linear equation

$$\sum_{i=1}^{n} \mathbf{y}_i(t_0) C_i = \mathbf{y}_0.$$

or

$$\mathbf{Y}(t_0) C = \mathbf{y}_0$$

where

$$\mathbf{Y}(t) = [\mathbf{y}_1(t), \mathbf{y}_2(t), \cdots, \mathbf{y}_n(t)], C = [C_1, \cdots, C_n]^t.$$

If $\mathbf{y}_1, \cdots, \mathbf{y}_n$ are independent, then C_i can be solved *uniquely*. Such a set of solutions $\{\mathbf{y}_1, \cdots, \mathbf{y}_n\}$ is called a fundamental solution of (4.8). So our main task is to find a set of fundamental solutions.

The basic idea is to try the decompose the system into smaller systems which can be solved easily. We shall learn this through examples first, then develop general theory.

4.4 Decouping the systems

4.4.1 Linear systems in three dimensions

Consider the 3×3 linear system

$$\mathbf{y}' = \mathbf{A}\mathbf{y},$$

where

$$\mathbf{y} = \begin{pmatrix} y^1 \\ y^2 \\ y^3 \end{pmatrix}, \ \mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}.$$

We look for three independent solutions of the form $e^{\lambda t}\mathbf{v}$. By plugging this into the equation, we find that λ and $\mathbf{v}$ have to be an eigenvalue and eigenvector of $\mathbf{A}$:

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}.$$

The eigenvalue satisfies the characteristic equation

$$\det(\lambda\mathbf{I} - \mathbf{A}) = 0.$$

This is a third order equation because we have a 3×3 system. One of its roots must be real. The other two roots can be both real or complex conjugate. We label the first one by λ_3 and the other two by λ_1 and λ_2. The corresponding eigenvectors are denoted by $\mathbf{v}_i$, $i = 1, 2, 3$. It is possible that $\lambda_1 = \lambda_2$. In this case, $\mathbf{v}_1$ and $\mathbf{v}_2$ are the vectors to make $\mathbf{A}$ in Jordan block. That is

$$\begin{aligned} \mathbf{A}\mathbf{v}_1 &= \lambda_1\mathbf{v}_1 \\ \mathbf{A}\mathbf{v}_2 &= \lambda_1\mathbf{v}_2 + \mathbf{v}_1 \end{aligned}$$

The general solution is

$$\mathbf{y}(t) = C_1\mathbf{y}_1(t) + C_2\mathbf{y}_2(t) + C_3\mathbf{y}_3(t).$$

The solution $\mathbf{y}_1$ and $\mathbf{y}_2$ are found exactly the same way as that in two dimension. The solution $\mathbf{y}_3(t) = e^{\lambda_3 t}\mathbf{v}_3$. If $\lambda_3 < 0$, then the general solution tends to the plane spanned by $\mathbf{v}_1$ and $\mathbf{v}_2$. Let us denote this plane by $< \mathbf{v}_1, \mathbf{v}_2 >$. On the other hand, if $\lambda_3 > 0$, the solution leaves the plane $< \mathbf{v}_1, \mathbf{v}_2 >$.

Example.

Consider

$$\mathbf{A} = \begin{pmatrix} 0 & 0.1 & 0 \\ 0 & 0 & 0.2 \\ 0.4 & 0 & 0 \end{pmatrix}.$$

The characteristic equation is

$$\lambda^3 - 0.008 = 0.$$

The roots are

$$\lambda_3 = 0.2,\ \lambda_1 = 0.2e^{i2\pi/3},\ \lambda_2 = 0.2e^{-i2\pi/3}.$$

The eigenvectors are

$$\mathbf{v}_3 = \begin{pmatrix} 1/2 \\ 1 \\ 1 \end{pmatrix},\ \mathbf{v}_1 = \begin{pmatrix} -1+i\sqrt{3} \\ -2-i2\sqrt{3} \\ 4 \end{pmatrix},\ \mathbf{v}_2 = \begin{pmatrix} -1-i\sqrt{3} \\ -2+i2\sqrt{3} \\ 4 \end{pmatrix}.$$

We denote $\mathbf{v}_1 = \mathbf{u}_1 + i\mathbf{u}_2$ and $\mathbf{v}_2 = \mathbf{u}_1 - i\mathbf{u}_2$. We also denote $\lambda_1 = \alpha + i\omega$, where $\alpha = -0.1$ and $\omega = \sqrt{0.03}$. Then the fundamental solutions are

$$\begin{aligned} \mathbf{y}_1(t) &= e^{\alpha t}(\cos(\omega t)\mathbf{u}_1 - \sin(\omega t)\mathbf{u}_2) \\ \mathbf{y}_2(t) &= e^{\alpha t}(\sin(\omega t)\mathbf{u}_1 + \cos(\omega t)\mathbf{u}_2) \\ \mathbf{y}_3(t) &= e^{\lambda_3 t}\mathbf{v}_3 \end{aligned}$$

4.4.2 Rotation in three dimensions

An important example for 3×3 linear system is the rotation in three dimensions. The governing equation is

$$\begin{aligned} \mathbf{y}'(t) &= \mathbf{\Omega} \times \mathbf{y} \\ &= \begin{pmatrix} 0 & -\omega_3 & -\omega_2 \\ \omega_3 & 0 & -\omega_1 \\ \omega_2 & \omega_1 & 0 \end{pmatrix} \mathbf{y} \end{aligned}$$

We have many examples in the physical world represented with the same equation.

- Top motion in classical mechanics: $\mathbf{y}$ is the angular momentum and $\mathbf{\Omega} \times \mathbf{y}$ is the torque.
- Dipole motion in a magnetic field: $\mathbf{y}$ is the angular momentum which is proportional to the magnetic dipole
- A particle motion under Coriolis force: $\mathbf{y}$ is the velocity and $-2\mathbf{\Omega} \times \mathbf{y}$ is the Coriolis force.
- Charge particle motion in magnetic field: $\mathbf{y}$ is the velocity. The term $\mathbf{\Omega} \times \mathbf{y}$ is a force pointing to the direction perpendicular to $\mathbf{y}$ and $\mathbf{\Omega}$. This is the Lorentz force in the motion of a charge particle in magnetic field Ω.
- Spin motion in magnetic field: $\mathbf{y}$ is the spin and $\mathbf{\Omega}$ is the magnetic field.

We may normalize $\mathbf{\Omega} = \omega\hat{z}$. In this case, the equation becomes

$$\begin{aligned} y^{1'} &= -\omega y^2 \\ y^{2'} &= \omega y^1 \\ y^{3'} &= 0 \end{aligned}$$

The solution reads:

$$\mathbf{y}(t) = \mathbf{R}(t)\mathbf{y}(0),\ \begin{pmatrix} \cos\omega t & -\sin\omega t & 0 \\ \sin\omega t & \cos\omega t & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

It is a rotation about the z axis with angular velocity ω.

Motion of a charge particle in constant electric magnetic field The force exerted by a charged particle is known as the Lorentz force

$$F = q(E + v \times B)$$

The motion of the charged particle in this E-M field is governed by

$$m\ddot{r} = F.$$

Suppose the EM field is constant with E only in z direction and B in x direction. Then the motion is on y-z plane if it is so initially. We write the equation in each components:

$$m\ddot{y} = qB\dot{z},\ m\ddot{z} = qE - qB\dot{y}.$$

Let

$$\omega := \frac{qB}{m},$$

the equations are rewritten as

$$\ddot{y} = \omega\dot{z},\ \ddot{z} = \omega\left(\frac{E}{B} - \dot{y}\right).$$

The particle started from zero vecolity has the trajectory

$$y(t) = \frac{E}{\omega B}(\omega t - \sin\omega t),\ z(t) = \frac{E}{\omega B}(1 - \cos\omega t).$$

This is a *cycloid*.

Homeworks

1. Complete the above calculation for motion of charge particle in electro-magnetic field.
2. Consider the equation

   $$p\left(\frac{d}{dt}\right)y(t) = 0,$$

   where y is scalar. Let us consider

   $$p(s) = (s-1)^3.$$

   Show that

   $$y_1(t) = e^t,\ y_2(t) = te^t,\ y_3(t) = t^2e^t.$$

   are three independent solutions.

3. Solve the system

$$\mathbf{y}' = \mathbf{A}\mathbf{y}, \mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & -1 \\ -3 & 2 & 4 \end{pmatrix}.$$

Ref. B-D pp. 429, problem 17.

4.4.3 Decoupling the spring-mass systems

Let us consider a spring-mass system which consists of 3 masses connected by 4 springs. We assume the masses have equal mass m, the springs have equal spring constant k, and there is no gravitational force. Let y_j be the displacement of mass i. Then the resulting differential equation is

$$\mathbf{M}\ddot{\mathbf{y}} + \mathbf{K}\mathbf{y} = \mathbf{0}.$$

where

$$\mathbf{M} = m\mathbf{I}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}, \quad \mathbf{K} := k \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$$

The idea to solve this system is to decouple it. That is, we will try to diagonalize this system. We find that the eigenvalues of $\mathbf{K}$ are $\lambda_1 = 2 - \sqrt{2}$, $\lambda_2 = 2$, $\lambda_3 = 2 + \sqrt{2}$. The corresponding eigenvectors are

$$\mathbf{v}_1 = \begin{pmatrix} 1/2 \\ 1/\sqrt{2} \\ 1/2 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 1/\sqrt{2} \\ 0 \\ -1/\sqrt{2} \end{pmatrix}, \quad \mathbf{v}_3 = \begin{pmatrix} 1/2 \\ -1/\sqrt{2} \\ 1/2 \end{pmatrix}.$$

Let us take the ansatz

$$\mathbf{y}(t) = \sum_{i=1}^{3} C_i(t)\mathbf{v}_i$$

Plug this ansatz to the equation, we get

$$m\sum_{i=1}^{3} \ddot{C}_i(t)\mathbf{v}_i = k\mathbf{K}(\sum_{i=1}^{3} C_i(t)\mathbf{v}_i) = k\sum_{i=1}^{3} \lambda_i C_i(t)\mathbf{v}_i.$$

Since $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ are independent, we get

$$m\ddot{C}_i(t) = k\lambda_i C_i(t), i = 1, 2, 3.$$

The system is decoupled! Their solutions are

$$C_i(t) = A_i \cos\left(\sqrt{\frac{k\lambda_i}{m}}t\right) + B_i \sin\left(\sqrt{\frac{k\lambda_i}{m}}t\right).$$

Thus, the general solutions can be expressed as

$$\mathbf{y}(t) = \sum_{i=1}^{3} \left(A_i \cos\left(\sqrt{\frac{k\lambda_i}{m}} t \right) + B_i \sin\left(\sqrt{\frac{k\lambda_i}{m}} t \right) \right) \mathbf{v}_i.$$

The coefficients A_i, B_i are determined by the initial conditions:

$$\sum_{i=0}^{3} A_i \mathbf{v}_i = \mathbf{y}(0), \quad \sum_{i=0}^{3} \sqrt{\frac{k\lambda_i}{m}} B_i \mathbf{v}_i = \mathbf{y}'(0).$$

Since $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ are orthonormal, we can obtain the coefficients easily:

$$A_i = \langle \mathbf{y}(0), \mathbf{v}_i \rangle, \quad B_i = \sqrt{\frac{k\lambda_i}{m}} \langle \mathbf{y}(0), \mathbf{v}_i \rangle.$$

Remark It is worth noting that $\mathbf{v}_i$ can be expressed as

$$\mathbf{v}_1 = \begin{pmatrix} \sin(\pi/4) \\ \sin(\pi/2) \\ \sin(3\pi/4) \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} \sin(2\pi/4) \\ \sin(\pi) \\ \sin(6\pi/4) \end{pmatrix}, \quad \mathbf{v}_3 = \begin{pmatrix} \sin(3\pi/4) \\ \sin(6\pi/4) \\ \sin(9\pi/4) \end{pmatrix}.$$

Homework Let us consider a spring-mass system consisting of $n-1$ masses connecting by n springs with two ends fixed. We assume the masses have equal mass m, the springs have equal spring constant k, and there is no gravitational force. Let y_j be the displacement of mass i. Then the resulting differential equation is

$$\mathbf{M}\ddot{\mathbf{y}} + \mathbf{K}\mathbf{y} = \mathbf{0}.$$

where

$$\mathbf{M} = m\mathbf{I}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \end{pmatrix}, \quad \mathbf{K} := k \begin{pmatrix} 2 & -1 & 0 & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & 0 & -1 & 2 \end{pmatrix}_{(n-1)\times(n-1)}$$

Prove that $\mathbf{K}$ can be diagonalized by

$$\mathbf{v}_\ell = \begin{pmatrix} \sin(\ell\pi/n) \\ \sin(2\ell\pi/n) \\ \vdots \\ \sin((n-1)\ell\pi/n) \end{pmatrix}, \qquad \ell = 1, \cdots, n-1.$$

What are the corresponding eigenvalues? Find the explicit expression of general solutions.

4.5 Jordan canonical form

4.5.1 Jordan matrix

In the 2×2 system $\mathbf{y}' = \mathbf{A}\mathbf{y}$, we have seen that when $\mathbf{A}$ has multiple eigenvalue, it may be similar to a special 2×2 matrix

$$\mathbf{V}^{-1}\mathbf{A}\mathbf{V} = \mathbf{J} = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}.$$

Such matrix is called a Jordan matrix. If we define $\mathbf{z} = \mathbf{V}^{-1}\mathbf{y}$, then

$$\mathbf{z}' = \mathbf{J}\mathbf{z},$$

which can be solved easily. For $n \times n$ system $\mathbf{y}' = \mathbf{A}\mathbf{y}$, we also want to transform it to such kind of system which we can solve easily.

A matrix $\mathbf{J}$ is called a Jordan normal form of a matrix $\mathbf{A}$ if we can find matrix $\mathbf{V}$ such that

$$\mathbf{A}\mathbf{V} = \mathbf{V}\mathbf{J},$$

where

$$\mathbf{J} = \mathbf{J}_{k_1} \otimes \cdots \otimes \mathbf{J}_{k_p} := \begin{pmatrix} \mathbf{J}_{k_1} & & & \\ & \mathbf{J}_{k_2} & & \\ & & \ddots & \\ & & & \mathbf{J}_{k_p} \end{pmatrix}, \quad \mathbf{V} = [\mathbf{V}_{k_1}, \mathbf{V}_{k_2}, \cdots, \mathbf{V}_{k_p}].$$

$$\mathbf{J}_k(\lambda_k) = \begin{pmatrix} \lambda_k & 1 & & & \\ & \lambda_k & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda_k & 1 \\ & & & & \lambda_k \end{pmatrix}_{k\times k}, \quad \mathbf{V}_k = [\mathbf{v}_k^1, \cdots, \mathbf{v}_k^k], \quad k = k_1, ..., k_s,$$

$$\sum_{i=1}^{s} k_i = n.$$

Here, λ_{k_i} are the eigenvalues of $\mathbf{A}$, $\mathbf{v}_k^j \in \mathbb{C}^n$ are called the generalized eigenvectors of $\mathbf{A}$, the matrices $\mathbf{J}_k$ are called Jordan blocks of size k of $\mathbf{A}$. The matrix $\mathbf{V}_k = [\mathbf{v}_k^1, \cdots, \mathbf{v}_k^k]$ is an $n \times k$ matrix. We can restrict $\mathbf{A}$ to $\mathbf{V}_k$, $k = k_1, ..., k_s$ as

$$\mathbf{A}\mathbf{V}_k = \mathbf{A}[\mathbf{v}_k^1, \cdots, \mathbf{v}_k^k] = [\mathbf{v}_k^1, \cdots, \mathbf{v}_k^k]\mathbf{J}_k, \ k = k_1, ..., k_s.$$

For each generalized vector,

$$\begin{aligned} (\mathbf{A} - \lambda_k\mathbf{I})\mathbf{v}_k^1 &= 0 \\ (\mathbf{A} - \lambda_k\mathbf{I})\mathbf{v}_k^2 &= \mathbf{v}_k^1 \\ &\vdots \\ (\mathbf{A} - \lambda_k\mathbf{I})\mathbf{v}_k^k &= \mathbf{v}_k^{k-1}, \quad k = k_1, ..., k_s. \end{aligned}$$

This implies

$$\begin{aligned}(\mathbf{A}-\lambda_k\mathbf{I})\,\mathbf{v}_k^1 &= 0\\ (\mathbf{A}-\lambda_k\mathbf{I})^2\mathbf{v}_k^2 &= 0\\ &\vdots\\ (\mathbf{A}-\lambda_k\mathbf{I})^k\mathbf{v}_k^k &= 0, \quad k=k_1,...,k_s.\end{aligned}$$

We will see later from the construction that the set $\{\mathbf{v}_{k_i}^j\}$ form a basis in $\mathbb{C}^n$. Therefore, $\mathbf{V}$ is invertible, and

$$\mathbf{A}=\mathbf{V}\mathbf{J}\mathbf{V}^{-1}.$$

We call $\mathbf{A}$ is similar to $\mathbf{J}$, and is denoted by $\mathbf{A}\sim\mathbf{J}$.

The matrix $\mathbf{N}_k := \mathbf{J}_k - \lambda_k\mathbf{I}$ is called a Nilpotent matrix, which has the form

$$\mathbf{N}_k=\begin{pmatrix}0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1\\ & & & 0\end{pmatrix}_{k\times k}.$$

It is easy to check that

$$\mathbf{N}_k^2=\begin{pmatrix}0 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 0 & 0 & 1\\ & & & 0 & 0\\ & & & & 0\end{pmatrix}_{k\times k}, \quad \cdots, \quad ,\mathbf{N}_k^k=\mathbf{0}.$$

Theorem 4.2. *Any matrix $\mathbf{A}$ over $\mathbb{C}$ is similar to a Jordan normal form. The structure of this Jordan normal form is unique.*

Before we develop general theory, let us study some examples first. These examples tell us

- how to find the structure of the Jordan matrix,
- how to find the generalized eigenvectors $\mathbf{v}_1,...,\mathbf{v}_n$.

We shall consider the case where the characteristic polynomial $p_A(\lambda) := det(\lambda\mathbf{I}-\mathbf{A})$ has only one eigenvalue with multiple multiplicity.

Example Suppose $\mathbf{A}$ is a 2×2 matrix with double eigenvalue λ. Let $\mathcal{N}_1 = Ker(\mathbf{A}-\lambda\mathbf{I})$ and $\mathcal{N}_2 = Ker(\mathbf{A}-\lambda\mathbf{I})^2$.

- Determine the structure of the Jordan block. For 2×2 matrix, there are only two possible structures: $\mathbf{J}_1\otimes\mathbf{J}_1$, or $\mathbf{J}_2$. This can be determined by the dimensions of $\mathcal{N}_1$. f $dim\mathcal{N}_1 = 2$, then $\mathbf{A}$ must similar to $\lambda\mathbf{I}$ (why?).

- Let us consider the other case: $dim\mathcal{N}_1 = 1$. We shall find generalized vectors and transform $\mathbf{A}$ to a Jordan form $\mathbf{J}_2$.

  1. First, by Caley-Hamilton theorem, $dim\mathcal{N}_2 = 2$. We have $\mathcal{N}_1 \subset \mathcal{N}_2$. Let us choose any $\mathbf{v}_2 \in \mathcal{N}_2 \setminus \mathcal{N}_1$.
  2. We define $\mathbf{v}_1 = (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_2$. Then $(\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_1 = (\mathbf{A} - \lambda\mathbf{I})^2\mathbf{v}_2 = 0$. Thus, $\mathbf{v}_1 \in \mathcal{N}_1$ and $\mathbf{v}_2 \in \mathcal{N}_2 \setminus \mathcal{N}_1$. We get that $\mathbf{v}_1$ and $\mathbf{v}_2$ are independent. Under $[\mathbf{v}_1, \mathbf{v}_2]$, the matrix $\mathbf{A}$ is transformed to $\mathbf{J}_2(\lambda)$.

You may wonder whether the choice of $\mathbf{v}_1$ and $\mathbf{v}_2$ is unique? It is clear that the choice of $\mathbf{v}_1$ is unique (up to scalar multiplication) (why?). Otherwise both of them will be independent eigenvectors corresponding to λ and thus $\mathbf{A} = \lambda\mathbf{I}$. How about we choose $\bar{\mathbf{v}}_2 = \mathbf{v}_2 + \beta\mathbf{v}_1$. Define $\bar{\mathbf{v}}_1 = (\mathbf{A} - \lambda\mathbf{I})\bar{\mathbf{v}}_2$. We see that $(\mathbf{A} - \lambda\mathbf{I})\bar{\mathbf{v}}_1 = (\mathbf{A} - \lambda\mathbf{I})^2\bar{\mathbf{v}}_2 = 0$. Also, $\bar{\mathbf{v}}_1$ and $\bar{\mathbf{v}}_2$ are independent. Thus, both $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2]$ and $\bar{\mathbf{V}} = [\bar{\mathbf{v}}_1, \bar{\mathbf{v}}_2]$ can transform $\mathbf{A}$ to the same Jordan form. The choice of $\mathbf{V}$ is not unique. You may check that the matrix

$$\bar{\mathbf{V}} = [\mathbf{v}_1, \mathbf{v}_2 + \beta\mathbf{v}_1] = [\mathbf{v}_1, \mathbf{v}_2]\mathbf{S}, \quad \mathbf{S} = \begin{pmatrix} 1 & \beta \\ 0 & 1 \end{pmatrix}.$$

Then you can double check that

$$\mathbf{S}^{-1}\mathbf{J}_2\mathbf{S} = \mathbf{J}_2.$$

From this 2×2 system, we conclude that the structure of Jordan is unique, but the choice of the similarity transform is not unique.

Example Suppose $\mathbf{A}$ is a 6×6 matrix with only one eigenvalue λ which has multiplicity 6. We have two tasks:

- Determine the structure of the Jordan form of $\mathbf{A}$;
- Find generalized vectors to transform $\mathbf{A}$ to a Jordan form $\mathbf{J}$.

Determine the structure There are many possible Jordan forms corresponding to $\mathbf{A}$. For instance, $\mathbf{J}_6$, $\mathbf{J}_1 \otimes \mathbf{J}_5$, $\mathbf{J}_2 \otimes \mathbf{J}_4$, $\mathbf{J}_1 \otimes \mathbf{J}_2 \otimes \mathbf{J}_3$, $\mathbf{J}_3 \otimes \mathbf{J}_3$, etc. Notice that if $\mathbf{A} \sim \mathbf{J}$, then $\mathbf{A}$ and $\mathbf{J}$ have the same Jordan block structure. But the structure of a Jordan form $\mathbf{J}(\lambda)$ can be read easily from the dimensions of the Kernel of $(\mathbf{J}(\lambda) - \lambda\mathbf{I})^k$. Let us call

$$\mathcal{N}_k := Ker((\mathbf{J} - \lambda\mathbf{I})^k), \quad d_k = dim\mathcal{N}_k.$$

We define $\mathcal{N}_0 = \{0\}$ and $d_0 = 0$. Let us investigate how d_k reflects the structure of Jordan blocks. The kernels $\mathcal{N}_k$ have have the following properties:

- $\mathcal{N}_{k-1} \subset \mathcal{N}_k$.
- there exists a number d such that $\mathcal{N}_{d-1} \neq \mathcal{N}_d$ but $\mathcal{N}_d = \mathcal{N}_{d+1}$.

- $d_k - d_{k-1}$ is the number of Jordan blocks of size at least k;
- the number of Jordan blocks of size k is $m_k := (d_k - d_{k-1}) - (d_{k+1} - d_k)$.

Let us explain these statements by the following examples:

1. Suppose $\mathbf{J} = \mathbf{J}_1 \otimes \mathbf{J}_5$. That is

$$\begin{aligned} \mathbf{J}\mathbf{e}_1 &= \lambda \mathbf{e}_1 & \mathbf{J}\mathbf{e}_2 &= \lambda \mathbf{e}_2 \\ \mathbf{J}\mathbf{e}_3 &= \lambda \mathbf{e}_3 + \mathbf{e}_2 & \mathbf{J}\mathbf{e}_4 &= \lambda \mathbf{e}_4 + \mathbf{e}_3 \\ \mathbf{J}\mathbf{e}_5 &= \lambda \mathbf{e}_5 + \mathbf{e}_4 & \mathbf{J}\mathbf{e}_6 &= \lambda \mathbf{e}_6 + \mathbf{e}_5. \end{aligned}$$

   Thus,

$$\begin{aligned} \mathcal{N}_1 &=< \mathbf{e}_1, \mathbf{e}_2 >, & \mathcal{N}_2 &=< \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3 > \\ \mathcal{N}_3 &=< \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4 >, & \mathcal{N}_4 &=< \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5 > \\ \mathcal{N}_5 &=< \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5, \mathbf{e}_6 >= \mathcal{N}_6 \end{aligned}$$

   Hence

$$d_0 = 0, \quad d_1 = 2, \quad d_2 = 3, \quad , d_3 = 4, \quad d_4 = 5, \quad d_5 = d_6 = 6.$$

2. Suppose $\mathbf{J} = \mathbf{J}_1 \otimes \mathbf{J}_2 \otimes \mathbf{J}_3$. That is

$$\begin{aligned} \mathbf{J}\mathbf{e}_1 &= \lambda \mathbf{e}_1 & \mathbf{J}\mathbf{e}_2 &= \lambda \mathbf{e}_2 \\ \mathbf{J}\mathbf{e}_3 &= \lambda \mathbf{e}_3 + \mathbf{e}_2 & \mathbf{J}\mathbf{e}_4 &= \lambda \mathbf{e}_4 \\ \mathbf{J}\mathbf{e}_5 &= \lambda \mathbf{e}_5 + \mathbf{e}_4 & \mathbf{J}\mathbf{e}_6 &= \lambda \mathbf{e}_6 + \mathbf{e}_5. \end{aligned}$$

   Thus,

$$\begin{aligned} \mathcal{N}_1 &=< \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_4 >, \\ \mathcal{N}_2 &=< \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_4, \mathbf{e}_3, \mathbf{e}_5 > \\ \mathcal{N}_3 &=< \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5, \mathbf{e}_6 > . \end{aligned}$$

   Hence

$$d_0 = 0, \quad d_1 = 3, \quad d_2 = 5, \quad , d_3 = 6.$$

Given a matrix $\mathbf{A}$ with eigenvalue λ with miltiplicity 6, we can read its Jordan block structure from

$$d_k := dim \mathcal{N}_k, \quad \mathcal{N}_k := \text{ Ker } (\mathbf{A} - \lambda \mathbf{I})^k.$$

Find generalized eigenvectors to transform A to J

1. Suppose $d_1 = 2, d_2 = 3, ..., d_5 = 6$. This is equivalent to $m_1 = 1$, $m_2 = \cdots = m_4 = 0$ and $m_5 = 1$. That is $\mathbf{A} \sim \mathbf{J}_1 \otimes \mathbf{J}_5$. To find generalized eigenvectors

$$\begin{aligned} (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_1 &= 0 & (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_2 &= 0 \\ (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_3 &= \mathbf{v}_2 & (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_4 &= \mathbf{v}_3 \\ (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_5 &= \mathbf{v}_4 & (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_6 &= \mathbf{v}_5 \end{aligned}$$

   we see that $\mathcal{N}_1 =< \mathbf{v}_1, \mathbf{v}_2 >, \mathcal{N}_2 =< \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3 >, ..., \mathcal{N}_5 =< \mathbf{v}_1, ..., \mathbf{v}_6 >$.

   (a) We choose $\mathbf{v}_6 \in \mathcal{N}_5 \setminus \mathcal{N}_4$,

   (b) We set
$$\mathbf{v}_i = (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_{i+1}, i = 5, 4, 3, 2.$$

   (c) You can check that $\mathbf{v}_2 \in \mathcal{N}_1$. Since $dim\mathcal{N}_1 = 2$, we can find another $\mathbf{v}_1 \in \mathcal{N}_1$ which is independent of $\mathbf{v}_1$.

2. Suppose $d_1 = 3, d_2 = 5$ and $d_3 = 6$. That is $m_1 = 2\times3-0-5 = 1$, $m_2 = 2\times5-3-6 = 1$ and $m_3 = 2\times6-5-6 = 1$, or $\mathbf{A} \sim \mathbf{J}_1 \otimes \mathbf{J}_2 \otimes \mathbf{J}_3$. We want to find generalized eigenvectors $\mathbf{v}_1, ..., \mathbf{v}_6$ satisfying

$$\begin{aligned} (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_1 &= 0 \\ (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_2 &= 0 \\ (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_3 &= \mathbf{v}_2 \\ (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_4 &= 0 \\ (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_5 &= \mathbf{v}_4 \\ (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_6 &= \mathbf{v}_5 \end{aligned}$$

   That is

$$\begin{aligned} \mathcal{N}_1 &=< \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_4 > \\ \mathcal{N}_2 &=< \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_4, \mathbf{v}_3, \mathbf{v}_5 > \\ \mathcal{N}_3 &=< \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_4, \mathbf{v}_3, \mathbf{v}_5, \mathbf{v}_6 > . \end{aligned}$$

   (a) We start from $\mathcal{N}_1 = Ker(\mathbf{A} - \lambda\mathbf{I})$. From $dim\mathcal{N}_1 = 3$, we find three independent vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_4 \in \mathcal{N}_1$ by solving $(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}$.

   (b) Next we solve
$$(\mathbf{A} - \lambda\mathbf{I})^2\mathbf{v} = 0.$$
   The dimension of this solution space $\mathcal{N}_2$ is 5 by our assumption. From $\mathcal{N}_1 \subset \mathcal{N}_2$, we choose two independent vectors $\mathbf{v}_3, \mathbf{v}_5 \in \mathcal{N}_2 \setminus \mathcal{N}_1$ and reset $\mathbf{v}_2 := (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_3$. Finally, the space $\mathcal{N}_3$ is the whole space $\mathbb{C}^6$. We choose $\mathbf{v}_6 \in N_3 \setminus \mathcal{N}_2$ and reset $\mathbf{v}_5 = (\mathbf{A}-\lambda\mathbf{I})\mathbf{v}_6$ and $\mathbf{v}_4 = (\mathbf{A}-\lambda\mathbf{I})\mathbf{v}_5$. With these choices of $\mathbf{v}_1, ..., \mathbf{v}_6$, $\mathbf{A}$ is transformed to $\mathbf{J}_1\otimes\mathbf{J}_2\otimes\mathbf{J}_3$.

As you can see from the construction, the choice of $[\mathbf{v}_1, \cdots, \mathbf{v}_6]$ is not unique. But the structure of the Jordan blocks $\mathbf{A} \sim \mathbf{J}_1 \otimes \mathbf{J}_2 \otimes \mathbf{J}_3$ is unique. (Why?)

Homework Find $\mathbf{V}$ to transform $\mathbf{A}$ to its Jordan normal form:

1. $\mathbf{A} = \begin{bmatrix} 2 & 2 & 3 \\ 1 & 1 & 3 \\ -1 & -1 & -2 \end{bmatrix}$.

2. $\mathbf{A} = \begin{bmatrix} -1 & 2 & -3 \\ 7 & 4 & 7 \\ -1 & -1 & 2 \end{bmatrix}$.

3. $\mathbf{A} = \begin{bmatrix} -2 & 3 & 3 \\ 2 & -2 & 2 \\ -3 & -3 & -8 \end{bmatrix}$.

4. $\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & -1 \\ 1 & -1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & -1 \\ -1 & 1 & 0 & 0 & 1 \end{bmatrix}$.

5. $\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \\ -1 & -1 & 1 & 1 & -1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$.

4.5.2 Outline of Spectral Theory

We assume $\mathbf{A}$ is an $n \times n$ matrix in $\mathbb{C}^n$.

Theorem 4.3 (Caley-Hamilton)**.** *Let $p_A(\lambda) := det(\lambda\mathbf{I} - \mathbf{A})$ be the characteristic polynomial of $\mathbf{A}$. Then $p_A(\mathbf{A}) = \mathbf{0}$.*

Proof. 1. We use the adjugate matrix property. The adjugate maytrix $adj(M)$ of a matrix M is defined to be the transpose of the cofactor matrix of M. The i-j entry of the cofactor matrix M_{ij} is the determinant of the $(n-1) \times (n-1)$ matrix which eliminate the ith row and jth column of the matrix M. The adjugate matrix has the following property:

$$adj(\mathbf{M}) \cdot \mathbf{M} = \mathbf{M} \cdot adj(\mathbf{M}) = det(\mathbf{M})\mathbf{I}_n.$$

Applying this property to $\mathbf{M} = \lambda\mathbf{I}_n - \mathbf{A}$, we get

$$(\lambda\mathbf{I}_n - \mathbf{A}) \cdot adj(\lambda\mathbf{I} - \mathbf{A}) = det(\lambda\mathbf{I}_n - \mathbf{A})\mathbf{I}_n.$$

2. The right-hand side is
$$det(\lambda \mathbf{I}_n - \mathbf{A})\mathbf{I}_n = \sum_{i=0}^{n} \lambda^i c_i \mathbf{I}_n.$$

3. Notice that the matrix $adj(\lambda \mathbf{I} - \mathbf{A})$ can be expressed as polynomial in λ of degree $(n-1)$:
$$adj(\lambda \mathbf{I} - \mathbf{A}) = \sum_{i=0}^{n-1} \mathbf{B}_i \lambda^i.$$
Thus, the left-hand side is
$$(\lambda \mathbf{I}_n - \mathbf{A}) \cdot adj(\lambda \mathbf{I} - \mathbf{A}) = \sum_{i=0}^{n-1} (\lambda \mathbf{I} - \mathbf{A}) \cdot \mathbf{B}_i \lambda^i$$
$$= \lambda^n \mathbf{B}_{n-1} + \sum_{i=1}^{n-1} \lambda^i (\mathbf{B}_{i-1} - \mathbf{A}\mathbf{B}_i) - \mathbf{A}\mathbf{B}_0.$$

4. By comparing both polynomials, we obtain
$$\mathbf{I}_n = \mathbf{B}_{n-1}, \quad c_i \mathbf{I}_n = \mathbf{B}_{i-1} - \mathbf{A}\mathbf{B}_i, 1 \le i \le n-1, \quad c_0 \mathbf{I}_n = -\mathbf{A}\mathbf{B}_0.$$

5. Multiply the above ith equation by $\mathbf{A}^i$ them sum over i from 0 to n, we obtain
$$\sum_{i=0}^{n} c_i \mathbf{A}^i = \mathbf{A}^n \mathbf{B}_{n-1} + \sum_{i=1}^{n-1} \mathbf{A}^i (\mathbf{B}_{i-1} - \mathbf{A}\mathbf{B}_i) - \mathbf{A}\mathbf{B}_0 = 0.$$

□

Theorem 4.4. *There exists a minimal polynomial p_m which is a factor of p_A and $p_m(\mathbf{A}) = 0$.*

Theorem 4.5 (Fundamental Theorem of Algebra)**.** *Any polynomial $p(\lambda)$ over $\mathbb{C}$ of degree m can be factorized as*
$$p(\lambda) = a \prod_{i=1}^{m} (\lambda - \lambda_i)$$
for some constant $a \neq 0$ and $\lambda_1, ..., \lambda_m \in \mathbb{C}$. This factorization is unique.

Definition 4.6. *Let $\mathbf{A} : \mathbb{C}^n \to \mathbb{C}^n$. A subspace $\mathcal{V} \subset \mathbb{C}^n$ is called an invariant subspace of the linear map $\mathbf{A}$ if $\mathbf{A}\mathcal{V} \subset \mathcal{V}$.*

Definition 4.7. *A vector space $\mathcal{V}$ is said to be the direct sum of its two subspaces $\mathcal{V}_1$ and $\mathcal{V}_2$ if for any $\mathbf{v} \in \mathcal{V}$ there exist two unique vectors $\mathbf{v}_i \in \mathcal{V}_i$, $i = 1, 2$ such that $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$. We denote it by $\mathcal{V} = \mathcal{V}_1 \oplus \mathcal{V}_2$.*

Remark 4.1. *We also use the notation $\mathcal{V} = \mathcal{V}_1 + \mathcal{V}_2$ for the property: any $\mathbf{v} \in \mathcal{V}$ can be written as $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$ for some $\mathbf{v}_i \in \mathcal{V}_i$, $i = 1, 2$. Notice that $\mathcal{V} = \mathcal{V}_1 \oplus \mathcal{V}_2$ if and only if $\mathcal{V} = \mathcal{V}_1 + \mathcal{V}_2$ and $\mathcal{V}_1 \cap \mathcal{V}_2 = \{0\}$.*

Lemma 4.1. *Suppose p and q are two polynomials over $\mathbb{C}$ and are relatively prime (i.e. no common roots). Then there exist two other polynomials a and b such that*

$$ap + bq = 1.$$

Lemma 4.2. *Suppose p and q are two polynomials over $\mathbb{C}$ and are relatively prime (i.e. no common roots). Let $\mathcal{N}_p := Ker(p(\mathbf{A}))$, $\mathcal{N}_q := Ker(q(\mathbf{A}))$ and $\mathcal{N}_{pq} := Ker(p(\mathbf{A})q(\mathbf{A}))$. Then*

$$\mathcal{N}_{pq} = \mathcal{N}_p \oplus \mathcal{N}_q.$$

Proof. From $ap + bq = 1$ we get

$$a(\mathbf{A})p(\mathbf{A}) + b(\mathbf{A})q(\mathbf{A}) = \mathbf{I}.$$

For any $\mathbf{v} \in \mathcal{N}_{pq}$, acting the above operator formula to $\mathbf{v}$, we get

$$\mathbf{v} = a(\mathbf{A})p(\mathbf{A})\mathbf{v} + b(\mathbf{A})q(\mathbf{A})\mathbf{v} := \mathbf{v}_2 + \mathbf{v}_1.$$

We claim that $\mathbf{v}_1 \in \mathcal{N}_p$, whereas $\mathbf{v}_2 \in \mathcal{N}_q$. This is because

$$p(A)\mathbf{v}_1 = p(A)b(\mathbf{A})q(\mathbf{A})\mathbf{v} = b(\mathbf{A})p(\mathbf{A})q(\mathbf{A})\mathbf{v} = 0.$$

Similar argument for proving $\mathbf{v}_2 \in \mathcal{N}_q$. To see this is a direct sum, suppose $\mathbf{v} \in \mathcal{N}_p \cap \mathcal{N}_q$. Then

$$\mathbf{v} = a(\mathbf{A})p(\mathbf{A})\mathbf{v} + b(\mathbf{A})q(\mathbf{A})\mathbf{v} = 0.$$

Hence $\mathcal{N}_p \cap \mathcal{N}_q = \{0\}$. □

Corollary 4.2. *Suppose a polynomial p is factorized as $p = p_1 \cdots p_s$ with $p_1, ..., p_s$ are relatively prime (no common roots). Let $\mathcal{N}_{p_i} := Kerp_i(\mathbf{A})$. Then*

$$\mathcal{N}_p = \mathcal{N}_{p_1} \oplus \cdots \oplus \mathcal{N}_{p_s}.$$

Theorem 4.6 (Spectral Decomposition)**.** *Let p_m be the minimal polynomial of $\mathbf{A}$. Suppose p_m can be factorized as*

$$p_m(\lambda) = \prod_{i=1}^{s} p_i(\lambda) = \prod_{i=1}^{s} (\lambda - \lambda_{k_i})^{m_i}$$

with $\lambda_{k_i} \neq \lambda_{k_j}$ for $i \neq j$. Let $\mathcal{N}_{k_i} = Ker(\mathbf{A} - \lambda_{k_i}\mathbf{I})^{m_i}$. Then

- *$\mathcal{N}_{k_i}$ is invariant under $\mathbf{A}$,*
- *$\mathbb{C}^n = \mathcal{N}_{k_1} \oplus \cdots \oplus \mathcal{N}_{k_s}$.*

Important application of spectral decomposition of a matrix is to compute $e^{t\mathbf{A}}$, which will be the fundamental solution of the ODE: $\mathbf{y}' = \mathbf{A}\mathbf{y}$. It is easy to compute $e^{t\mathbf{A}}$ if $\mathbf{A}$ is a diagonal matrix or a Jordan matrix. Through spectral decomposition, we can compute $e^{t\mathbf{A}}$ for general $\mathbf{A}$.

4.6 Fundamental Matrices and $\exp(t\mathbf{A})$

4.6.1 Fundamental matrices

We have seen that the general solution to the initial value problem:

$$\mathbf{y}'(t) = \mathbf{A}\mathbf{y}(t),\ \mathbf{y}(0) = \mathbf{y}_0,$$

can be express as $y(t) = C_1\mathbf{y}_1(t) + \cdots + C_n\mathbf{y}_n$, where $\mathbf{y}_1, .., \mathbf{y}_n$ are n independent solutions. The matrix $\mathbf{Y}(t) = [\mathbf{y}_1(t), \cdots, \mathbf{y}_n(t)]$ is called a fundamental matrix. The solution $\mathbf{y}(t)$ is expressed as $\mathbf{y}(t) = \mathbf{Y}(t)\mathbf{C}$, where $\mathbf{C} = (C_1, ..., C_n)^t$. By plugging $\mathbf{y}(t) = \mathbf{Y}(t)\mathbf{C}$ into the equation $\mathbf{y}' = \mathbf{A}\mathbf{y}$, we obtain

$$\mathbf{Y}'\mathbf{C} = \mathbf{A}\mathbf{Y}\mathbf{C}$$

This is valid for all $\mathbf{C}$. We conclude that the fundamental matrix satisfies

$$\mathbf{Y}'(t) = \mathbf{A}\mathbf{Y}(t). \tag{4.9}$$

From $\mathbf{y}(0) = \mathbf{Y}(0)\mathbf{C}$, we obtain $\mathbf{C} = \mathbf{Y}(0)^{-1}\mathbf{y}(0)$. Thus,

$$\mathbf{y}(t) = \mathbf{Y}(t)\mathbf{Y}(0)^{-1}\mathbf{y}(0).$$

The matrix $\Phi(t) := \mathbf{Y}(t)\mathbf{Y}(0)^{-1}$ is still a fundamental matrix and satisfies $\Phi(0) = \mathbf{I}$. We shall see that $\Phi(t) = \exp(t\mathbf{A})$ in the next section.

Homework

1. Consider an $n \times n$ matrix ODE
$$\mathbf{Y}'(t) = \mathbf{A}\mathbf{Y}(t)$$
Let $W(t) = det\mathbf{Y}(t)$. Show that
$$W'(t) = tr(A)W(t)$$
where $tr(A) := \sum_i a_{ii}$.
Hint: $(det\mathbf{A})' = \sum_{i,j} a'_{ij}A_{ij}$, where A_{ij} is the cofactor of $\mathbf{A}$.

4.6.2 Computing $\exp(\mathbf{A})$

The exponential function is defined as a power series, which involves the concept of norm and limit in the space of $n \times n$ matrices. Let $\mathcal{M}_n = \{\mathbf{A} | \mathbf{A}$ is an $n \times n$ complex-valued matrix$\}$. We define a *norm* on $\mathcal{M}_n$ by

$$\|A\| := \left(\sum_{i,j} |a_{ij}|^2\right)^{1/2}.$$

The norm $\|\cdot\|$ has the following properties:

- $\|\mathbf{A}\| \geq 0$ and $\|\mathbf{A}\| = 0$ if and only if $\mathbf{A} = 0$.
- $\|\alpha\mathbf{A}\| = |\alpha|\|\mathbf{A}\|$ for any $\alpha \in \mathbb{C}$.
- $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$.

In addition, the matrix space $\mathcal{M}_n$ is an algebra with the matrix multiplication. It satisfies

- $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$.

The proof of the last assertion is the follows.

$$\begin{aligned}
\|\mathbf{AB}\|^2 &= \sum_{i,j} |\sum_k a_{ik}b_{kj}|^2 \\
&\leq \sum_{i,j} (\sum_k |a_{ik}|^2)(\sum_k |b_{kj}|^2) \\
&= \sum_i (\sum_k |a_{ik}|^2) \sum_j (\sum_k |b_{kj}|^2) \\
&= \|\mathbf{A}\|^2\|\mathbf{B}\|^2
\end{aligned}$$

With this norm, we can talk about theory of convergence. The space $\mathcal{M}_n$ is equivalent to $\mathbb{C}^{n^2}$. Thus, it is complete. This means that every Cauchy sequence converges to a point in $\mathcal{M}_n$.

Now we define the exponential function in $\mathcal{M}_n$ as the follows.

$$\exp(\mathbf{A}) := \sum_{n=0}^{\infty} \frac{1}{n!}\mathbf{A}^n. \tag{4.10}$$

Theorem 4.7. *The exponential function has the following properties:*

- $\exp(\mathbf{A})$ *is well-defined.*
- *The function* $\exp(t\mathbf{A})$ *is differentiable and* $\frac{d}{dt}\exp(t\mathbf{A}) = \mathbf{A}\exp(t\mathbf{A})$.
- $\exp(\mathbf{0}) = \mathbf{I}$.

Proof. 1. This series converges because $\mathcal{M}_n$ is complete and this series is a Cauchy series:

$$\|\sum_n^m \frac{1}{k!}\mathbf{A}^k\| \leq \sum_n^m \frac{1}{k!}\|\mathbf{A}\|^k < \varepsilon,$$

if $n < m$ are sufficiently enough.

2. Notice that the series

$$\exp(t\mathbf{A}) = \sum_{n=0}^{\infty} \frac{1}{n!}t^n\mathbf{A}^n.$$

convergence uniformly for t in any bounded set in $\mathbb{R}$. Further, the function $\exp(t\mathbf{A})$ is differentiable in t. This is because the series obtained by the term-by-term differentiation

$$\sum_{n=1}^{\infty} \frac{1}{(n-1)!} t^{n-1} \mathbf{A}^n$$

converges uniformly for t in any bounded set in $\mathbb{R}$. And the derivative of $\exp(t\mathbf{A})$ is the term-by-term differentiation of the original series:

$$\begin{aligned} \frac{d}{dt}\exp(t\mathbf{A}) &= \sum_{n=1}^{\infty} \frac{1}{(n-1)!} t^{n-1} \mathbf{A}^n \\ &= \sum_{n=1}^{\infty} \frac{1}{(n-1)!} t^{n-1} \mathbf{A}^{n-1}\mathbf{A} \\ &= \mathbf{A}\exp(t\mathbf{A}) = \exp(t\mathbf{A})\mathbf{A}. \end{aligned}$$

□

We have seen that the fundamental solution $\mathbf{Y}(t)$ of the equation $\mathbf{y}' = \mathbf{A}\mathbf{y}$ satisfies $\mathbf{Y}' = \mathbf{A}\mathbf{Y}$. From the above theorem, we see that $\exp(t\mathbf{A})$ is a fundamental solution satisfying $\exp(\mathbf{0}) = \mathbf{I}$.

Below, we compute $\exp(t\mathbf{A})$ for some special $\mathbf{A}$.

1. $\mathbf{A} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$. In this case,

$$\mathbf{A}^n = \begin{pmatrix} \lambda_1^n & 0 \\ 0 & \lambda_2^n \end{pmatrix}$$

and

$$\exp(t\mathbf{A}) = \begin{pmatrix} e^{t\lambda_1} & 0 \\ 0 & e^{t\lambda_2} \end{pmatrix}.$$

If λ_1 and λ_2 are complex conjugate and $\lambda_1 = \alpha + i\omega$, then

$$\exp(t\mathbf{A}) = e^{\alpha t} \begin{pmatrix} \cos\omega t + i\sin\omega t & 0 \\ 0 & \cos\omega t - i\sin\omega t \end{pmatrix}.$$

2. $\mathbf{A} = \begin{pmatrix} 0 & -\omega \\ \omega & 0 \end{pmatrix}$. In this case,

$$\begin{aligned} \mathbf{A}^2 &= \begin{pmatrix} -\omega^2 & 0 \\ 0 & -\omega^2 \end{pmatrix} \\ \mathbf{A}^3 &= \begin{pmatrix} 0 & \omega^3 \\ -\omega^3 & 0 \end{pmatrix} \\ \mathbf{A}^4 &= \begin{pmatrix} \omega^4 & 0 \\ 0 & \omega^4 \end{pmatrix} \end{aligned}$$

Hence,

$$\exp(t\mathbf{A}) = \sum_n \frac{1}{n!} t^n \mathbf{A}^n = \begin{pmatrix} \cos\omega t & -\sin\omega t \\ \sin\omega t & \cos\omega t \end{pmatrix}$$

3. $\mathbf{A} = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$. The matrix $\mathbf{A} = \lambda\mathbf{I} + \mathbf{N}$, where

$$\mathbf{N} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

is called a nilponent matrix. $\mathbf{N}$ has the property

$$\mathbf{N}^2 = 0.$$

Thus,

$$\mathbf{A}^n = (\lambda\mathbf{I} + \mathbf{N})^n = \lambda^n\mathbf{I} + n\lambda^{n-1}\mathbf{N}$$

With this,

$$\begin{aligned} \exp(t\mathbf{A}) &= \sum_{n=0}^{\infty} \frac{1}{n!} t^n \mathbf{A}^n \\ &= \sum_{n=0}^{\infty} \frac{1}{n!} t^n (\lambda^n\mathbf{I} + n\lambda^{n-1}\mathbf{N}) \\ &= \exp(\lambda t)\mathbf{I} + \sum_{n=1}^{\infty} \frac{1}{(n-1)!} \lambda^{n-1} t^n \mathbf{N} \\ &= \exp(\lambda t)\mathbf{I} + t\exp(t\lambda)\mathbf{N} \\ &= \begin{pmatrix} e^{\lambda t} & te^{\lambda t} \\ 0 & e^{\lambda t} \end{pmatrix} \end{aligned}$$

For general 2×2 matrices $\mathbf{A}$, we have seen that there exists a matrix $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2]$ such that

$$\mathbf{AV} = \mathbf{V\Lambda}$$

where $\mathbf{\Lambda}$ is either diagonal matrix (case 1) or a Jordan matrix (Case 3). Notice that

$$\mathbf{A}^n = (\mathbf{V\Lambda V}^{-1})^n = \mathbf{V\Lambda}^n\mathbf{V}^{-1}$$

Hence, the corresponding exponential function becomes

$$\begin{aligned} \exp(t\mathbf{A}) &= \sum_{n=0}^{\infty} \frac{1}{n!} t^n \mathbf{A}^n \\ &= \sum_{n=0}^{\infty} \frac{1}{n!} t^n \mathbf{V\Lambda}^n\mathbf{V}^{-1} \\ &= \mathbf{V}(\sum_{n=0}^{\infty} \frac{1}{n!} t^n \mathbf{\Lambda}^n)\mathbf{V}^{-1} \\ &= \mathbf{V}\exp(t\mathbf{\Lambda})\mathbf{V}^{-1} \end{aligned}$$

Revisit fundamental matrix

1. We recall that a fundamental matrix of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ is

$$\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n],$$

where $\mathbf{y}_1, ..., \mathbf{y}_n$ are independent solutions of $\mathbf{y}' = \mathbf{A}\mathbf{y}$. The relation between $\mathbf{Y}(t)$ and $\exp(t\mathbf{A})$ is

$$\mathbf{Y}(t)\mathbf{Y}(0)^{-1} = \exp(t\mathbf{A}).$$

This is because any solution can be expressed uniquely by

$$\mathbf{y}(t) = \sum_{i=1}^{n} c_i \mathbf{y}_i(t) = \mathbf{Y}\mathbf{c}.$$

where $\mathbf{c}$ is given by

$$\mathbf{y}(0) = \mathbf{Y}(0)\mathbf{c}, \quad \text{or } \mathbf{c} = \mathbf{Y}(0)^{-1}.$$

Thus, $\mathbf{y}(t)$ can be represented as

$$\mathbf{y}(t) = \mathbf{Y}(t)\mathbf{Y}(0)^{-1}\mathbf{y}_0.$$

We have also seen that

$$\mathbf{y}(t) = \exp(t\mathbf{A})\mathbf{y}_0.$$

This is valid for all $\mathbf{y}_0$. Thus, $\exp(t\mathbf{A}) = \mathbf{Y}(t)\mathbf{Y}(0)^{-1}$.

2. The fundamental matrix is not unique. If both $\mathbf{Y}(t)$ and $\mathbf{Z}(t)$ are fundamental matrices of $\mathbf{y}' = \mathbf{A}\mathbf{y}$, there must be a constant matrix $\mathbf{C}$ such that $\mathbf{Z}(t) = \mathbf{Y}(t)\mathbf{C}$. This follows from the previous result:

$$\mathbf{Y}(t) = \mathbf{Y}(0)^{-1} = \mathbf{Z}(t) = \mathbf{Z}(0)^{-1} = \exp(t\mathbf{A}).$$

Thus, $\mathbf{C} = \mathbf{Y}(o)^{-1}\mathbf{Z}(0)$. In particular, $\exp(t\mathbf{A})$ is a fundamental matrix.

3. A particular fundamental matrix is

$$\mathbf{Y}(t) = \mathbf{V}\exp(t\mathbf{\Lambda}).$$

This is because

$$\mathbf{V}\exp(t\mathbf{\Lambda}) = \exp(t\mathbf{A})\mathbf{V}$$

and $\exp(t\mathbf{A})$ is a fundamental matrix. For 2×2 system, in the case of Jordan form, the fundamental matrix $\mathbf{Y}(t)$ is given by

$$\begin{aligned}
[\mathbf{y}_1(t), \mathbf{y}_2(t)] &= \mathbf{Y}(t) = [\mathbf{v}_1, \mathbf{v}_2]\exp(t\mathbf{\Lambda}) \\
&= [\mathbf{v}_1, \mathbf{v}_2]\begin{pmatrix} e^{\lambda t} & te^{\lambda t} \\ 0 & e^{\lambda t} \end{pmatrix} \\
&= [e^{\lambda t}\mathbf{v}_1, te^{\lambda t}\mathbf{v}_1 + e^{\lambda t}\mathbf{v}_2].
\end{aligned}$$

This is identical to the fundamental solution we obtained before.

Homeworks.

1. Find $\exp(t\mathbf{J}_k(\lambda))$ for general k. Here, $\mathbf{J}_k(\lambda)$ is the Jordan matrix of size k.
2. Compute $\exp(t\mathbf{A})$ with
$$\mathbf{A} = \begin{pmatrix} 0 & -\omega_3 & -\omega_2 \\ \omega_3 & 0 & -\omega_1 \\ \omega_2 & \omega_1 & 0 \end{pmatrix}$$
3. B-D, pp. 420: 3,18
4. B-D, pp. 428, 6,17,18
5. Show that if $\mathbf{AB} = \mathbf{BA}$, then $\exp(\mathbf{A}+\mathbf{B}) = \exp(\mathbf{A})\exp(\mathbf{B})$. In particular, use this result to show $\exp((t-s)\mathbf{A}) = \exp(t\mathbf{A})\exp(s\mathbf{A})^{-1}$.
6. If $\mathbf{A} \neq \mathbf{B}$, show that $\exp(t(\mathbf{A}+\mathbf{B})) - \exp(t\mathbf{A})\exp(t\mathbf{B}) = O(t^2)$ for small t.

4.6.3 Linear Stability Analysis

Consider the $n \times n$ linear system with constant coefficients

$$\mathbf{y}' = \mathbf{A}\mathbf{y}. \tag{4.11}$$

The state $\mathbf{0}$ is an equilibrium state of this system.

Definition 4.8. *The equilibrium $\mathbf{0}$ of (4.11) is called stable if for any $\epsilon > 0$, there exists a $\delta > 0$ such that any solution $\mathbf{y}(\cdot, \mathbf{y}_0)$ starting from $\mathbf{y}_0$ with $|\mathbf{y}_0| < \delta$, we have $|\mathbf{y}(t)| \leq \epsilon$ for all $t > 0$. It is called asymptotically stable if it is stable, in addition, there exists a neighborhood $|\mathbf{y}| < \delta$ such that any solution $\mathbf{y}(\cdot, \mathbf{y}_0)$ starting from $\mathbf{y}_0$ with $|\mathbf{y}_0| < \delta$, then $\mathbf{y}(t) \to \mathbf{0}$ as $t \to \infty$. If, in addition, $|\mathbf{y}(t)| \leq Ce^{-\alpha t}$ for some positive constantts C and α, we say $\mathbf{y}(t)$ converges to $\mathbf{0}$ at exponential rate.*

Remark. For 2×2 linear system:

1. The centers are stable, but not asymptotic stable.
2. The sources, spiral sources and saddle points are unstable.
3. The sinks and the spiral sinks are asymptotic stable.

Theorem 4.8. *Consider the linear system with constant coefficients:*

$$\mathbf{y}' = \mathbf{A}\mathbf{y}.$$

1. *The state $\mathbf{0}$ is asymptotically stable if and only if all eigenvalues $\lambda(\mathbf{A})$ satisfy $Re(\lambda(\mathbf{A})) < 0$.*

2. *The state* $\mathbf{0}$ *is stable if and only if all eigenvalues* $\lambda(\mathbf{A})$ *are either (i)* $Re(\lambda(\mathbf{A})) < 0$*, or (ii)* $Re(\lambda(\mathbf{A})) = 0$ *but it is simple.*

Proof. 1. Let us decompose the space $\mathbb{C}^n$ (or $\mathbb{R}^n$) into the invariant subspaces. The matrix $\mathbf{A}$ is just a Jordan block as restricted to these invariant subspaces. The stability or asymptotic stability of the state $\mathbf{0}$ in the whole space $\mathbb{C}^n$ (or $\mathbb{R}^n$) is equivalent to that in all invariant subspaces. Thus, we only need to discuss the case that $\mathbf{A}$ is a Jordan block $\mathbf{J}$.

2. For a Jordan block $\mathbf{J}(\lambda)$ of size k, the corresponding fundamental solutions are

$$\mathbf{v}_1 e^{\lambda t}, (t\mathbf{v}_1 + \mathbf{v}_2)e^{\lambda t}, \cdots, \left(\frac{t^{k-1}}{(k-1)!}\mathbf{v}_1 + \cdots + \mathbf{v}_k\right) e^{\lambda t}$$

where $\mathbf{v}_i$ are the generalized eigenvectors corresponding to $\mathbf{J}$. If $Re(\lambda) < 0$, then $t^j e^{\lambda t}$ decays exponentially fast to 0 for any $j \geq 0$. Thus, $\mathbf{0}$ is asymptotic stable if $Re(\lambda) < 0$ for all eigenvalues λ.

3. Conversely, suppose $\mathbf{0}$ is asymptotic stable. Since the solutions only have the form $t^j e^{\lambda t}$ with $j \geq 0$, and such solution can tend to $\mathbf{0}$ as $t \to \infty$ only when $Re(\lambda) < 0$. Thus $\mathbf{0}$ is asymptotic stable only when all eigenvalues satisfy $Re(\lambda) < 0$.

4. If there exists a λ such that $Re(\lambda) > 0$, then the solution $e^{\lambda t}\mathbf{v} \to \infty$ as $t \to \infty$. The converse is also true.

5. If there exists a λ such that $Re(\lambda) = 0$ and λ is simple, then the corresponding solution is $e^{\lambda t}\mathbf{v}$, where $\mathbf{v}$ is the corresponding eigenvector. Such solution stays bounded.

6. If $Re(\lambda) = 0$ and with multiplicity $k > 1$, then there is a solution of the form $te^{\lambda t}\mathbf{v}$, where $\mathbf{v}$ is a generalized eigenvector. This solution tends to infinity as t tends to infinity. Thus, $\mathbf{0}$ is not stable.

□

4.7 Non-homogeneous Linear Systems

We consider the inhomogeneous linear systems:

$$\mathbf{y}'(t) = \mathbf{A}\mathbf{y}(t) + \mathbf{f}(t),\ \mathbf{y}(0) = \mathbf{y}_0. \tag{4.12}$$

We use variation of parameters to solve this equation. Let $\mathbf{\Phi}(t) = \exp(t\mathbf{A})$ be the fundamental solution for the homogeneous equation. To find a particular solution for the inhomogeneous equation, we consider

$$\mathbf{y}(t) = \mathbf{\Phi}(t)\mathbf{u}(t).$$

We plug this into equation. We get

$$\mathbf{\Phi}'\mathbf{u} + \mathbf{\Phi}\mathbf{u}' = \mathbf{A}\mathbf{\Phi}\mathbf{u} + \mathbf{f}$$

Using $\mathbf{\Phi}' = \mathbf{A}\mathbf{\Phi}$, we get

$$\mathbf{\Phi}\mathbf{u}' = \mathbf{f}$$

Hence, a particular of $\mathbf{u}$ is

$$\mathbf{u}(t) = \int_0^t \mathbf{\Phi}(s)^{-1}\mathbf{f}(s)\, ds$$

Thus a particular solution $\mathbf{y}_p(t)$ is

$$\mathbf{y}_p(t) = \mathbf{\Phi}(t)\int_0^t \mathbf{\Phi}^{-1}(s)\mathbf{f}(s)\, ds = \int_0^t \mathbf{\Phi}(t)\mathbf{\Phi}(s)^{-1}\mathbf{f}(s)\, ds$$

This special solution has 0 initial data. The solution for initial condition $\mathbf{y}(0) = \mathbf{y}_0$ has the following expression:

$$\mathbf{y}(t) = \mathbf{\Phi}(t)\mathbf{y}_0 + \int_0^t \mathbf{\Phi}(t)\mathbf{\Phi}(s)^{-1}\mathbf{f}(s)\, ds \tag{4.13}$$

Notice that the matrix exponential function also satisfies the exponential laws. We can rewrite the above expression as

$$\mathbf{y}(t) = \mathbf{\Phi}(t)\mathbf{y}_0 + \int_0^t \mathbf{\Phi}(t-s)\mathbf{f}(s)\, ds. \tag{4.14}$$

Homeworks.

1. B-D pp. 439: 11, 12.
2. Consider the example of circuit system in subsection 4.2.2. Now, we add another node, say 5 and edges $(1,5),(5,2)$. On edge $(1,5)$, we add a power supply $I(t)$. Derive the equation, find its solution formula. (B-D 439, Figure 7.9.1, problem 13)
3. B-D pp. 422: 14.

Chapter 5

Methods of Laplace Transforms

The method of Laplace transform converts a linear ordinary differential equation with constant coefficients to an algebraic equation. The core of the this differential equation then lies in the roots of the corresponding algebraic equation. In applications, the method of Laplace transform is particular useful to handle general source terms.

5.1 Laplace transform

For function f defined on $[0, \infty)$, we define its Laplace transformation by

$$\mathcal{L}f(s) = F(s) := \int_0^\infty f(t)e^{-st}\,dt.$$

$\mathcal{L}$ is a linear transformation which maps f to F. For those functions f such that

$$|f(t)| \le Ce^{\alpha t} \tag{5.1}$$

for some positive constants C and α, the above improper integral converges *uniformly and absolutely* for complex number s lies in a compact set in $\{s \in \mathbb{C} | Re(s) > \alpha\}$:

$$\int_0^\infty |f(t)e^{-st}|\,dt \le C\int_0^\infty e^{\alpha t}e^{-st}\,dt = \frac{C}{s-\alpha}.$$

Here, we have used that

$$\lim_{t\to\infty} e^{-(s-\alpha)t} = 0$$

due to $Re(s) > \alpha$. We call functions with this growth condition (5.1) admissible. Since the integration allows f being discontinuous, the admissible functions include all piecewise continuous functions. We summarize the class of these admissible functions are those f such that

1. f is bounded and piecewise continuous functions on $[0, \infty)$;
2. there exists an $\alpha \in \mathbb{R}$ and a constant $C > 0$ such that $|f(t)| \le Ce^{\alpha t}$ for all $t \ge 0$.

The image space of the Laplace transform are those (analytic) function $F(s)$ defined on $s \in \mathbb{C}$ with $Re(s) > \alpha$ for some α.

5.1.1 Examples

1. When $f(t) \equiv 1$, $\mathcal{L}(1) = 1/s$.

2. $\mathcal{L}(e^{\lambda t}) = 1/(s-\lambda)$. This is because

$$\mathcal{L}(e^{\lambda t}) = \int_0^\infty e^{\lambda t} e^{-st}\, dt = \int_0^\infty e^{-(s-\lambda)t}\, dt = \frac{1}{s-\lambda}.$$

   Indeed, this is valid for any complex number λ and $s \in \mathbb{C}$ with $Re(s) > \lambda$.

3. When $f(t) = t^n$,

$$\begin{aligned}
\mathcal{L}(t^n) &= \int_0^\infty t^n e^{-st}\, dt = \frac{-1}{s}\int_0^\infty t^n\, de^{-st} \\
&= \frac{-1}{s}(t^n e^{-st})_0^\infty - \int_0^\infty n t^{n-1} e^{-st}\, dt \\
&= \frac{n}{s}\mathcal{L}(t^{n-1}) = \frac{n}{s}\frac{(n-1)}{s}\cdots\frac{1}{s}\mathcal{L}(1) = \frac{n!}{s^{n+1}}.
\end{aligned}$$

   Alternatively,

$$\begin{aligned}
\mathcal{L}(t^n) &= \int_0^\infty t^n e^{-st}\, dt = \int_0^\infty (-\frac{d}{ds})^n e^{-st}\, dt \\
&= (-\frac{d}{ds})^n \int_0^\infty e^{-st}\, dt = (-\frac{d}{ds})^n \frac{1}{s} = \frac{n!}{s^{n+1}}
\end{aligned}$$

4. $\mathcal{L}(t^n e^{\lambda t}) = \frac{n!}{(s-\lambda)^{n+1}}$. Indeed,

$$\mathcal{L}(t^n e^{\lambda t}) = \int_0^\infty t^n e^{\lambda t} e^{-st}\, ds = \int_0^\infty t^n e^{-(s-\lambda)t}\, ds = \frac{n!}{(s-\lambda)^{n+1}}$$

5. $\mathcal{L}(\cos\omega t) = \frac{s}{s^2+\omega^2},\ \mathcal{L}(\sin\omega t) = \frac{\omega}{s^2+\omega^2}$. Indeed,

$$\mathcal{L}(\cos\omega t) = \frac{1}{2}\mathcal{L}(e^{i\omega t} + e^{-i\omega t}) = \frac{1}{2}(\frac{1}{s-i\omega} + \frac{1}{s+i\omega}) = \frac{s}{s^2+\omega^2}.$$

6. The function

$$h(t) = \begin{cases} 1 & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases}$$

   is called the Heaviside function. It has a discontinuity at $t = 0$ with jump $h(0+) - h(0-) = 1$. Its translation $h(t-a)$ has jump at a. The corresponding Laplace transform is

$$\mathcal{L}(h(t-a)) = \int_0^\infty h(t-a) e^{-st}\, dt = \int_a^\infty e^{-st}\, dt = e^{-as}\mathcal{L}(1) = \frac{e^{-as}}{s},$$

   for any $a \geq 0$.

7. We shall apply the method of Laplace transform to solve the initial value problem:

$$y' + y = t,\ y(0) = y_0.$$

We apply Laplace transform both sides.

$$\mathcal{L}(y') = \int_0^\infty e^{-st} y'(t)\, dt = -y(0) + s\int_0^\infty e^{-st} y(t)\, dt$$

Let us denote $\mathcal{L}y = Y$. We have

$$sY - y_0 + Y = \frac{1}{s^2}$$

Hence

$$\begin{aligned} Y(s) &= \frac{1}{s+1}\left(y_0 + \frac{1}{s^2}\right) \\ &= \frac{y_0}{s+1} + \frac{1}{s^2} - \frac{1}{s} + \frac{1}{s+1} \end{aligned}$$

Hence

$$y(t) = y_0 e^{-t} + t - 1 + e^{-t}.$$

5.1.2 Properties of Laplace transform

Let us denote the Laplace transform of f by F. That is, $F = \mathcal{L}f$.

1. $\mathcal{L}$ is linear. This follows from the linearity of integration.

2. $\mathcal{L}$ is one-to-one, that is $\mathcal{L}(f) = 0$ implies $f = 0$. Hence, $\mathcal{L}^{-1}$ exists.
This is indeed not so obvious to prove. I leave it in the homework.

3. Translation: Given f in the admissible class. We set $f(t) = 0$ for $t < 0$. Then for any $a \geq 0$, we have

$$\mathcal{L}(f(t-a)) = e^{-as}F(s).$$

Thus, the term e^{-as} in the s-space represents a translation in the time domain.
On the other hand, a translation on the image space corresponds to a multiplication of an exponential function:

$$\mathcal{L}^{-1}F(s+\lambda) = e^{-\lambda t} f(t).$$

4. Dilation:

$$\mathcal{L}(f(bt)) = \frac{1}{b}F\left(\frac{s}{b}\right),\ \mathcal{L}^{-1}F(bs) = \frac{1}{b}f\left(\frac{t}{b}\right).$$

5. Differentiation:

$$\mathcal{L}(f'(t)) = sF(s) - f(0),\ \mathcal{L}^{-1}F'(s) = -tf(t). \tag{5.2}$$

6. Integration:

$$\mathcal{L}\left(\int_0^t f(\tau)\,d\tau\right) = \frac{F(s)}{s}, \mathcal{L}^{-1}\left(\int_s^\infty F(s_1)\,ds_1\right) = \frac{f(t)}{t},$$

7. Convolution:

$$\mathcal{L}(f * g) = \mathcal{L}(f) \cdot \mathcal{L}(g),$$

where

$$(f * g)(t) = \int_0^t f(\tau)g(t-\tau)\,d\tau$$

Proof.

$$\begin{aligned}
\mathcal{L}(f * g) &= \int_0^\infty e^{-st} \int_0^t f(\tau)g(t-\tau)\,d\tau\,dt \\
&= \int_0^\infty \int_0^t e^{-s\tau} f(\tau) e^{-s(t-\tau)} g(t-\tau)\,d\tau\,dt \\
&= \int_0^\infty d\tau \int_\tau^\infty dt \left(e^{-s\tau} f(\tau) e^{-s(t-\tau)} g(t-\tau)\right) \\
&= \int_0^\infty e^{-s\tau} f(\tau)\,d\tau \int_0^\infty e^{-st} g(t)\,dt = \mathcal{L}(f)\mathcal{L}(g)
\end{aligned}$$

Homeworks.

1. B-D, pp. 313: 26,27.
2. Find the Laplace transforms of
   (a) $\cosh(at)$ (ans. $s/(s^2 - a^2)$).
   (b) $\sinh(at)$, (ans. $a/(s^2 - a^2)$.)
   (c) $(-t)^n f(t)$ (ans. $F^{(n)}(s)$.)
3. B-D,pp. 331: 27.28
4. Find the Laplace transforms of
   (a) $B_0(2t) - B_0(2t-1)$, where $B_0(t) = 1$ for $0 \le t < 1$ and $B_0(t) = 0$ otherwise.
   (b) $f(t) = \sum_{k=0}^\infty B(2t-k)$.
   (c) Let $f_0(t) = t(1-t)$ for $0 \le t < 1$ and $f_0(t) = 0$ elsewhere. Let $f(t)$ be the periodic extension of f_0 with period 1. Find $\mathcal{L}f_0$, $\mathcal{L}f$, $\mathcal{L}f_0'$ and $\mathcal{L}f'$..
5. Prove

$$\mathcal{L}\left(\int_0^t f(\tau)\,d\tau\right) = \frac{F(s)}{s}, \mathcal{L}^{-1}\left(\int_s^\infty F(s_1)\,ds_1\right) = \frac{f(t)}{t},$$

6. Let $f(t)$ be a period function with period p. Let

$$f_0 = \begin{cases} f(t) & \text{for } 0 < t < p \\ 0 & \text{elsewhere.} \end{cases}$$

Let $F(s)$ denote for $\mathcal{L}f$. Show that

$$\mathcal{L}f_0 = \mathcal{L}f - e^{-ps}\mathcal{L}f = (1 - e^{-ps})F(s).$$

7. If $g(u)$ is a continuous function on $[0, 1]$ such that

$$\int_0^1 g(u)u^n \, du = 0 \quad \text{for all } n \geq 0,$$

show that $g(u) \equiv 0$.

8. If f is continuous function on $[0, \infty)$ and admissible, and $\mathcal{L}f = 0$. Show that $f \equiv 0$.
Hint: express $s = s_0 + n + 1$ and make a change of variable $u = e^t$ in the integral of the Laplace transform of f.

5.2 Laplace transform for differential equations

5.2.1 General linear equations with constant coefficients

A linear differential equations of order n with constant coefficients has the form:

$$(a_n D^n + a_{n-1}D^{n-1} + \cdots + a_1 D + a_0)y = f(t), \tag{5.3}$$

where $D = d/dt$. We may abbreviate this equation by

$$P(D)y = f.$$

For order n equations, We need to assume $a_n \neq 0$ and need impose n conditions. The initial value problem imposes the following conditions:

$$y(0) = y_0, y'(0) = y_1, \cdots, y^{(n-1)}(0) = y_{n-1}. \tag{5.4}$$

When the source term $f(t) \equiv 0$, the equation

$$P(D)y = 0 \tag{5.5}$$

is called the homogeneous equation. The equation (5.3) is called the inhomogeneous equation.

We shall accept that this initial value problem has a unique solution which exists for all time. Such existence and uniqueness theory is the same as that for the 2×2 systems of linear equations. Therefore, we will not repeat here. Instead, we are interested in the cases where the source terms have discontinuities or impulses. Such problems appear in circuit problems where a power supply is

only provided in certain period of time, or a hammer punches the mass of a mass-spring system suddenly, or a sudden immigration of population in the population dynamics. For linear systems with constant coefficients, the Laplace transform is a useful tool to get exact solution. The method transfers the linear differential equations with constant coefficients to an algebraic equation, where the source with discontinuities is easily expressed. The solution is found through solving the algebraic equation and by the inverse Laplace transform.

5.2.2 Laplace transform applied to differential equations

Given linear differential equation with constant coefficients (5.3):

$$P(D)y = f,$$

we perform Laplace transform both sides:

$$\mathcal{L}(P(D)y) = \mathcal{L}f.$$

We claim that

$$\boxed{\mathcal{L}(P(D)y) = P(s)\cdot Y(s) - I(s) = F(s),} \tag{5.6}$$

where

$$Y(s) = (\mathcal{L}y)(s), \quad F(s) = \mathcal{L}f(s),$$
$$I(s) = \sum_{i=1}^{n}\sum_{k=i}^{n} a_k y^{(k-i)}(0)s^{i-1}.$$

In other words, the function $Y(s)$ of the Laplace transform of y satisfies an algebraic equation.

To show this, we perform

$$\mathcal{L}(D^k y) = \int_0^\infty D^k y e^{-st}\,dt = \int_0^\infty e^{-st}\,dy^{(k-1)} = -y^{(k-1)}(0) + s\mathcal{L}(D^{k-1}y).$$

Thus,

$$\begin{aligned}\mathcal{L}(D^k y) &= -y^{(k-1)}(0) + s\mathcal{L}(D^{k-1}y)\\ &= -y^{(k-1)}(0) + s\left(-y^{(k-2)}(0) + s\mathcal{L}(D^{k-2}y\right)\\ &= (-y^{(k-1)}(0) - sy^{(k-2)}(0) - \cdots - s^{k-1}y(0)) + s^k\mathcal{L}y.\end{aligned}$$

Now, $P(D) = \sum_{k=0}^{n} a_k D^k$, we have

$$\mathcal{L}(P(D)y) = \sum_{k=0}^{n} a_k\mathcal{L}(D^k y) = -\sum_{k=1}^{n} a_k \sum_{i=1}^{k} y^{(k-i)}(0)s^{i-1} + P(s)\mathcal{L}y.$$

The equation

$$P(s)\cdot Y(s) - I(s) = F(s)$$

can be solved explicitly with

$$Y(s) = \frac{F(s) + I(s)}{P(s)}.$$

Let us call

$$\boxed{G(t) = \mathcal{L}^{-1}\left(\frac{1}{P(s)}\right)} \tag{5.7}$$

called the Green's function. Then in the case of $I(s) \equiv 0$, we have

$$\boxed{y(t) = \mathcal{L}^{-1}\left(\frac{1}{P(s)} \cdot F(s)\right) = (G * f)(t).}$$

Thus, the solution is the convolution of the Green's function and the source term.

Example

1. Solve $y'' + 4y' + 4y = te^{-2t}$, $y(0) = 1$, $y'(0) = 1$.
   Taking Laplace transform, we get

$$\begin{aligned}\mathcal{L}(Dy) &= -y(0) + sY(s)\\ \mathcal{L}(D^2y) &= -y'(0) + s\mathcal{L}(Dy) = -y'(0) + s(-y(0) + sY(s))\end{aligned}$$

   Hence,

$$\mathcal{L}[(D^2 + 4D + 4)y] = (s^2 + 4s + 4)Y(s) - [y'(0) + sy(0) + 4y(0)]$$

   The Laplace transform of the source term is

$$\mathcal{L}(te^{-2t}) = \frac{1}{(s+2)^2}.$$

   Thus, we get

$$(s^2 + 4s + 4)Y(s) - [y'(0) + sy(0) + 4y(0)] = \frac{1}{(s+2)^2},$$

$$\begin{aligned}Y(s) &= \frac{1}{(s+2)^2}\left([y'(0) + sy(0) + 4y(0)] + \frac{1}{(s+2)^2}\right)\\ &= \frac{y(0)}{s+2} + \frac{y'(0) + 2y(0)}{(s+2)^2} + \frac{1}{(s+2)^4}\end{aligned}$$

   Thus, its inverse Laplace transform is

$$y(t) = y(0)e^{-2t} + (y'(0) + 2y(0))te^{-2t} + \frac{1}{3!}t^3e^{-2t}.$$

2. Solve $y'' - y = f(t)$, $y(0) = y'(0) = 0$, where

$$f(t) = \begin{cases} t, & 0 \le t < 1 \\ 0, & 1 \le t < \infty \end{cases}$$

The Laplace transform of f is

$$F(s) = \mathcal{L}(f) = \int_0^1 te^{-st}\,dt = \frac{1}{s^2}(1 - (s+1)e^{-s})$$

The Laplace transform of the equation gives

$$(s^2 - 1)Y(s) = F(s).$$

Thus,

$$\begin{aligned} Y(s) = \frac{F(s)}{s^2-1} &= (\frac{1}{s^2-1}) \cdot \left(-\frac{s+1}{s^2}e^{-s} + \frac{1}{s^2}\right) \\ &= \left(-\frac{1}{(s-1)s^2}\right)e^{-s} + \frac{1}{s^2(s^2-1)} \\ &= \left(\frac{1}{s^2} + \frac{1}{s} - \frac{1}{s-1}\right)e^{-s} + \frac{1}{2}\left(\frac{1}{s-1} - \frac{1}{s+1}\right) - \frac{1}{s^2} \end{aligned}$$

The inverse Laplace transform of each term of Y is

$$\mathcal{L}^{-1}\left(\frac{1}{s^2} + \frac{1}{s} - \frac{1}{s-1}\right) = t + 1 - e^t$$

$$\mathcal{L}^{-1}\left[\left(\frac{1}{s^2} + \frac{1}{s} - \frac{1}{s-1}\right)e^{-s}\right] = h(t-1)\left((t-1) + 1 - e^{(t-1)}\right)$$

$$\mathcal{L}^{-1}\left[\frac{1}{2}\left(\frac{1}{s-1} - \frac{1}{s+1}\right) - \frac{1}{s^2}\right] = \frac{1}{2}(e^t - e^{-t}) - t.$$

Here $h(t)$ is the Heaviside function.

Homeworks.

1. B-D,pp.322: 24,27,36,38.

2. B-D,pp. 338: 21,22

5.2.3 Generalized functions and Delta function

The delta function $\delta(t)$ is used to represent an impulse which is defined to be

$$\delta(t) = \begin{cases} \infty & \text{for } t = 0 \\ 0 & \text{otherwise.} \end{cases} \quad \text{and} \quad \int_{-\infty}^{\infty} \delta(t)\, dt = 1.$$

The δ-function can be viewed as the limit of the finite impulses

$$\delta(t) = \lim_{\epsilon \to 0+} \frac{1}{\epsilon} B_0\left(\frac{t}{\epsilon}\right)$$

where $B_0(t) = 1$ for $0 \le t < 1$ and $B_0(t) = 0$ otherwise. This limit is taken in the integral sense. Namely, for any smooth function ϕ with finite support (i.e. the nonzero domain of ϕ is bounded), the meaning of the integral:

$$\int \delta(t)\phi(t)\, dt := \lim_{\epsilon \to 0+} \int_{-\infty}^{\infty} \left(\frac{1}{\epsilon} B_0\left(\frac{t}{\epsilon}\right)\right) \phi(t)\, dt.$$

Since the latter is $\phi(0)$, we therefore define δ to be the generalized function such that

$$\int \delta(t)\phi(t)\, dt = \phi(0)$$

for any smooth function ϕ with finite support. The function ϕ here is called a test function. Likewise, *a generalized function is defined how it is used.* Namely, it is defined how it acts on smooth test functions. For instance, the Heaviside function is a generalized function in the sense that

$$\int h(t)\phi(t)\, dt := \int_0^{\infty} \phi(t)\, dt.$$

The function $f(t) := a_1\delta(t - t_1) + a_2\delta(t - t_2)$ is a generalized function. It is defined by

$$\int f(t)\phi(t)\, dt := a_1\phi(t_1) + a_2\phi(t_2).$$

All ordinary functions are generalized functions. In particular, all piecewise smooth functions are generalized functions. For such a function f, it is un-important how f is defined at the jump points. All it matters is the integral

$$\int f(t)\phi(t)\, dt$$

with test function ϕ. For piecewise smooth function f, the jump point makes no contribution to the integration.

One can differentiate a generalized function. The *generalized derivative* of a generalized function is again a generalized function in the following sense:

$$\int D_t f(t)\phi(t)\, dt := -\int f(t)\phi'(t)\, dt.$$

The right-hand side is well-defined because f is a generalized function. You can check that $D_t h(t) = \delta(t)$. If f is a piecewise smooth function having jump at $t = a$ with jump height $[f]_a$ defined by $[f]_a := \lim_{t\to a+} f(t) - \lim_{t\to a-} f(t)$. Let $f'(t)$ be the ordinary derivative of f in the classical sense. Thus, $f'(t)$ is defined everywhere except at the jump $t = a$. This $f'(t)$ is a piecewise smooth function and hence it is a generalized function. From the definition of the generalized derivative, we claim that

$$\boxed{(D_t f)(t) = f'(t) + [f]_a \delta(t-a).}$$

To see this,

$$\int (D_t f)\phi\, dt := -\int_{-\infty}^{\infty} f(t)\phi'(t)\, dt = -(\int_{-\infty}^{a} + \int_{a}^{\infty}) f(t)\phi'(t)\, dt$$

These integrals are

$$-\int_{-\infty}^{a} f(t)\phi'(t)\, dt = -f(a-)\phi(a) + \int_{-\infty}^{a} f'(t)\phi(t)\, dt$$

$$-\int_{a}^{\infty} f(t)\phi'(t)\, dt = f(a+)\phi(a) + \int_{a}^{\infty} f'(t)\phi(t)\, dt$$

Hence,

$$\begin{aligned}\int (D_t f)\phi\, dt &= (f(a+) - f(a-))\phi(a) + \int_{-\infty}^{\infty} f'(t)\phi(t)\, dt \\ &= \int ([f]_a \delta(t-a) + f'(t))\phi(t)\, dt\end{aligned}$$

You can check that $D_t\delta$ is a generalized function. It is defined by

$$\int (D_t\delta)(t)\phi(t)\, dt := -\phi'(0).$$

Let us abbreviate $D_t\delta$ by $\delta'(t)$ in later usage.

Similarly, one can take indefinite integral of a generalized function.

$$\int \left(\int_{-\infty}^{t} f(\tau)\, d\tau\right) \phi(t)\, dt := \int f(\tau) \left(\int_{\tau}^{\infty} \phi(t)\, dt\right) d\tau$$

for any test function ϕ such that $\int \phi = 0$. The Heaviside function $h(t)$ can be viewed as the integral of the delta function, namely,

$$h(t) = \int_{0}^{t} \delta(\tau)\, d\tau.$$

Laplace transform of the delta-functions It is easy to check that

1. $\mathcal{L}\delta = \int \delta(t)e^{-st}\, dt = 1$.
2. $\mathcal{L}\delta' = s$.
3. $\mathcal{L}h = 1/s$.

5.2.4 Green's function

Let us go back to the differential equation:

$$P(D)y = f.$$

with initial data $y(0), \cdots, y^{(n-1)}(0)$ prescribed. We recall that the Laplace transform of this equation gives

$$\mathcal{L}(P(D)y) = P(s) \cdot Y(s) - I(s) = F(s) \tag{5.8}$$

where $Y(s) = (\mathcal{L}y)(s)$, $F(s) = \mathcal{L}f(s)$ and

$$I(s) = \sum_{i=1}^{n} \sum_{k=i}^{n} a_k y^{(k-i)}(0) s^{i-1}.$$

The Green's function is defined to be

$$G = \mathcal{L}^{-1}\left(\frac{1}{P(s)}\right). \tag{5.9}$$

There are two situations that produce Green's function as its solutions.

- Impulse source: $I(s) \equiv 0$ and $F(s) \equiv 1$: That is,

  $$\boxed{P(D)G(t) = \delta(t),\ G(0) = G'(0) = \cdots = G^{(n-1)}(0) = 0.}$$

  Taking the Laplace transform on both sides, using

  $$\mathcal{L}\delta = 1,$$

  we have $P(s)\mathcal{L}G = 1$, or $\mathcal{L}G = 1/P(s)$, or

  $$G = \mathcal{L}^{-1}\left(\frac{1}{P(s)}\right).$$

  The Green's function corresponds to solution with impulse source and zero initial data.

- Initial impulse: $I(s) = 1$ and $F(s) \equiv 0$: That is

  $$\boxed{P(D)G(t) = 0 \text{ for } t > 0,\ G(0) = G'(0) = \cdots = 0, G^{(n-1)}(0) = \frac{1}{a_n}.}$$

Remark. Notice that the Green's functions obtained by the above two methods are identical. Indeed, let us see the following simplest example. The function e^{at} is the solution (Green's function) of both problems:

(i) $y' - ay = \delta,\ y(0) = 0,$

(ii) $y' - ay = 0,\ y(0) = 1.$

Indeed, in the first problem, the equation should be realized for $t \in \mathbb{R}$. The corresponding initial data is $y(0-) = 0$. While in the second problem, the equation should be understood to be hold for $t > 0$ and the initial data understood to be $y(0+) = 1$. This is classical sense. With this solution e^{at}, if we define

$$y(t) = \begin{cases} e^{at} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

then $D_t y + ay = \delta$. This means that this extended function is a solution of (i) and the derivative in (i) should be interpreted as weak derivative.

Examples

1. Suppose $P(D) = (D+1)(D+2)$. Then

$$\frac{1}{P(s)} = \frac{1}{s+1} - \frac{1}{s+2}$$

Hence,

$$G(t) = e^{-t} - e^{-2t}.$$

2. If $P(D) = (D+1)^2$, then

$$G(t) = \mathcal{L}^{-1}\left(\frac{1}{(s+1)^2}\right) = \mathcal{L}^{-1}\left(\left(-\frac{d}{ds}\right)\frac{1}{(s+1)}\right) = t\mathcal{L}^{-1}\left(\frac{1}{s+1}\right) = te^{-t}.$$

3. Suppose $P(D) = (D^2 + \omega^2)$. Then

$$G(t) = \mathcal{L}^{-1}\left(\frac{1}{s^2+\omega^2}\right) = \frac{\sin \omega t}{\omega}$$

In these two examples, we notice that $G(0) = 0$ but $G'(0+) = 1$. This is consistent to $G'(0-) = 0$. Indeed, G' has a jump at $t = 0$ and the generalized derivative of G' produces the delta function.

Explicit form of the Green's function

Case 1. Suppose $P(s)$ has n distinct roots $\lambda_1, ..., \lambda_n$. Then

$$\frac{1}{P(s)} = \sum_{k=1}^{n} \frac{A_k}{s - \lambda_k}, \text{ where } A_k = \frac{1}{P'(\lambda_k)}.$$

The corresponding Green's function is

$$G(t) = \sum_{k=1}^{n} A_k e^{\lambda_k t}.$$

Case 2. When $P(s)$ has multiple roots, say $P(s) = \prod_{i=1}^{\ell}(s-\lambda_i)_i^k$. Then

$$\frac{1}{P(s)} = \sum_{i=1}^{\ell}\sum_{j=1}^{k_i}\sum_{m=0}^{j-1}\frac{A_{i,j,m}s^m}{(s-\lambda_i)^j},$$

It can be shown that (see (5.2))

$$\mathcal{L}^{-1}\left(\frac{s^m}{(s-\lambda_i)^j}\right) = \frac{d^m}{dt^m}\mathcal{L}^{-1}\left(\frac{1}{(s-\lambda_i)^j}\right).$$

On the other hand,

$$\begin{aligned}\mathcal{L}^{-1}\left(\frac{1}{(s-\lambda_i)^j}\right) &= \mathcal{L}^{-1}\left(\frac{1}{j!}(-\frac{d}{ds})^j\left(\frac{1}{s-\lambda_i}\right)\right)\\ &= \frac{1}{j!}t^j\mathcal{L}^{-1}\left(\frac{1}{s-\lambda_i}\right)\\ &= \frac{1}{j!}t^j e^{\lambda_i t}.\end{aligned}$$

Thus,

$$G(t) = \sum_{i=1}^{\ell}\sum_{j=1}^{k_i}\sum_{m=0}^{j-1}A_{i,j,m}\frac{1}{j!}\frac{d^m}{dt^m}\left(t^j e^{\lambda_i t}\right)$$

Representation of solutions in terms of Green's function

1. Contribution from the source term With the Green's function, using convolution, one can express the solution of the equation $P(D)y = f$ with zero initial condition by

$$y(t) = (G * f)(t) = \int_0^t G(t-\tau)f(\tau)\,d\tau.$$

A physical interpretation of this is that the source term $f(t)$ can be viewed as

$$f(t) = \int_0^t f(\tau)\delta(t-\tau)\,d\tau$$

the superposition of delta source $\delta(t-\tau)$ with weight $f(\tau)$. This delta source produces a solution $G(t-\tau)f(\tau)$. By the linearity of the equation, we have the solution is also the superposition of these solution:

$$y(t) = \int_0^t G(t-\tau)f(\tau)\,d\tau.$$

2. Contribution from the initial data. Next, let us see the case when $f \equiv 0$ and the initial data are not zero. We have seen that the contribution of the initial state is

$$Y(s) = \frac{I(s)}{P(s)}, \quad \text{where } I(s) = \sum_{i=1}^{n} \sum_{k=i}^{n} a_k y^{(k-i)}(0) s^{i-1}.$$

We have seen that $\mathcal{L}^{-1}(s^{i-1}/P(s)) = D^{i-1}\mathcal{L}^{-1}(1/P(s)) = D^{i-1}G(t)$ (5.2). With this, we can write the general solution as the follows.

Theorem 5.1. *The solution to the initial value problem*

$$P(D)y = f$$

with prescribed $y(0), ..., y^{(n-1)}$ *has the following explicit expression:*

$$\begin{aligned} y(t) &= \mathcal{L}^{-1}\left(\frac{I(s)}{P(s)} + \frac{F(s)}{P(s)}\right) \\ &= \sum_{i=1}^{n} \sum_{k=i}^{n} a_k y^{(k-i)}(0) G^{(i-1)}(t) + (G * f)(t) \end{aligned}$$

Homeworks.

1. B-D,pp. 344: 1, 10, 14,15,16
2. Prove $\mathcal{L}(\delta^{(i)}) = s^i$.
3. Find the Green's function for the differential operator $P(D) = (D^2 + \omega^2)^m$.
4. Find the Green's function for the differential operator $P(D) = (D^2 - k^2)^m$.
5. Suppose $G = \mathcal{L}^{-1}(1/P(s))$ is the Green's function. Show that

$$\mathcal{L}^{-1}\left(\frac{s^i}{P(s)}\right) = D_t^i G(t).$$

6. B-D, pp. 352: 13, 18,19,21,22,23

Chapter 6

Calculus of Variations

6.1 A short story about Calculus of Variations

The development of calculus of variations has a long history. It may goes back to the brachistochrone problem proposed by Johann Bernoulli (1696). This is an ancient Greek problem, which is to find a path (or a curve) connecting two points A and B with B lower than A such that it takes minimal time for a ball to roll from A to B under gravity. Hohann Bernoulli used Fermat principle (light travels path with shortest distance) to prove that the curve for solving the brachistochrone problem is the cycloid.

Euler (1707-1783) and Lagrange (1736-1813) are two important persons in the development of the theory of calculus of variations. I quote two paragraphs below from Wiki for you to know some story of Euler and Lagrange.

"Lagrange was an Italian-French Mathematician and Astronomer. By the age of 18 he was teaching geometry at the Rotal Artillery School of Turin, where he organized a discussion group that became the Turin Academy of Sciences. In 1755, Lagrange sent Euler a letter in which he discussed the Calculus of Variations. Euler was deeply impressed by Lagrange's work, and he held back his own work on the subject to let Lagrange publish first."

"Although Euler and Lagrange never met, when Euler left Berlin for St. Petersburg in 1766, he recommended that Lagrange succeed him as the director of the Berlin Academy. Over the course of a long and celebrated career (he would be lionized by Marie Antoinette, and made a count by Napoleon before his death), Lagrange published a systemization of mechanics using his calculus of variations, and did significant work on the three-body problem and astronomical perturbations."

6.2 Problems from Geometry

Geodesic curves Find the shortest path connecting two points A and B on the plane. Let $y(x)$ be a curve with $(a, y(a)) = A$ and $(b, y(b)) = B$. The geodesic curve problem is to minimize

$$\int_a^b \sqrt{1 + y'(x)^2}\, dx$$

among all paths $y(\cdot)$ connecting A to B.

Isoperimetric problem This was an ancient Greek problem. It is to find a closed curve with a given length enclosing the greatest area. Suppose the curve is described by $(x(t), y(t))$, $0 \le t \le T$. We may assume the total length is L. The isoperimetric inequality problem is to

$$\max \left\{ \frac{1}{2} \int_0^T (x(t)\dot{y}(t) - y(t)\dot{x}(t))\, dt \right\},$$

subject to

$$\int_0^T \sqrt{\dot{x}(t)^2 + \dot{y}(t)^2}\, dt = L.$$

Its solution is the circle with radius $R = L/(2\pi)$. Since the circle has the maximal enclosed area among all closed curves with arc length L, we then get so-called iso-perimetric inequality

$$4\pi A \le L^2.$$

The equality holds for circles. A geometric proof was given by Steiner (1838). An analytic proof was given by Weierstrass and by Edler. [1] The proof by Hurwitz (1902) using Fourier method can also be found in John Hunter and Bruno Nachtergaele's book, Applied Analysis. In later section, we shall give an ODE proof.

6.3 Euler-Lagrange Equation

Let us consider the following variational problem:

$$\min \mathcal{J}[y] := \int_a^b F(x, y(x), y'(x))\, dx,$$

subject to the boundary conditions

$$y(a) = y_a, y(b) = y_b.$$

The function $F : \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a smooth function. We call the set

$$\mathcal{A} = \left\{ y : [a, b] \to \mathbb{R} \in C^1[a, b] | y(a) = y_a, y(b) = y_b \right\}$$

an admissible class. Here, $C^1[a, b]$ denotes the set of functions from $[a, b]$ to $\mathbb{R}$ which are continuously differentiable. An element $y \in \mathcal{A}$ is a path connecting (a, y_a) to (b, y_b). The mapping $\mathcal{J} : \mathcal{A} \to \mathbb{R}$ is called a functional. It measures the cost of a path. Given a path $y \in \mathcal{A}$, we consider a variation of this path in the direction of v by

$$y(x, \epsilon) := y(x) + \epsilon v(x).$$

[1] You can read a review article by Alan Siegel, A historical review of isoperimetric theorem in 2-D, and its place in elementary plan geometry . For applications, you may find a book chapter from Fan in .

Here, v is a C^1 function with $v(a) = v(b) = 0$ in order to have $y(\cdot, \epsilon) \in \mathcal{A}$ for small ϵ. Such v is called a *variation*. Sometimes, it is denoted by δy. We can plug $y(\cdot, \epsilon)$ into $\mathcal{J}$. Suppose y is a local minimum of $\mathcal{J}$ in $\mathcal{A}$, then for any such variation v, $\mathcal{J}[y + \epsilon v]$ takes minimum at $\epsilon = 0$. This leads to a necessary condition:

$$\frac{d}{d\epsilon}\Big|_{\epsilon=0} \mathcal{J}[y + \epsilon v] = 0.$$

Let us compute this derivative

$$\begin{aligned}
\frac{d}{d\epsilon}\Big|_{\epsilon=0} \mathcal{J}[y + \epsilon v] &= \frac{d}{d\epsilon}\Big|_{\epsilon=0} \int_a^b F(x, y(x) + \epsilon v(x), y'(x) + \epsilon v'(x))\, dx \\
&= \int_a^b \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=0} F(x, y(x) + \epsilon v(x), y'(x) + \epsilon v'(x))\, dx \\
&= \int_a^b F_y(x, y(x), y'(x))v(x) + F_{y'}(x, y(x), y'(x))v'(x)\, dx
\end{aligned}$$

It is understood that $F_{y'}$ here means the partial derivative w.r.t. the third variable y'. For instance, suppose $F(y, y') = \frac{y^2}{2} + \frac{y'^2}{2}$, then $F_{y'} = y'$.

Theorem 6.1 (Necessary Condition)**.** *A necessary condition for $y \in \mathcal{A}$ to be a local minimum of $\mathcal{J}$ is*

$$\boxed{\int_a^b F_y(x, y(x), y'(x))v(x) + F_{y'}(x, y(x), y'(x))v'(x)\, dx = 0} \tag{6.1}$$

for all $v \in C^1[a, b]$ with $v(a) = v(b) = 0$.

If the solution $y \in C^2[a, b]$, then we can take integration by part on the second term to get

$$\int_a^b F_{y'}(x, y(x), y'(x))v'(x)\, dx = -\int_a^b \frac{d}{dx} F_{y'}(x, y(x), y'(x))v(x)\, dx.$$

Here, I have used $v(a) = v(b) = 0$. Thus, the necessary condition can be rewritten as

$$\int_a^b \left(F_y(x, y(x), y'(x)) - \frac{d}{dx} F_{y'}(x, y(x), y'(x)) \right) v(x)\, dx = 0$$

for all $v \in C^1[a, b]$ with $v(a) = v(b) = 0$. A fundamental theorem of calculus of variations is the following theorem.

Theorem 6.2. *If $f \in C[a, b]$ satisfies*

$$\int_a^b f(x)v(x)\, dx = 0$$

for all $v \in C^\infty[a, b]$ with $v(a) = v(b) = 0$, then $f \equiv 0$.

Proof. If $f(x_0) \neq 0$ for some $x_0 \in (a, b)$ (say $f(x_0) = C > 0$), then there is small neighborhood $(x_0 - \epsilon, x_0 + \epsilon)$ such that $f(x) > C/2$. We can choose v to be a hump such that $v(x) = 1$ for $|x - x_0| \leq \epsilon/2$ and $v(x) \geq 0$ and $v(x) = 0$ for $|x - x_0| \geq \epsilon$. The test function still satisfies the boundary constraint if ϵ is small enough. Using this v, we get

$$\int_a^b f(x)v(x)\,dx \geq \frac{C\epsilon}{2} > 0.$$

This contradicts to our assumption. We conclude $f(x_0) = 0$ for all $x_0 \in (a, b)$. Since f is continuous on $[a, b]$, we also have $f(a) = f(b) = 0$ by continuity of f. □

Thus, we obtain the following stronger necessary condition.

Theorem 6.3. *A necessary condition for a local minimum y of $\mathcal{J}$ in $\mathcal{A} \cap C^2$ is*

$$\boxed{\frac{\delta \mathcal{J}}{\delta y} := F_y(x, y(x), y'(x)) - \frac{d}{dx}F_{y'}(x, y(x), y'(x)) = 0.} \tag{6.2}$$

Equation 6.2 is called the *Euler-Lagrange equation* for the minimization problem $\min \mathcal{J}[y]$.

Example For the problem of minimizing arc length, the functional is

$$\mathcal{J}[y] = \int_a^b \sqrt{1 + y'^2}\,dx,$$

where $y(a) = y_a, y(b) = y_b$. The corresponding Euler-Lagrange equation is

$$\frac{d}{dx}F_{y'} = \frac{d}{dx}\left(\frac{y'}{\sqrt{1 + y'^2}}\right) = 0.$$

This yields

$$\frac{y'}{\sqrt{1 + y'^2}} = Const.$$

Solving y', we further get

$$y' = C \text{ (a constant)}.$$

Hence $y = Cx + D$. Applying boundary condition, we get

$$C = \frac{y_b - y_a}{b - a}, \quad D = \frac{by_a - ay_b}{b - a}.$$

Thus, the curves with minimal arc length on the plane are straight lines.

Homework

1. Compute $\delta\mathcal{J}/\delta y$ of the following functionals. We will neglect boundary effects if there is any.

   (a) $\mathcal{J}[y] = \int_a^b V(x)y(x)\,dx$.

   (b) $\mathcal{J}[y] = \int_a^b \alpha(x)y'(x)\,dx$.

   (c) $\mathcal{J}[y] = \int_a^b (\alpha(x)y'(x))^2\,dx$.

   (d) $\mathcal{J}[y] = \int_a^b \left(-\frac{y(x)^2}{2} + \frac{y(x)^4}{4}\right)\,dx$.

   (e) $\mathcal{J}[y] = \frac{1}{p}\int_a^b (y'(x))^p\,dx$, $1 < p < \infty$.

   (f) $\mathcal{J}[y] = \int_a^b (y''(x))^2\,dx$.

6.4 Problems from Mechanics

Least action principle In classical mechanics, the motion of a particle in $\mathbb{R}^3$ is described by

$$m\ddot{x} = -\nabla V(x) = F(x),$$

where, $V(x)$ is called a potential and F is called a (conservative) force. This is called Newton's mechanics. Typical examples of potentials are the potential $V(x) = gx$ with uniform force field, the harmonic potential $V(x) = \frac{k^2}{2}|x|^2$ for a mass-spring system, the Newtonian potential $V(x) = -\frac{G}{|x|}$ for solar-planet system, etc. Here, k is the spring constant, G, the gravitation constant.

The Newton mechanics was reformulated by Lagrange (1788) in variational form and was originally motivated by describing particle motion under constraints. Let us explain this variational formulation without constraint. First, let us introduce the concept of *virtual velocity* or *variation of position*. Given a path $x(t)$, $t_0 \le t \le t_1$, consider a family of paths

$$x_\epsilon(t) := x(t,\epsilon) := x(t) + \epsilon v(t), t_0 \le t \le t_1, -\epsilon_0 < \epsilon < \epsilon_0.$$

Here, $v(t)$ is called a virtual velocity and $x_\epsilon(\cdot)$ is called a small variation of the path $x(\cdot)$. Sometimes, we denote $v(\cdot)$, the variation of $x_\epsilon(\cdot)$, by δx. That is, $\delta x := \partial_\epsilon|_{\epsilon=0} x_\epsilon$.

Now, Newton's law of motion can be viewed as

$$\delta W = (F - m\ddot{x}) \cdot v = 0 \text{ for any virtual velocity } v.$$

The term δW is called the *total virtual work* in the direction v. The term $F \cdot v$ is the virtual work done by the external force F, while $m\ddot{x} \cdot v$ is the work done by the inertia force. The *d'Alembert principle of virtual work* states that *the virtual work is always zero along physical particle path under small perturbation* v.

If we integrate it in time from t_0 to t_1 with fixed $v(t_0) = v(t_1) = 0$, then we get

$$\begin{aligned} 0 &= \int_{t_0}^{t_1} -m\ddot{x} \cdot v - \nabla V(x) \cdot v \, d\tau \\ &= \int_{t_0}^{t_1} m\dot{x} \cdot \dot{v} - \nabla V(x) \cdot v \, d\tau \\ &= \int_{t_0}^{t_1} \partial_\epsilon|_{\epsilon=0} \left(\frac{1}{2} m|\dot{x}_\epsilon|^2 - V(x_\epsilon) \right) d\tau \\ &= \frac{d}{d\epsilon}\Big|_{\epsilon=0} \int_{t_0}^{t_1} L(x_\epsilon, \dot{x}_\epsilon) \, d\tau = \delta\mathcal{S}. \end{aligned}$$

Here,

$$L(x, \dot{x}) := \frac{1}{2} m|\dot{x}|^2 - V(x),$$

is called the *Lagrangian*, and the integral

$$\mathcal{S}[x] := \int_{t_0}^{t_1} L(x(\tau), \dot{x}(\tau)) \, d\tau$$

is called the *action*. Thus, $\delta\mathcal{S} = 0$ *along a physical path*. This is called the *Hamilton principle or the least action principle*. You can show that the corresponding Euler-Language equation is exactly the Newton's law of motion.

Theorem 6.4. *The following formulations are equivalent:*

- *Newton's equation of motion* $m\ddot{x} = -V'(x)$*;*
- *d'Alembert principle of virtual work:* $\int_{t_0}^{t_1} (m\dot{x} \cdot \dot{v} - V'(x)v) \, dt = 0$ *for all virtual velocity* v*;*
- *Hamilton's least action principle:* $\delta \int_{t_0}^{t_1} \left(\frac{m}{2}|\dot{x}|^2 - V(x) \right) dt = 0$.

Remarks

1. The meaning of the notation δ. In the path space, we vary $x(\cdot)$ by $x_\epsilon(\cdot)$. This means that they are a family of paths. We can express them as $x(t, \epsilon)$. A typical example is $x(t, \epsilon) = x(t) + \epsilon v(t)$. The variation of the path x_ϵ simply means

$$\delta x(t) = \frac{\partial}{\partial \epsilon}|_{\epsilon=0} x(t, \epsilon).$$

For the case $x_\epsilon = x + \epsilon v$, $\delta x = v$. Sometimes, we use prime to denote for $\frac{\partial}{\partial \epsilon}$, while dot denote for $\frac{\partial}{\partial t}$. The two differentiations commute. That is

$$\delta\dot{x} = \dot{x}' = \frac{d}{dt}\delta x.$$

2. When we consider a variation of path x_ϵ, the functional $\mathcal{S}[x_\epsilon]$ becomes a function of ϵ as well:

$$S(\epsilon) := \mathcal{S}[x_\epsilon] = \int_{t_0}^{t_1} L(x(\tau,\epsilon), \dot{x}(\tau,\epsilon))\, d\tau.$$

We can take differentiation of S w.r.t. ϵ at $\epsilon = 0$:

$$\begin{aligned}
\frac{dS}{d\epsilon}(0) &= \frac{d}{d\epsilon}|_{\epsilon=0} \int_{t_0}^{t_1} L(x(\tau,\epsilon), \dot{x}(\tau,\epsilon))\, d\tau \\
&= \int_{t_0}^{t_1} \left(\frac{\partial}{\partial \epsilon} L(x(\tau,\epsilon), \dot{x}(\tau,\epsilon)) \right) d\tau \\
&= \int_{t_0}^{t_1} \left(L_x x' + L_{\dot{x}} \dot{x}' \right) d\tau \\
&= \int_{t_0}^{t_1} \left(L_x \delta x - \frac{d}{d\tau} L_{\dot{x}} \delta x \right) d\tau \\
&= \int_{t_0}^{t_1} \frac{\delta \mathcal{S}}{\delta x}(\tau) \delta x(\tau)\, d\tau.
\end{aligned}$$

Thus, the notation $\frac{\delta \mathcal{S}}{\delta x}$ is

$$\frac{\delta \mathcal{S}}{\delta x}(t) = L_x(x(t), \dot{x}(t)) - \frac{d}{dt} L_{\dot{x}}(x(t), \dot{x}(t)).$$

is the variation of $\mathcal{S}$ w.r.t. the path x. Sometimes, we write

$$\delta \mathcal{S} = \frac{\delta \mathcal{S}}{\delta x} \cdot \delta x.$$

One advantage of variational formulation – existence of first integral One advantage of this variational formulation is that it is easy to find some invariants (or so-called integrals) of the system. One exmple is the existence of the first integral.

Theorem 6.5. *When the Lagrangian $L(x, \dot{x})$ is independent of t, then the quantity (called the first integral)*

$$I(x, \dot{x}) := \dot{x} \cdot \frac{\partial L}{\partial \dot{x}} - L(x, \dot{x})$$

is independent of t along physical trajectories.

Proof. We differentiate $I(x(\cdot), \dot{x}(\cdot))$ along a physical trajectory $x(\cdot)$:

$$\begin{aligned}
\frac{d}{dt} \left[\dot{x} L_{\dot{x}} - L \right] &= \ddot{x} L_{\dot{x}} + \dot{x} \frac{d}{dt} L_{\dot{x}} - L_x \dot{x} - L_{\dot{x}} \ddot{x} \\
&= \dot{x} \left(\frac{d}{dt} L_{\dot{x}} - L_x \right) = 0.
\end{aligned}$$

□

Remarks.

1. For the Newton mechanics where $L(x, \dot{x}) = \frac{1}{2}m|\dot{x}|^2 - V(x)$, this first integral is indeed the total energy. Indeed, we obtain

$$I(x, \dot{x}) = \frac{1}{2}m|\dot{x}|^2 + V(x).$$

2. In Newton's equation:

$$m\ddot{x} = -\nabla V(x),$$

we multiply both sides by $\dot{x}$ and obtain

$$m\ddot{x}\dot{x} + \nabla V(x)\dot{x} = 0.$$

This can be written as

$$\frac{d}{dt}\left(\frac{1}{2}m|\dot{x}|^2 + V(x)\right) = 0.$$

Thus,

$$\frac{1}{2}m|\dot{x}|^2 + V(x) = E.$$

for some constant E. This is another equivalent derivation, called energy method for Newton's mechanics with conservative force field.

3. If the particle motion is in one dimension, that is, $x(\cdot) \in \mathbb{R}$, then the first integral

$$\frac{m}{2}\dot{x}^2 + V(x) = E$$

determines trajectories on the phase plane. Let us see the following example.

   (a) Harmonic oscillator: $V(x) = \frac{k}{2}x^2$. The conservation of energy gives

$$\frac{m}{2}\dot{x}^2 + \frac{k}{2}x^2 = E.$$

   Each fixed E determines an ellipse on the phase plane $(x, \dot{x})$. Given an initial state $(x(0), \dot{x}(0))$, it also determines a unique $E_0 = \frac{m}{2}\dot{x}(0)^2 + \frac{k}{2}x(0)^2$. This E_0 determines a trajectory from $\frac{m}{2}\dot{x}^2 + \frac{k}{2}x^2 = E$, which is exactly the trajectory with the initial state $(x(0), \dot{x}(0))$.

   (b) Simple pendulum: A simple pendulum has a mass m hanging on a massless rod with length ℓ. The rod is fixed at one end and the mass m swings at the other end by the gravitational force, which is mg. Let θ be the angle of the rod and the negative vertical direction $(0, -1)$. The locus the mass travels is on the circle centered at the fixed end of the rod. Thus, we have

   - mass position: $\ell(\sin\theta, -\cos\theta)$,
   - tangential direction of the motion: $(\cos\theta, \sin\theta)$

- tangential velocity: $v = \ell\dot{\theta}$,
- tangential acceleration: $a = \ell\ddot{\theta}$,
- the gravitation force: $\mathbf{F} = mg(0, -1)$,
- the force in the tangential direction: $-mg\sin\theta$.

The Newton's law of motion gives

$$m\ell\ddot{\theta} = -mg\sin\theta.$$

We eliminate m and get

$$\ddot{\theta} = -\frac{g}{\ell}\sin\theta.$$

The conservation of energy reads

$$\frac{1}{2}\dot{\theta}^2 - \frac{g}{\ell}\cos\theta = E.$$

Each E determines a trajectory on the phase plane $(\theta, \dot{\theta})$. Here are some special trajectories.

- The stable equilibria: $\theta = 2n\pi$, $\dot{\theta} = 0$. The corresponding $E_0 = -\frac{g}{\ell}$.
- The unstable equilibria: $\theta = (2n+1)\pi$, $\dot{\theta} = 0$. The corresponding energy is $E_1 = \frac{g}{\ell}$.
- The heteroclinic orbit: it connects two neighboring unstable equilibria: it satisfies
$$\frac{1}{2}\dot{\theta}^2 - \frac{g}{\ell}\cos\theta = E_1,$$
but it is not an equilibrium state.
- For $E_0 < E < E_1$, the corresponding orbit is a closed curve. For $E > E_1$, the corresponding is an unbounded orbit.

6.5 Method of Lagrange Multiplier

In variational problems, there are usually accompanied with some constraints. As we have seen that the iso-perimetric problem. Lagrange introduced auxiliary variable, called the Lagrange multiplier, to solve these kinds of problems. Below, we use the hanging rope problem to explain the method of Lagrange multiplier.

Hanging rope problem A rope given by $y(x)$, $a \le x \le b$ hangs two end points (a, y_a) and (b, y_b). Suppose the rope has length ℓ and density $\rho(x)$. Suppose the rope is in equilibrium, then it minimizes its potential energy, which is

$$\mathcal{J}[y] = \int_0^\ell \rho g y\, ds = \int_a^b \rho g y\sqrt{1 + y'^2}\, dx.$$

The rope is subject to the length constraint

$$\mathcal{W}[y] = \int_a^b \sqrt{1 + y'^2}\, dx = \ell.$$

Method of Lagrange multiplier In dealing with such problems, it is very much like the optimization problems in finite dimensions with constraints. Let us start with two dimensional examples. Suppose we want to minimize $f(x, y)$ with constraint $g(x, y) = 0$. The method of Lagrange multiplier states that a necessary condition for (x_0, y_0) being such a solution is that, if $\nabla g(x_0, y_0) \neq 0$, then $\nabla f(x_0, y_0) \parallel \nabla g(x_0, y_0)$. This means that there exists a constant λ_0 such that $\nabla f(x_0, y_0)+\lambda_0 \nabla g(x_0, y_0) = 0$. In other words, (x_0, y_0, λ_0) is an extremum of the unconstraint function $F(x, y, \lambda) := f(x, y) + \lambda g(x, y)$. That is, (x_0, y_0, λ_0) solves

$$\frac{\partial F}{\partial x} = 0, \ \frac{\partial F}{\partial y} = 0, \ \frac{\partial F}{\partial \lambda} = 0.$$

The first two is equivalent to $\nabla f(x_0, y_0) \parallel \nabla g(x_0, y_0)$. The last one is equivalent to the constraint $g(x_0, y_0) = 0$. The advantage is that the new formulation is an *unconstrained minimization problem.*

For constrained minimization problem in n dimensions, we have same result. Let $\mathbf{y} = (y^1, ..., y^n)$. $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$. Consider

$$\min f(\mathbf{y}) \quad \text{subject to} \quad g(\mathbf{y}) = 0.$$

A necessary condition for $\mathbf{y}_0$ being such a solution is that, if $\nabla g(\mathbf{y}_0) \neq 0$, then there exists λ_0 such that $(\mathbf{y}_0, \lambda_0)$ is an extremum of the unconstraint function $F(\mathbf{y}, \lambda) := f(\mathbf{y}) + \lambda g(\mathbf{y})$. That is, $(\mathbf{y}_0, \lambda_0)$ solves

$$\frac{\partial F}{\partial \mathbf{y}}(\mathbf{y}_0, \lambda_0) = 0, \quad \frac{\partial F}{\partial \lambda}(\mathbf{y}_0, \lambda_0) = 0.$$

For variational problem, we have much the same. Let us consider a variational problem in an abstract form:

$$\min \mathcal{J}[y] \quad \text{subject to} \quad \mathcal{W}[y] = 0$$

in some admissible class $\mathcal{A} = \{y : [a, b] \to \mathbb{R} | y(a) = y_a, y(b) = y_b\}$ in some function space. We approximate this variational problem to a finite dimensional problem. For any large n, we partition $[a, b]$ into n even subintervals:

$$x_i = a + i\frac{b-a}{n}, i = 0, ..., n.$$

We approximate $y(\cdot) \in \mathcal{A}$ by piecewise linear continuous function $\tilde{y}$ with

$$\tilde{y}(x_i) = y(x_i), i = 0, ..., n.$$

The function $\tilde{y} \in \mathcal{A}$ has an one-to-one correspondence to $\mathbf{y} := (y^1, ..., y^{n-1}) \in \mathbb{R}^{n-1}$. We approximate $\mathcal{J}[y]$ by $J(\mathbf{y}) := \mathcal{J}[\tilde{y}]$, and $\mathcal{W}[y]$ by $W(\mathbf{y}) = \mathcal{W}[\tilde{y}]$. Then the original constrained variational problem is approximated by a constrained optimization problem in finite dimension. Suppose $\mathbf{y}_0$ is such a solution. According to the method of Lagrange multiplier, if $\nabla W(\mathbf{y}_0) \neq 0$, then there exists a λ_0 such that $(\mathbf{y}_0, \lambda_0)$ solves the variational problem: $J(\mathbf{y}) + \lambda W(\mathbf{y})$.

Notice that the infinite dimensional gradient $\delta\mathcal{W}/\delta y$ can be approximated by the finite dimensional gradient $\nabla W(\mathbf{y})$. That is

$$\frac{\delta \mathcal{W}}{\delta y}[y] \approx \frac{\delta \mathcal{W}}{\delta y}[\tilde{y}] = \frac{\partial W}{\partial \mathbf{y}} = \nabla W(\mathbf{y}).$$

We summarize the above intuitive argument as the following theorem.

Theorem 6.6. *If y_0 is an extremum of $\mathcal{J}[\cdot]$ subject to the constraint $\mathcal{W}[y] = 0$, and if $\delta\mathcal{W}/\delta y \neq 0$, then there exists a constant λ_0 such that (y_0, λ_0) is an extremum of the functional $\mathcal{J}[y] + \lambda\mathcal{W}[y]$ with respect to (y, λ).*

***Remark.** A more serious proof is the follows.

1. We consider two-parameter variations

$$z(x) = y(x) + \epsilon_1 h_1(x) + \epsilon_2 h_2(x).$$

The variation h_i should satisfy the boundary conditions: $h_i(a) = h_i(b) = 0$ in order to have z satisfy the boundary conditions: $z(a) = y_a$ and $z(b) = y_b$. For arbitrarily chosen such variations h_i, we should also require ϵ_i satisfying

$$W(\epsilon_1, \epsilon_2) = \mathcal{W}[y + \epsilon_1 h_1 + \epsilon_2 h_2] = 0.$$

On the variational subspaces spanned by h_i, $i = 1, 2$, the functional $\mathcal{J}$ becomes

$$J(\epsilon_1, \epsilon_2) := \mathcal{J}[y + \epsilon_1 h_1 + \epsilon_2 h_2].$$

Thus the original problem is reduced to

$$\min J(\epsilon_1, \epsilon_2) \quad \text{subject to} \quad W(\epsilon_1, \epsilon_2) = 0$$

on this variational subspace. By the method of Lagrange multiplier, there exists a λ such that an extremum of the original problem solves the unconstraint optimization problem $\min J + \lambda W$. This leads to three equations

$$\begin{aligned} 0 &= \frac{\partial}{\partial \epsilon_1}(J + \lambda W) = \left(\frac{\delta \mathcal{J}}{\delta y} + \lambda \frac{\delta \mathcal{W}}{\delta y}\right) \cdot h_1 \\ 0 &= \frac{\partial}{\partial \epsilon_2}(J + \lambda W) = \left(\frac{\delta \mathcal{J}}{\delta y} + \lambda \frac{\delta \mathcal{W}}{\delta y}\right) \cdot h_2 \\ 0 &= \frac{\partial}{\partial \lambda}(J + \lambda W) = \mathcal{W}[y] \end{aligned}$$

2. Notice that the Lagrange multiplier λ so chosen, depends on h_1 and h_2. We want o show that it is indeed a constant. This is proved below.

3. Since $\delta\mathcal{W}/\delta y(x) \neq 0$, we choose x_1 where $\delta\mathcal{W}/\delta y(x_1) \neq 0$. For any $x_2 \in (a, b)$, we consider $h_i = \delta(x - x_i)$, $i = 1, 2$. Here, δ is the Dirac delta function. It has the property: for any continuous function f,

$$\int f(x)\delta(x - x_0)\, dx = f(x_0).$$

By choosing such h_i, we obtain that there exists a λ_{12} such that

$$\frac{\delta \mathcal{J}}{\delta y}(x_1) + \lambda_{12}\frac{\delta \mathcal{W}}{\delta y}(x_1) = 0$$

$$\frac{\delta \mathcal{J}}{\delta y}(x_2) + \lambda_{12}\frac{\delta \mathcal{W}}{\delta y}(x_2) = 0$$

In other words, the constant

$$\lambda_{12} = -\frac{\frac{\delta \mathcal{J}}{\delta y}(x_1)}{\frac{\delta \mathcal{W}}{\delta y}(x_1)}.$$

For any arbitrarily chosen x_2, we get the same constant. Thus, λ_{12} is independent of x_2. In fact, the above formula shows

$$\frac{\frac{\delta \mathcal{J}}{\delta y}(x_1)}{\frac{\delta \mathcal{W}}{\delta y}(x_1)} = \frac{\frac{\delta \mathcal{J}}{\delta y}(x_2)}{\frac{\delta \mathcal{W}}{\delta y}(x_2)},$$

for any $x_2 \neq x_1$. This means that there exists a constant λ such that

$$\frac{\delta \mathcal{J}}{\delta y}(x) + \lambda\frac{\delta \mathcal{W}}{\delta y}(x) = 0 \text{ for all } x \in (a, b).$$

6.6 Examples

6.6.1 The hanging rope problem

Let us go back to investigate the hanging rope problem. By the method of Lagrange multiplier, we consider the extremum problem of new Lagrangian

$$L(y, y', \lambda) = \rho g y\sqrt{1 + y'^2} + \lambda\sqrt{1 + y'^2}.$$

The Lagrangian is independent of x, thus it admits the first integral $L - y'L_{y'} = C$, or

$$(\rho g y + \lambda)\left(\sqrt{1 + y'^2} - \frac{y'^2}{\sqrt{1 + y'^2}}\right) = C.$$

Solving for y' gives

$$y' = \pm\frac{1}{C}\sqrt{(\rho g y + \lambda)^2 - C^2}.$$

Using method of separation of variable, we get

$$\frac{dy}{\sqrt{(\rho g y + \lambda)^2 - C^2}} = \pm\frac{dx}{C}.$$

Change variable $u = \rho g y + \lambda$, we get

$$\frac{1}{\rho g}\cosh^{-1}\left(\frac{u}{C}\right) = \pm\frac{x}{C} + C_1.$$

Hence

$$y = -\frac{\lambda}{\rho g} + \frac{C}{\rho g}\cosh\left(\frac{\rho g x}{C} + C_2\right).$$

The constraints C, C_2 and the Lagrange multiplier λ are then determined by the two boundary conditions and the constraint. The shape of this hanging rope is called a *catenary*.

6.6.2 Isoperimetric inequality

We recall that the isoperimetric inequality is to find a closed curve with a given length enclosing the greatest area. Suppose the curve is described by $(x(t), y(t))$, where t is a parameter on the curve, $0 \le t \le T$. The iso-perimetric problem is to maximize the area

$$\mathcal{A}[x, y] := \frac{1}{2}\int_0^T \left(x(s)\dot{y}(t) - y(t)\dot{x}(t)\right)\, dt$$

subject to

$$\mathcal{L}[x, y] := \int_0^T \sqrt{\dot{x}(t)^2 + \dot{y}(t)^2}\, dt = 2\pi.$$

This is a constrained maximization problem. We use method of Lagrange multiplier, there exists a constant λ such that the solution satisfies

$$\begin{aligned}
\delta(\mathcal{A} - \lambda\mathcal{L}) &= \frac{1}{2}\int_0^T \dot{y}\delta x - y\delta\dot{x} + x\delta\dot{y} - \dot{x}\delta y\, dt - \lambda\int_0^T \frac{\dot{x}\delta\dot{x} + \dot{y}\delta\dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2}}\, dt \\
&= \int_0^T \dot{y}\delta x - \dot{x}\delta y\, dt + \lambda\int_0^T \frac{d}{dt}\left(\frac{\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2}}\right)\delta x + \frac{d}{dt}\left(\frac{\dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2}}\right)\delta y\, dt \\
&= \int_0^T \left(\dot{y} + \lambda\frac{d}{dt}\left(\frac{\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2}}\right)\right)\delta x + \left(-\dot{x} + \lambda\frac{d}{dt}\left(\frac{\dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2}}\right)\right)\delta y\, dt = 0.
\end{aligned}$$

This is valid for any δx and δy. Thus,

$$\begin{aligned}
\frac{d}{dt}\left(\frac{\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2}}\right) &= -\frac{1}{\lambda}\dot{y} \\
\frac{d}{dt}\left(\frac{\dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2}}\right) &= \frac{1}{\lambda}\dot{x}.
\end{aligned}$$

We claim that this means that the curve $(x(\cdot), y(\cdot))$ has *constant curvature*, and such curves must be circles.

To see this, let us review some plane curve theory. On the curve $(x(t), y(t))$, we may parametrize it by the arc length

$$s = \int_0^t \sqrt{\dot{x}(\tau)^2 + \dot{y}(\tau)^2}\, d\tau.$$

Since we assume the total arc length is L, we have $0 \le s \le L$. We have $ds = \sqrt{\dot{x}(t)^2 + \dot{y}(t)^2}dt$. Let us denote the differentiation in s by prime. The tangent and normal of the curve are

$$\mathbf{t} := (x', y') = \left(\frac{\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2}}, \frac{\dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2}} \right),$$
$$\mathbf{n} := (-y', x') = \left(\frac{-\dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2}}, \frac{\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2}} \right).$$

It is clearly that $\mathbf{t} \perp \mathbf{n}$, $\mathbf{t} \cdot \mathbf{t} = 1$, and $\mathbf{n} \cdot \mathbf{n} = 1$. Differentiate $\mathbf{t} \cdot \mathbf{t} = 1$ in s, we get $\frac{d\mathbf{t}}{ds} \perp \mathbf{t} = 0$. Since $\mathbf{t} \perp \mathbf{n}$, we have $\frac{d\mathbf{t}}{ds} \parallel \mathbf{n}$. The curvature of a curve K is defined by

$$\frac{d\mathbf{t}}{ds} = K\mathbf{n}.$$

This equation, as expressed in terms of the parameter t, reads

$$\frac{1}{\sqrt{\dot{x}^2 + \dot{y}^2}} \frac{d}{dt} \left(\frac{\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2}} \right) = K \frac{-\dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2}},$$
$$\frac{1}{\sqrt{\dot{x}^2 + \dot{y}^2}} \frac{d}{dt} \left(\frac{\dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2}} \right) = K \frac{\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2}}.$$

Comparing this equation and the Euler-Lagrange equation corresponding iso-perimetric inequality problem, we conclude that $K = 1/\lambda$ is a constant. The quantity $\lambda = 1/K$ is called the radius of curvature.

Let us denote (x', y') by (ξ, η). The above equation is

$$\xi' = -K\eta$$
$$\eta' = K\xi.$$

This gives $\xi = -\sin(Ks)$, $\eta = \cos(Ks)$. Here, I have normalized $(\xi, \eta) = (0, 1)$ at $s = 0$. Notice that (ξ, η) is a unit vector. From $(x', y') = (-\sin(Ks), \cos(Ks))$, we get

$$x(s) = x_0 + \frac{1}{K}\cos(Ks)$$
$$y(s) = y_0 + \frac{1}{K}\sin(Ks).$$

Since the total length of this curve is L, we get

$$L = \frac{2\pi}{K}.$$

The area enclosed by the circle is $A^* = \pi \frac{1}{K^2}$, which has the maximal area among all closed curves with arc length L. Thus, for any closed curve with arc length L, the enclosed area satisfies

$$A \le A^* = \frac{1}{4\pi} L^2.$$

This is the iso-perimetric inequality.

6.6.3 The Brachistochrone

The Brachistochrone problem is to find a curve on which a ball sliding down under gravitation to a point $A(0,0)$ to another point $B(x_b, y_b)$ takes least time. The word "brachistochrone" means the "the shortest time delay" in Greek. It was one of the oldest problem in Calculus of Variation. Its solution is a section of a cycloid. This was founded by Leibnitz, L'Hospital, Newton and two Bernoullis.

Suppose the curve is given by $(x(\cdot), y(\cdot))$ starts from $A = (0,0)$. Let s be the arc length of the curve. We can parametrize this curve by this arc length s, i.e. $(x(s), y(s))$. The gravitation force is $-mg(0,1)$, where m is the mass of the ball and g is the gravitation constant. We project the force to the tangential direction of the curve, which is $(x'(s), y'(s))$, and get the tangential force is $-mgy'(s)$. Thus, the equation of motion (in the tangential direction) is

$$m\ddot{s} = -mgy'(s).$$

Here, dot means d/dt, whereas prime means d/ds. We multiply both sides by $\dot{s}$, we then find

$$m\ddot{s}\dot{s} + mgy'(s)\dot{s} = 0.$$

which is

$$\frac{d}{dt}\left(\frac{1}{2}m\dot{s}^2 + mgy\right) = 0.$$

This gives the conservation of energy

$$\frac{1}{2}m\dot{s}^2 + mgy(s) = E.$$

At point $A(0,0)$, we take $s = 0$, $\dot{s} = 0$ and $y(0) = 0$. With this normalization, $E = 0$. Thus, the conservation of energy gives the speed

$$v = \dot{s} = \sqrt{-2gy}.$$

Notice that $y \leq 0$ under our consideration. It is more convenient to work on positive y. Thus, we change y to $-y$ and the y stays positive. The traveling time from A to B is given by

$$T_A^B = \int_0^s \frac{1}{v}\, ds = \int_0^s \frac{1}{\sqrt{2gy}}\, ds,$$

where the distance s is not known yet. To find this curve$(x(s), y(s))$, we now parameterize it by x. That is, we look for $y(x)$, $x \in (0, x_b)$. The term $ds = \sqrt{dx^2 + dy^2} = \sqrt{1 + y'(x)^2}dx$. From now on, the prime means d/dx. Now,

$$T_A^B = \int_0^s \frac{ds}{\sqrt{2gy}} = \int_0^{x_b} \frac{\sqrt{1 + y'(x)^2}}{\sqrt{2gy}}\, dx.$$

We may move the constant $\sqrt{2g}$ to the left-hand side:

$$\sqrt{2g}T_A^B = \int_0^{x_b} F(y, y')\, dx := \int_0^{x_b} \sqrt{\frac{1 + y'^2}{y}}\, dx.$$

The corresponding Euler-Lagrange equation is

$$\frac{d}{dx}F_{y'} - F_y = 0.$$

Since the Lagrangian $F(y, y')$ is independent of x, the first integral exists. We derive it again below. We multiply this equation by y', we arrive

$$\begin{aligned} 0 &= (F_y - \frac{d}{dx}F_{y'})y' \\ &= \left(\frac{d}{dx}F - F_{y'}y''\right) - \left(\frac{d}{dx}(F_{y'}y') - F_{y'}y''\right) \\ &= \frac{d}{dx}\left(F - F_{y'}y'\right). \end{aligned}$$

The quantity $F - F_{y'}y'$, the first integral, is a constant. That is,

$$\sqrt{\frac{1+y'^2}{y}} - \frac{y'^2}{\sqrt{y(1+y'^2)}} = C.$$

This leads to

$$y(1+y'^2) = \frac{1}{C^2} = A.$$

After rearrangement, we get

$$\frac{dy}{dx} = \pm\sqrt{\frac{A-y}{y}}.$$

There are positive and negative branches. We can choose positive branch, because the other branch can be obtained by replacing x by $-x$. Using separation of variable, we get

$$x = \int \sqrt{\frac{y}{A-y}}\, dy.$$

Taking the substitution

$$y = A(1-\cos\theta) = 2A\sin^2(\frac{\theta}{2}),$$

we get

$$\begin{aligned} x &= \int \sqrt{\frac{y}{A-y}}\, dy = \int \sqrt{\frac{\sin^2\frac{\theta}{2}}{1-\sin^2\frac{\theta}{2}}} 2A\sin\frac{\theta}{2}\cos\frac{\theta}{2}\, d\theta \\ &= 2A\int \sin^2\frac{\theta}{2}\, d\theta = A(\theta - \sin\theta) + B. \end{aligned}$$

Here, A, B are constants and can be determined from the boundary conditions. At $(x, y) = (0, 0)$, we get

$$0 = y(\theta_a) = 2A\sin^2\theta_a \quad \Rightarrow \quad \theta_a = 0,$$

$$0 = x(\theta_a) = A(\theta - \sin\theta) + B \quad \Rightarrow \quad B = 0.$$

At $(x, y) = (x_b, y_b)$, we solve θ_b and A from

$$\begin{cases} x_b = A(\theta_b - \sin\theta_b) \\ y_b = A(1 - \cos\theta_b). \end{cases}$$

Thus, the solution is a cycloid given in parametric form:

$$\begin{aligned} x &= A(\theta - \sin\theta) \\ y &= A(1 - \cos\theta). \end{aligned}$$

6.6.4 Phase field model

A multi-phase material is a material consisting of more than one phase. For instance, steam water can have liquid and gas phases. Alloy is made of two or more metallic elements. Two-phase material can be modeled by so-called phase field model. It characterizes the material property by an order parameter ϕ through minimizing an energy functional

$$\mathcal{E}[\phi] := \int_{-a}^{b} \left(\frac{\epsilon^2}{2}\phi_x^2(x) + F(\phi(x)) \right) dx.$$

Here, we assume the domain is $[a, b]$. The energy $\int_a^b \frac{\epsilon^2}{2}\phi_x^2(x)\,dx$ is called kinetic energy. It means that the variation of ϕ causes higher energy. The quantity ϵ is a parameter which measures the length of transition from one phase to another phase. The second energy $\int_a^b F(\phi(x))\,dx$ is called the bulk energy. One example of the bulk energy is

$$F(\phi) = -\frac{\delta}{2}\phi^2 + \frac{1}{4}\phi^4.$$

The Euler-Lagrange equation is

$$-\epsilon^2\phi_{xx} + F'(\phi) = -\epsilon^2\phi_{xx} - \phi + \phi^3 = 0. \tag{6.3}$$

In this phase field model, we are interested in the interface profile which connects the two equilibrium phases: ± 1. The interface profile satisfies (6.3) for $x \in \mathbb{R}$ with boundary condition

$$\phi(\pm\infty) = \pm 1, \quad \phi'(\pm\infty) = 0.$$

On the phase-plane (ϕ, ϕ'), you can check that the two equilibria $(\pm 1, 0)$ are saddles. The interface profile ϕ is a heteroclinic orbit connecting these two equilibria.

To find the heteroclinic orbit, we use energy method. First, we can rescale x by replacing it by x'/ϵ. Then the equation becomes

$$\phi_{x'x'} + \phi - \phi^3 = 0.$$

Let us denote $\phi_{x'}$ by ϕ'. We multiply both sides by ϕ' to get

$$\phi'\phi'' + (\phi - \phi^3)\phi' = 0.$$

This can be written as

$$\left(\frac{1}{2}\phi'^2 + \frac{\phi^2}{2} - \frac{\phi^4}{4}\right)' = 0.$$

Thus,

$$\frac{1}{2}\phi'^2 + \frac{\phi^2}{2} - \frac{\phi^4}{4} = C.$$

We plug the end conditions: $\phi(\pm\infty) = \pm 1$ and $\phi'(\pm\infty) = 0$, we get $C = 1/4$. This leads to

$$\frac{1}{2}\phi'^2 = \frac{1}{4} - \frac{\phi^2}{2} + \frac{\phi^4}{4}.$$

$$\phi' = \pm\sqrt{\frac{1}{2} - \phi^2 + \frac{\phi^4}{2}}.$$

$$\frac{d\phi}{1-\phi^2} = \frac{1}{\sqrt{2}}dx'$$

We integrate it and get

$$\frac{1}{2}\ln\left|\frac{1+\phi}{1-\phi}\right| = \frac{1}{\sqrt{2}}x' + C.$$

We look for ϕ satisfy $-1 < \phi < 1$. Thus, we get

$$\frac{1+\phi}{1-\phi} = C_1 e^{\sqrt{2}x'}.$$

We can absorb C_1 into exponential function:

$$C_1 e^{\sqrt{2}x'} = e^{\sqrt{2}(x'-x'_0)}, \quad x_0 \text{ is a constant.}$$

We solve ϕ and get

$$\phi(x') = \frac{e^{\sqrt{2}(x'-x'_0)} - 1}{e^{\sqrt{2}(x'-x'_0)} + 1} = \tanh\left(\frac{x'-x'_0}{\sqrt{2}}\right).$$

Or

$$\phi(x) = \tanh\left(\frac{x-x_0}{\sqrt{2}\epsilon}\right).$$

This is the interface shape function connecting two equilibrium phases $\phi = -1$ and $\phi = +1$.

Homeworks

1. Determine the function $y(x)$ which connects two points (x_0, y_0) and (x_1, y_1) and has minimum surface of revolution rotating about the x-axis.

2. Solve the Euler-Lagrange corresponding to the functional

$$J[y] = \int_1^2 \frac{\sqrt{1+y'^2}}{x}\,dx, \quad y(1) = 0, y(2) = 1.$$

3. Find the general solutions of the Euler-Lagrange corresponding to the functional

$$J[y] = \int_a^b f(x)\sqrt{1+y'^2}\,dx.$$

4. Find the extremal of the functional

$$J[y] = \int \sqrt{x^2+y^2}\sqrt{1+y'^2}\,dx.$$

5. Consider a water droplet on the table. We are interested in the shape of the droplet and its contact angle to the table surface. Let us imagine the droplet is two dimensional. Thus, the shape of the droplet is described by a curve $y(x)$ $0 \le x \le a$ with $y(0) = y(a) = 0$ and $y(x) \ge 0$ for $0 < x < a$. We are interested to determine the shape function $y(x)$, the contact angles $\tan^{-1}(y'(0))$ and $\tan^{-1}(y'(b))$. These quantities are determined by minimizing an energy functional to be described below. Let us denote the water phase by (1), the air phase by (0), and the table phase by (2). The energy consists of three parts: the surface energy between (0) and (1), (1) and (2), (0) and (2). Let us denote the energy per unit length between two phases (i) and (j) by ϵ_{ij}, where $0 \le i \ne j \le 2$. The energy functional is

$$\mathcal{E} = \mathcal{E}_{0,1} + \mathcal{E}_{1,2} + \mathcal{E}_{0,2},$$

where

$$\mathcal{E}_{0,1} = \int_0^a \epsilon_{01}\sqrt{1+y'(x)^2}\,dx,$$

$$\mathcal{E}_{1,2} = \int_0^a \epsilon_{12}\,dx = \epsilon_{12}a$$

$$\mathcal{E}_{0,2} = \epsilon_{02}(L-a).$$

Here, we assume the length of table is L and the droplet occupies the region $(0, a)$. This minimization also subject to an area (volume) constraint: the area of the droplet is fixed:

$$\int_0^a y(x)\,dx = A.$$

Here, ϵ_{01}, ϵ_{02}, ϵ_{12} are given physical parameters. A and L are also given. The unknowns are $y(\cdot)$, a, and the contact angles.

(a) Find the Euler-Lagrange equation of this system.

(b) Prove that the shape function is a portion of a circle.

(c) Derive the formula that the contact angles should satisfy.

Chapter 7

Examples of Nonlinear Systems

There are rich classes of dynamical systems, even in two dimensions. We will introduce

- Hamiltonian systems
- Gradient systems
- Dissipative systems.

7.1 Hamiltonian systems

7.1.1 Motivation

We have seen in the last that a conservative mechanical system:

$$m\ddot{x} = -\nabla V(x)$$

has the first integral $H = \frac{1}{2}m\dot{x}^2 + V(x)$, which is invariant for all t. It can be derived by multiplying the Newton law of motion by $\dot{x}$, we get

$$0 = m\ddot{x}\dot{x} + \nabla V(x)\dot{x} = \frac{d}{dt}\left(\frac{1}{2}m\dot{x}^2 + V(x)\right).$$

From this, we get

$$\frac{1}{2}m|\dot{x}|^2 + V(x) = E$$

for some constant E. When x is a scalar, we can obtain

$$\dot{x} = \pm\sqrt{\frac{2}{m}(E - V(x))}$$

Then $x(t)$ can be obtained by method of separation of variable:

$$\int \frac{dx}{\sqrt{\frac{2}{m}(E - V(x))}} = \pm \int dt.$$

In the derivation above, the quantity $H = \frac{1}{2}m|\dot{x}|^2 + V(x)$ plays a key role. We can express H in a more symmetric way. Define $p = m\dot{x}$, called momentum. Express

$$H(x,p) = \frac{p^2}{2m} + V(x).$$

Then Newton's mechanics is equivalent to

$$\begin{cases} \dot{x} &= H_p(x,p) \\ \dot{p} &= -H_x(x,p). \end{cases} \tag{7.1}$$

An advantage to express the Newton mechanics in this form is that it is easier to find invariants of the flow.

Definition 7.1. *A quantity $f(x,p)$ is called an invariant of the Hamiltonian flow (7.1) if*

$$\frac{d}{dt}f(x(t),p(t)) = 0.$$

From chain rule, we see that f is invariant under the Hamiltonian flow (7.1) if and only if

$$\frac{d}{dt}f(x(t),p(t)) = f_x H_p - f_p H_x = 0.$$

Theorem 7.1. *If a Hamiltonian H is independent of t, then H is invariant under the corresponding Hamiltonian flows (7.1).*

Remarks

- General Lagrangian mechanics:

  $$\delta_x \int_{t_0}^{t_1} L(x_1(t),...,x_n(t),\dot{x}_1(t),...,\dot{x}_n(t))\,dt = 0$$

  can also be written in the form of Hamiltonian system

  $$\begin{aligned} \dot{x}_i &= H_{p_i}(x_1,...,x_n,p_1,...,p_n), \\ \dot{p}_i &= -H_{x_i}(x_1,...,x_n,p_1,...,p_n). \end{aligned}$$

  This will be derived in later section.

- If H is invariant under certain group action, then there are corresponding invariants of the Hamiltonian flow. For instance, if the flow is in two dimensions, say $H(x_1,x_2,p_1,p_2)$. Suppose H is invariant under $x_1 \mapsto x_1 + c$ for any c, that is,

  $$H(x_1+c,x_2,p_1,p_2) = H(x_1,x_2,p_1,p_2), \text{ for any } c \in \mathbb{R}.$$

  An immediate consequence is that

  $$\dot{p}_1 = -H_{x_1} = 0.$$

  Thus, p_1 is an invariant of this Hamiltonian flow. We can eliminate it right away. The system becomes smaller! We shall come back to this point at the end of this chapter.

7.1.2 Trajectories on Phase Plane

A conservative quantity of a time-independent Hamiltonian flow

$$\begin{cases} \dot{x} = H_y(x, y) \\ \dot{y} = -H_x(x, y) \end{cases} \tag{7.2}$$

is the Hamiltonian H itself. That is, along any trajectory $(x(t), y(t))$ of (7.2), we have

$$\frac{d}{dt}H(x(t), y(t)) = H_x\dot{x} + H_y\dot{y} = H_xH_y + H_y(-H_x) = 0.$$

In two dimensions, *the trajectories of a Hamiltonian system in the phase plane are the level sets of its Hamiltonian*.

Example Some linear and nonlinear oscillators are governed by a restoration potential V. The equation of motion in Newton's mechanics is

$$m\ddot{x} = -V'(x)$$

where V is a restoration potential. Define the momentum $y = mv$ and the total energy

$$H(x, y) = \frac{y^2}{2m} + V(x),$$

1. Harmonic oscillator: $H(x, y) = \frac{1}{2}y^2 + \frac{k}{2}x^2$.
2. Duffing oscillator: $H(x, y) = \frac{1}{2}y^2 - \frac{\delta}{2}x^2 + \frac{x^4}{4}$.
3. Cubic potential: $H(x, y) = \frac{1}{2}\left(y^2 - x^2 + x^3\right)$.
4. Simple pendulum: $H(x, y) = \frac{1}{2}y^2 - \frac{g}{l}\cos x$.

You can plot the level sets of H to see the trajectories. In particular, you should pay more attentions on critical points, homoclinic and heteroclinic orbits.

Example Consider fluid flows in a two dimensional domain Ω. The flow is represented as a vector field $\mathbf{V} : \Omega \to \mathbb{R}^2$, or in component form: $\mathbf{V}(x, y) = (u(x, y), v(x, y))$. The flow is called a potential flow if it is *incompressible and irrotational* . That is

$$\nabla \cdot \mathbf{V} = 0, \quad \nabla \times \mathbf{V} = 0.$$

In component form, they are

$$\begin{aligned} u_x + v_y &= 0, \quad \text{incompressible} \\ v_x - u_y &= 0, \quad \text{irrotational.} \end{aligned}$$

From divergence theorem, the first equation yields that there exists a function called *stream function* $\psi(x, y)$ such that

$$u(x, y) = \psi_y(x, y),\ v(x, y) = -\psi_x(x, y).$$

Indeed, from this divergence free condition, we can define the stream function $\psi(x, y)$ by the line integral:

$$\psi(x, y) = \int^{(x,y)} (-v(x, y)dx + u(x, y)dy).$$

1. The starting point of the line integral is not important. What is relevant is the derivatives of ψ. We can choose any point as our starting point. The corresponding ψ is defined up to a constant, which disappears after taking differentiation.

2. By the divergence theorem, the integral is independent of path in a simply connected domain. Hence, ψ is well-defined on simply connected domain. You can check that $\psi_y = u$ and $\psi_x = -v$. If the domain is not simply connected, the steam function may be a multiple valued function. We shall not study this case now.

The second equation for the velocity field yields that

$$\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} = 0.$$

This equation is called a potential equation.

The particle trajectory (which flows with fluid flow) is governed by

$$\begin{aligned} \dot{x} &= u(x, y) = \psi_y(x, y) \\ \dot{y} &= v(x, y) = -\psi_x(x, y). \end{aligned}$$

This is a Hamiltonian flow with Hamiltonian $\psi(x, y)$. The theory of potential flow can be analyzed by complex analysis. You can learn this from text books of complex variable or elementary fluid mechanics.

Here are two examples for the potential flow: let $z = x + iy$

1. $\psi(z) = Im(z^2) = 2xy$,

2. $\psi(z) = Im(z + 1/z) = y - \frac{y}{x^2+y^2}$.

The first one represent a jet. The second is a flow passing a circle .

Example The magnetic field $\mathbf{B}$ satisfies div$\mathbf{B} = 0$. For two-dimensional steady magnetic field $\mathbf{B} = (u, v)$, this reads

$$u_x + v_y = 0.$$

The magnetic field lines are the curves which are tangent to $\mathbf{B}$ at every points on this line. That is, it satisfies

$$\begin{aligned} \dot{x} &= u(x, y) = \psi_y(x, y) \\ \dot{y} &= v(x, y) = -\psi_x(x, y) \end{aligned}$$

where ψ is the stream function corresponding to the divergent free field $\mathbf{B}$.

Example Linear hamiltonian flow. If we consider

$$H(x,y) = \frac{ax^2}{2} + bxy + \frac{cy^2}{2}$$

the corresponding Hamiltonian system is

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} b & c \\ -a & -b \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{7.3}$$

7.1.3 Equilibria of a Hamiltonian system

In this subsection, we want to investigate the property of the equilibria of the Hamiltonian flows. These equilibria are the critical points of the Hamiltonian H.

Definition 7.2. *If $\nabla H(\bar{x},\bar{y}) = 0$, then $(\bar{x},\bar{y})$ is called a critical point of H. Such a critical point is said to be non-degenerate if the hessian of H at $(\bar{x},\bar{y})$ (i.e. the matrix $d^2H(\bar{x},\bar{y})$) is non-singular.*

Since H is usually convex in y variable in mechanical problems, we may assume that $H_{yy} > 0$ at the equilibrium. Notice that this assumption eliminates the possibility of any local maximum of H.

To study the stability of an equilibrium $(\bar{x},\bar{y})$ of the Hamiltonian system (7.2), we linearize it around $(\bar{x},\bar{y})$ to get

$$\dot{\mathbf{u}} = \mathbf{A}\mathbf{u},$$

where $\mathbf{A}$ is the Jacobian of the linearized system of (7.2) at an equilibrium $(\bar{x},\bar{y})$

$$\mathbf{A} = \begin{pmatrix} H_{yx} & H_{yy} \\ -H_{xx} & -H_{xy} \end{pmatrix}_{(\bar{x},\bar{y})}.$$

Since the trace part T of $\mathbf{A}$ is zero, its eigenvalues are

$$\lambda_i = \pm\frac{1}{2}\sqrt{H_{yx}^2 - H_{xx}H_{yy}|_{(\bar{x},\bar{y})}}, i = 1,2.$$

We have the following possibilities.

- H has minimum at $(\bar{x},\bar{y})$. This is equivalent to $H_{xx}H_{yy} - H_{xy}^2 > 0$ at $(\bar{x},\bar{y})$ because we already have $H_{yy} > 0$ from assumption. This is also equivalent to λ_i $i = 1,2$ are pure imaginary. Thus, $(\bar{x},\bar{y})$ is a *center*.
- H has a saddle at $(\bar{x},\bar{y})$. This is equivalent to $H_{xx}H_{yy} - H_{xy}^2 < 0$ at $(\bar{x},\bar{y})$. The corresponding two eigenvalues are real and with opposite signs. Hence the equilibrium is a saddle.
- H cannot have a local maximum at $(\bar{x},\bar{y})$ because the assumption $H_{yy} > 0$.

We summarize it by the following theorem.

Theorem 7.2. *Assuming that $(\bar{x}, \bar{y})$ is a non-degenerate critical point of a Hamiltonian H and assuming $H_{yy}(\bar{x}, \bar{y}) > 0$. Then*

1. *$(\bar{x}, \bar{y})$ is a local minimum of H iff $(\bar{x}, \bar{y})$ is a center of the corresponding Hamiltonian flow.*

2. *$(\bar{x}, \bar{y})$ is a saddle of H iff $(\bar{x}, \bar{y})$ is a saddle of the corresponding Hamiltonian flow..*

The examples we have seen are

1. Simple pendulum: $H(x, p) = \frac{1}{2}p^2 - \frac{g}{l}\cos x$.

2. Duffing oscillator: $H(x, p) = \frac{1}{2}p^2 - \frac{\delta}{2}x^2 + \frac{x^4}{4}$.

3. Cubic potential: $H(x, p) = \frac{1}{2}\left(p^2 - x^2 + x^3\right)$.

In the case of simple pendulum, $(2n\pi, 0)$ are the centers, whereas $(2(n+1)\pi, 0)$ are the saddles. In the case of Duffing oscillator, $(\pm\sqrt{\delta}, 0)$ are the centers, while $(0, 0)$ is the saddle. In the last example, the Hamiltonian system reads

$$\begin{cases} \dot{x} &= p \\ \dot{p} &= x - \frac{3}{2}x^2. \end{cases} \tag{7.4}$$

The state $(0, 0)$ is a saddle, whereas $(3/2, 0)$ is a center.

Below, we use Maple to plot the contour curves the Hamiltonian. These contour curves are the orbits.

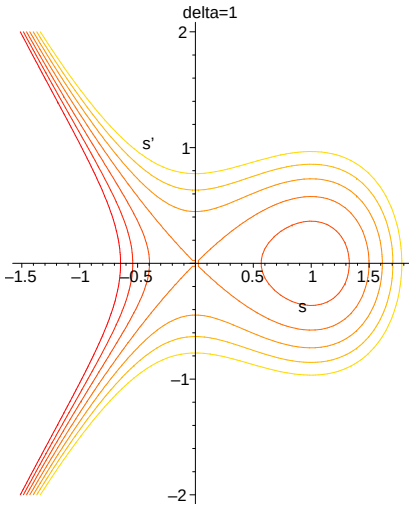
```
> with(DEtools):
> with(plots):

> E := y^2/2+x^3/3-delta*x^2/2;
```

$$E := \frac{1}{2}y^2 + \frac{1}{3}x^3 - \frac{1}{2}\delta x^2$$

Plot the level set for the energy. Due to conservation of energy, these level sets are the orbits.

```
> contourplot(subs(delta=1,E),x=-2..2,y=-2..2,grid=[80,80],contours
> =[-0.3,-0.2,-0.1,0,0.1,0.2,0.3],scaling=CONSTRAINED,labels=[`s`,`s'`],
> title=`delta=1`);
```



7.2 Gradient Flows

In many applications, we look for a strategy to find a minimum of some energy function or entropy function. This minimal energy state is called the ground state. One efficient way is to start from any state then follow the negative gradient direction of the energy function. Such a method is called the steepest descent method. The corresponding flow is called a (negative) gradient flow. To be precise, let us consider an energy function $\psi(x, y)$. We consider the ODE system:

$$\begin{cases} \dot{x} &= -\psi_x(x, y) \\ \dot{y} &= -\psi_y(x, y). \end{cases} \tag{7.5}$$

Along any of such a flow $(x(t), y(t))$, we have

$$\frac{d\psi}{dt}(x(t), y(t)) = \psi_x\dot{x} + \psi_y\dot{y} = -(\psi_x^2 + \psi_y^2) < 0,$$

unless the flow reaches a minimum of ψ.

The gradient flow of ψ is always orthogonal to the Hamiltonian flow of ψ. For if

$$\begin{cases} \dot{x} &= \psi_y(x, y) \\ \dot{y} &= -\psi_x(x, y) \end{cases} \quad \begin{cases} \dot{\xi} &= -\psi_x(\xi, \eta) \\ \dot{\eta} &= -\psi_y(\xi, \eta) \end{cases}$$

then

$$\dot{x}(t) \cdot \dot{\xi}(t) + \dot{y}(t) \cdot \dot{\eta}(t) = 0.$$

Thus, the two flows are orthogonal to each other. We have seen that ψ is an integral of the Hamiltonian flow. Suppose ϕ is an integral of the gradient flow (7.5) (that is, the gradient flows are the level sets of ϕ), then the level sets of ψ and ϕ are orthogonal to each other.

Example 1. Let $\psi = (x^2 - y^2)/2$. Then the gradient flow satisfies

$$\begin{cases} \dot{x} &= -x \\ \dot{y} &= +y. \end{cases}$$

Its solutions are given by $x = x_0 e^{-t}$ and $y = y_0 e^t$. We can eliminate t to obtain that the function $\phi(x, y) := 2xy$ is an integral. If we view these functions on the complex plane: $z = x + iy$, we see that $\psi(z) + i\phi(z) = z^2$.

Example 2. Let $\psi(x, y) = (x^2 + y^2)/2$. The gradient flows are given by

$$\begin{cases} \dot{x} &= -x \\ \dot{y} &= -y. \end{cases}$$

Its solutions are given by $x = x_0 e^{-t}$ and $y = y_0 e^{-t}$. An integral is $\phi = \tan^{-1}(y/x)$. On the other hand, the Hamiltonian flow is given by

$$\begin{cases} \dot{x} &= \psi_y = y \\ \dot{y} &= -\psi_x = -x \end{cases}$$

Its solutions are given by $x = A\sin(t + t_0)$, $y = A\cos(t + t_0)$. The integral is $\psi = (x^2 + y^2)/2$. In fact, $\frac{1}{2}\ln(x^2 + y^2)$ is also an integral of the Hamiltonian flow. The complex valued function $\psi + i\phi = \ln z$.

Example 3. In general, the hamiltonian

$$\psi(x, y) = \frac{ax^2}{2} + bxy + \frac{cy^2}{2}$$

the corresponding Hamiltonian system is

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} b & c \\ -a & -b \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

The gradient flow is

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = -\begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

Find the corresponding integral ϕ of the gradient flow by yourself.

Example 4. Let

$$\psi(x, y) = \frac{y^2}{2} - \frac{x^2}{2} + \frac{x^4}{4}.$$

The gradient flow is

$$\begin{cases} \dot{x} &= -\psi_x = x - x^3 \\ \dot{y} &= -\psi_y = -y \end{cases}$$

The trajectory satisfies

$$\frac{dy}{dx} = \frac{\frac{dy}{dt}}{\frac{dx}{dt}} = \frac{y}{-x + x^3}$$

By the separation of variable

$$\frac{dy}{y} = \frac{dx}{-x + x^3},$$

we get

$$\ln y = \int \frac{dx}{-x + x^3} = -\ln|x| + \frac{1}{2}\ln|1 - x| + \frac{1}{2}\ln|1 + x| + C.$$

Hence, the solutions are given by

$$\phi(x, y) := \frac{x^2 y^2}{1 - x^2} = C_1.$$

Remarks.

- We notice that if ψ is an integral of an ODE system, so is the composition function $h(\psi(x, y))$ for any function h. This is because

$$\frac{d}{dt}h(\psi(x(t), y(t)) = h'(\psi)\frac{d}{dt}\psi(x(t), y(t)) = 0.$$

- If $(0, 0)$ is the center of ψ, then $(0, 0)$ is a sink of the corresponding gradient flow.
- If $(0, 0)$ is a saddle of ψ, it is also a saddle of ϕ.

The properties of a gradient system are shown in the next theorem.

Theorem 7.3. *Consider the gradient system*

$$\begin{cases} \dot{x} &= -\psi_x(x, y) \\ \dot{y} &= -\psi_y(x, y) \end{cases}$$

Assume that the critical points of ψ are isolated and non-degenerate. Then the system has the following properties.

- *The equilibrium is either a souce, a sink, or a saddle. It is impossible to have spiral structure.*
- *If $(\bar{x}, \bar{y})$ is an isolated minimum of ψ, then $(\bar{x}, \bar{y})$ is a sink.*
- *If $(\bar{x}, \bar{y})$ is an isolated maximum of ψ, then $(\bar{x}, \bar{y})$ is a source.*
- *If $(\bar{x}, \bar{y})$ is an isolated saddle of ψ, then $(\bar{x}, \bar{y})$ is a saddle.*

To show these, we see that the Jacobian of the linearized equation at $(\bar{x}, \bar{y})$ is the Hessian of the function ψ at $(\bar{x}, \bar{y})$: is

$$-\begin{pmatrix} \psi_{xx} & \psi_{xy} \\ \psi_{xy} & \psi_{yy} \end{pmatrix}$$

Its eigenvalues $\lambda_i, i = 1, 2$ are

$$-\frac{1}{2}\left(T \pm \sqrt{T^2 - 4D}\right),$$

where $T = \psi_{xx} + \psi_{yy}$, $D = \psi_{xx}\psi_{yy} - \psi_{xy}^2$. From

$$T^2 - 4D = (\psi_{xx} - \psi_{yy})^2 + 4\psi_{xy}^2 \geq 0$$

we have that the imaginary part of the eigenvalues λ_i are 0. Hence the equilibrium can only be a sink, a source or a saddle.

Recall from Calculus that whether the critical point $(\bar{x}, \bar{y})$ of ψ is a local maximum, a local minimum, or a saddle, is completed determined by $\lambda_1, \lambda_2 < 0$, $\lambda_1, \lambda_2 > 0$, or $\lambda_1\lambda_2 < 0$, respectively. On the other hand, whether the equilibrium $(\bar{x}, \bar{y})$ of (7.5) is a source, a sink, or a saddle, is also completed determined by the same conditions.

Homeworks.

1. Consider a linear ODE
$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$
   (a) Show that the system is a hamiltonian system if and only if $a + d = 0$. Find the corresponding hamiltonian.
   (b) Show that the system is a gradient system if and only if $b = c$, i,e. the matrix is symmetric.

7.3 Simple pendulum

Motion on a given curve in a plane A curve $(x(s), y(s))$ in a plane can be parametrized by its arc length s. If the curve is prescribed as we have in the case of simple pendulum, then the motion is described by just a function $s(t)$. By Newton's law, the motion is governed by

$$m\ddot{s} = f(s),$$

where $f(s)$ is the force in the tangential direction of the curve. For instance, suppose the curve is given by $y = y(s)$, and suppose the force is the uniform garvitational force $-mg(0, 1)$, then the force in the tangential direction is

$$f(s) = (\frac{dx}{ds}, \frac{dy}{ds}) \cdot [-mg(0, 1)] = -mg\frac{dy}{ds}.$$

Thus, the equation of motion is

$$\ddot{s} = -g\frac{dy}{ds}. \tag{7.6}$$

For simple pendulum, $s = l\theta$, $(x(\theta), y(\theta)) = (l\sin\theta, -l\cos\theta)$, and

$$\frac{dy}{ds} = \frac{dy}{d\theta}\frac{d\theta}{ds} = -g\sin\theta$$

Hence, the equation of motion is

$$ml\ddot{\theta} = -mg\sin\theta,$$

or in terms of s,

$$m\ddot{s} = -mg\sin\left(\frac{s}{l}\right).$$

7.3.1 global structure of phase plane

We are interested in all possible solutions as a function of its parameters E and t_0. The constant t_0 is unimportant. For the system is autonomous, that is its right-hand side $F(y)$ is independent of t. This implies that if $y(t)$ is a solution, so is $y(t-t_0)$ for any t_0. The trajectories $(y(t), \dot{y}(t))$ and $(y(t-t_0), \dot{y}(t-t_0))$ are the same curve in the phase plane (i.e. y-$\dot{y}$ plane). So, to study the trajectory on the phase plane, the relevant parameter is E. We shall take the simple pendulum as a concrete example for explanation. In this case, $V(\theta) = -\cos(\theta)g/l$.

As we have seen that

$$\frac{\dot{\theta}^2}{2} + V(\theta) = E, \tag{7.7}$$

the total conserved energy. We can plot the equal-energy curve on the phase plane.

$$C_E := \{(\theta, \dot{\theta}) \mid \frac{\dot{\theta}^2}{2} - \frac{g}{l}\cos\theta = E\} \tag{7.8}$$

This is the trajectory with energy E. These trajectories can be classified into the follow categories.

1. **No trajectory**: For $E < -g/l$, the set $\{(\theta, \dot{\theta}) | \frac{\dot{\theta}^2}{2} - \frac{g}{l}\cos\theta = E\}$ is empty. Thus, there is no trajectory with such E.

2. **Equilibria**: For $E = -g/l$, the trajectories are isolated points $(2n\pi, 0)$, $n \in \mathbb{Z}$. These correspond to equibria, namely they are constant state solutions

   $$\theta(t) = 2n\pi, \text{ for all } t.$$

3. **Bounded solutions.** For $-g/l < E < g/l$, the trajectories are bounded closed orbits. Due to periodicity of the cosine function, we see from (7.8) that $(\theta, \dot{\theta})$ is on C_E if and only if $(\theta + 2n\pi, \dot{\theta})$ is on C_E. We may concentrate on the branch of the trajectory lying between $(-\pi, \pi)$, since others are simply duplications of the one in $(-\pi, \pi)$ through the mapping $(\theta, \dot{\theta}) \mapsto (\theta + 2n\pi, \dot{\theta})$.

   For $\theta \in (-\pi, \pi)$, we see that the condition

   $$\frac{\dot{\theta}^2}{2} - \frac{g}{l}\cos\theta = E$$

   implies

   $$E + \frac{g}{l}\cos\theta \geq 0,$$

   or

   $$\cos\theta \geq -\frac{El}{g}.$$

   This forces θ can only stay in $[-\theta_1, \theta_1]$, where

   $$\theta_1 = \cos^{-1}(-El/g).$$

The condition $-g/l < E < g/l$ is equivalent to $0 < \theta_1 < \pi$. The branch of the trajectory C_E in the region $(-\pi, \pi)$ is a closed orbit:

$$\dot{\theta} = \begin{cases} \sqrt{2(E + \frac{g}{l}\cos\theta)} & \text{for } \dot{\theta} > 0, \\ -\sqrt{2(E + \frac{g}{l}\cos\theta)} & \text{for } \dot{\theta} < 0 \end{cases}$$

The solution is bounded in $[-\theta_1, \theta_1]$. The two end states of this orbit are $(\pm\theta_1, 0)$, where the velocity $\dot{\theta} = 0$ and the corresponding angle $\theta = \theta_1$, the largest absolute value. The value θ_1 is called the amplitude of the pendulum.

We integrate the upper branch of this closed orbit by using the method of separation of variable:

$$\int_0^\theta \frac{d\theta}{\sqrt{2(E + \frac{g}{l}\cos\theta)}} = \int dt = \pm(t - t_0)$$

We may normalize $t_0 = 0$ because the system is autonomous (that is, the right-hand side of the differential equation is independent of t). Let us denote

$$t_1 := \int_0^{\theta_1} \frac{d\theta}{\sqrt{2(E + \frac{g}{l}\cos\theta)}}.$$

Let us call

$$\psi(\theta) := \int_0^{\theta} \frac{d\theta}{\sqrt{2(E + \frac{g}{l}\cos\theta)}}.$$

Then $\psi(\theta)$ is defined for $\theta \in [-\theta_1, \theta_1]$ with range $[-t_1, t_1]$. The function ψ is monotonic increasing (because $\psi'(\theta) > 0$ for $\theta \in (-\theta_1, \theta_1)$) Hence, its inversion $\theta(t) = \phi(t)$ is well-defined for $t \in [-t_1, t_1]$. This is the solution $\theta(t)$ in the upper branch of C_E in $(-\pi, \pi)$. We notice that at the end point of this trajectory, $\dot{\theta}(t_1) = 0$. Therefore, for $t > t_1$, we can go to the lower branch *smoothly*:

$$-\int_{\theta_1}^{\theta} \frac{d\theta}{\sqrt{2(E + \frac{g}{l}\cos\theta)}} = t - t_1.$$

This yields

$$-\left(\int_{\theta_1}^{0} + \int_0^{\theta}\right) \frac{d\theta}{\sqrt{2(E + \frac{g}{l}\cos\theta)}} = t - t_1,$$

The first integral is t_1, whereas the second integral is $-\psi(\theta)$. Thus,

$$\psi(\theta) = 2t_1 - t.$$

As θ varies from θ_1 to $-\theta_1$, $2t_1 - t$ varies from t_1 to $-t_1$, or equivalently, t varies from t_1 to $3t_1$. Hence, the solution for $t \in [t_1, 3t_1]$ is

$$\theta(t) := \phi(2t_1 - t).$$

We notice that

$$\theta(t) = \phi(2t_1 - t) = \theta(2t_1 - t) \text{ for } t \in [2t_1, 3t_1]$$

At $t = 3t_1$, $\theta(3t_1) = -\theta_1$ and $\dot{\theta}(3t_1) = 0$. We can continue the time by integrating the upper branch of C_E again. This would give the same orbit. Therefore, we can extend θ periodically with period $T = 2t_1$ by:

$$\theta(t) = \theta(t - 2nT) \text{ for } 2nT \le t \le 2(n+1)T.$$

4. **Another equilibria**: For $E = g/l$, the set C_E contains isolated equilibria:

$$\{((2n+1)\pi, 0) | n \in \mathbb{Z}\} \subset C_E = \{(\theta, \dot{\theta}) \mid \frac{\dot{\theta}^2}{2} - \frac{g}{l}\cos\theta = E\}$$

These equilibria are saddle points, which can be seen by linearizing the system at these equilibria. The nonlinear system is $\ddot{\theta} + \frac{g}{l}\sin\theta = 0$. Near the equilibrium $((2n+1)\pi, 0)$, we write the solution $(\theta, \dot{\theta}) = ((2n+1)\pi + u, \dot{u})$, where $(u, \dot{u})$ is the perturbation, which is small. Plug into the equation, we get

$$\ddot{u} + \frac{g}{l}\sin((2n+1)\pi + u) = 0.$$

For small u, we get the linearized equation

$$\ddot{u} - \frac{g}{l}u \approx 0.$$

The characteristic roots of this linearized system are $\pm\sqrt{g/l}$. Thus, $\mathbf{0}$ is a saddle point of the linearized system.

5. **Heteroclinic orbits**: We can connect two neighboring saddle points $(-\pi, 0)$ and $(\pi, 0)$. This can be thought as a limit of the above case with $E \to g/l$ from below. For $E = g/l$,

$$\dot{\theta} = \pm\sqrt{2(1+\cos(\theta))\frac{g}{l}} = \pm 2\sqrt{\frac{g}{l}}\cos\left(\frac{\theta}{2}\right).$$

Using $t' = \sqrt{\frac{g}{l}}t$, we have

$$\frac{d\theta}{2dt'} = \cos\left(\frac{\theta}{2}\right).$$

Using the polar stereographic projection:

$$u = \tan\left(\frac{\theta}{4}\right), \quad \frac{d\theta}{du} = \frac{4}{1+u^2}, \quad \cos\frac{\theta}{2} = \frac{1-u^2}{1+u^2},$$

we obtain

$$dt' = \frac{2du}{1+u^2}\frac{1+u^2}{1-u^2} = \frac{2du}{1-u^2}.$$

Integrate this, we get

$$t' = \ln\left|\frac{1-u}{1+u}\right|.$$

Here, we normalize a constant $t'_0 = 0$, which is just a shift of time. It is nothing to do with the orbit. Solve u, we obtain

$$u = \frac{1-e^{t'}}{1+e^{t'}} = -\tanh\left(\frac{t'}{2}\right).$$

Since $u = \tan(\theta/4)$, we get

$$\theta = 4\tan^{-1}\left(\tanh\left(\sqrt{\frac{g}{4l}}t\right)\right).$$

This is the orbit connecting $(-\pi, 0)$ to $(\pi, 0)$.

6. **Unbounded solution**: For $E > g/l$, there are two branches of C_E, the upper one ($\dot\theta > 0$) and the lower one ($\dot\theta < 0$). The upper branch: $\dot\theta = \sqrt{2(E + \cos(\theta)g/l)} > 0$ is defined for all $\theta \in \mathbb{R}$. By using the method of separation of variable, we get

$$\int_0^\theta \frac{d\theta}{\sqrt{2\left(E + \frac{g}{l}\cos(\theta)\right)}} = t$$

Let us call the left-hand side of the above equation by $\psi(\theta)$. Notice that $\psi(\theta)$ is a monotonic increasing function defined for $\theta \in (-\infty, \infty)$, because $\psi'(\theta) > \frac{1}{2(E-g/l)} > 0$. The range of ψ is $(-\infty, \infty)$. Its inversion $\phi(t)$ is the solution $\theta = \phi(t)$. Let

$$T := \int_0^{2\pi} \frac{d\theta}{\sqrt{2\left(E + \frac{g}{l}\cos(\theta)\right)}}$$

From the periodicity of the cosine function, we have for $2n\pi \le \theta \le 2(n+1)\pi$,

$$t = \psi(\theta) = \left(\int_0^{2\pi} + \cdots + \int_{2(n-1)\pi}^{2n\pi} + \int_{2n\pi}^{\theta}\right) \frac{d\theta}{\sqrt{2\left(E + \frac{g}{l}\cos(\theta)\right)}}$$

This yields

$$t = nT + \psi(\theta - 2n\pi).$$

Or

$$\theta(t) = 2n\pi + \phi(t - nT), \text{ for } t \in [nT, (n+1)T].$$

7.3.2 Period

Let us compute the period for case 3 in the previous subsection. Recall that

$$\begin{aligned} T &= \int_{-\theta_1}^{\theta_1} \frac{d\theta}{\sqrt{2\left(E+\frac{g}{l}\cos(\theta)\right)}} = \sqrt{\frac{l}{2g}}\int_{-\theta_1}^{\theta_1}\frac{d\theta}{\sqrt{\frac{El}{g}+\cos(\theta)}} \\ &= \sqrt{\frac{l}{2g}}\int_{-\theta_1}^{\theta_1}\frac{d\theta}{\sqrt{\cos(\theta)-\cos(\theta_1)}} = \sqrt{\frac{l}{g}}\int_{-\theta_1}^{\theta_1}\frac{d\theta}{\sqrt{\sin^2\frac{\theta_1}{2}-\sin^2\frac{\theta}{2}}} \end{aligned}$$

where $0 < \theta_1 = \arccos(-El/g) < \pi$ is the amptitude of the pendulum. By the substitution

$$u = \frac{\sin(\theta/2)}{\sin(\theta_1/2)},$$

the above integral becomes

$$T = 2\sqrt{\frac{l}{g}}\int_{-1}^{1}\frac{du}{\sqrt{(1-u^2)(1-k^2u^2)}} \tag{7.9}$$

where $k = \sin(\theta_1/2)$. This integral is called an elliptic integral. This integral cannot be expressed as an elementary function. But we can estimate the period by using

$$1 \geq 1-k^2u^2 \geq 1-k^2$$

for $-1 \leq u \leq 1$ and using $\int_{-1}^{1} 1/\sqrt{1-u^2}\,du = \pi$, the above elliptic integral becomes

$$2\pi\sqrt{\frac{l}{g}} \leq T \leq 2\pi\sqrt{\frac{l}{g}}\left(\frac{1}{1-k^2}\right) \tag{7.10}$$

Homework.

Using Taylor expansion for $(1-k^2u^2)^{-1/2}$, expand the elliptic integral

$$f(k) = \int_{-1}^{1}\frac{du}{\sqrt{(1-u^2)(1-k^2u^2)}}$$

in Taylor series in k for k near 0. You may use Maple to do the integration.

7.4 Cycloidal Pendulum – Tautochrone Problem

7.4.1 The Tautochrone problem

The period of a simple pendulum depends on its amptitude y_1[1]. A question is that can we design a pendulum such that its period is independent of its amptitude. An ancient Greek problem called

[1] Indeed, $k = \sin(y_1/2)$

tautochrone problem answers this question. The tautochrone problem is to find a curve down which a bead placed anywhere will fall to the bottom in the same amount of time. Thus, such a curve can provide a pendulum with period independent of its amptitude. The answer is the cycloid. The cycloidal pendulum oscillates on a cycloid. The equation of a cycloid is

$$\begin{cases} x &= l(\theta + \pi + \sin\theta). \\ y &= -l(1+\cos\theta) \end{cases}$$

Its arc length is

$$\begin{aligned} s &= \int \sqrt{(dx/d\theta)^2 + (dy/d\theta)^2}\, d\theta \\ &= l\int \sqrt{(1+\cos\theta)^2 + \sin^2\theta}\, d\theta \\ &= 2l\int \cos\left(\frac{\theta}{2}\right)\, d\theta \\ &= 4l\sin\left(\frac{\theta}{2}\right). \end{aligned}$$

The force

$$\frac{dy}{ds} = \frac{dy}{d\theta}\frac{d\theta}{ds} = \frac{l\sin\theta}{2l\cos\left(\frac{\theta}{2}\right)} = \sin\left(\frac{\theta}{2}\right) = \frac{s}{4l}.$$

The equation of motion on cycloidal pendulum is

$$\ddot{s} = -\frac{g}{4l}s,$$

a linear equation! Its period is $T = 2\pi\sqrt{l/g}$, which is independent of the amplitude of the oscillation.

Which planar curves produce linear oscillators?

The equation of motion on a planar curve is

$$\ddot{s} = -g\frac{dy}{ds}.$$

The question is: what kind of curve produce linear oscillator. In other word, which curve gives $dy/ds = ks$. This is an ODE for $y(s)$. Its solution is

$$y(s) = \frac{k}{2}s^2.$$

Since s is the arc length of the curve, we have

$$x'(s)^2 + y'(s)^2 = 1.$$

Hence, $x'(s) = \pm\sqrt{1-k^2s^2}$. We use the substitution: $s = \sin(\theta/2)/k$. Then

$$y = \frac{k}{2}s^2 = \frac{1}{2k}\sin^2\left(\frac{\theta}{2}\right) = \frac{1}{4k}(1-\cos\theta).$$

$$x = \int \sqrt{1-k^2s^2}\,ds = \frac{1}{2k}\int \cos^2\left(\frac{\theta}{2}\right)\,d\theta = \frac{1}{4k}\int (1+\cos\theta)\,d\theta = \frac{1}{4k}\left(\theta+\sin\theta\right).$$

Thus, the planar curve that produces linear restoration tangential force is a cycloid.

Ref. http://mathworld.wolfram.com

7.4.2 Construction of a cycloidal pendulum

To construct a cycloidal pendulum [2] , we take $l = 1$ for explanation. We consider the *evolute* of the cycloid

$$x = \pi + \theta + \sin\theta, \;\; y = -1 - \cos\theta. \tag{7.11}$$

In geometry, the *evolute* E of a curve C is the set of all centers of curvature of that curve. On the other hand, if E is the evolute of C, then C is the involute of E. An involute of a curve E can be constructed by the following process. We first wrape E by a thread with finite length. One end of the thread is fixed on E. We then unwrape the thread. The trajectory of the other end as you unwrape the thread forms the *involute* of E. We shall show below that the evolute E of a cycloid C is again a cycloid. With this, we can construct a cycloidal pendulum as follows. We let the mass P is attached by a thread of length 4 to one of the cusps of the evolute E. Under the tension, the thread is partly coincide with the evolute and lies along a tangent to E. The mass P then moves on the cycloid C.

Next, we show that the motion of the mass P lies on the cycloid C. The proof consists of three parts.

1. The evolute of a cycloid is again a cycloid. Suppose C is expressed by $(x(\theta), y(\theta))$. We recall that the curvature of C at a particular point $P = (x(\theta), y(\theta))$ is defined by $d\alpha/ds$, where $\alpha = \arctan(\dot{y}(\theta)/\dot{x}(\theta))$ is the inclined angle of the tangent of C and $ds = \sqrt{\dot{x}^2+\dot{y}^2}\,d\theta$ is the infinitesimal arc length. Thus, the curvature, as expressed by parameter θ, is given by

$$\kappa = \frac{d\alpha}{ds} = \frac{d\alpha}{d\theta}\frac{d\theta}{ds} = \frac{\frac{\dot{x}\ddot{y}-\ddot{x}\dot{y}}{\dot{x}^2}}{1+\left(\frac{\dot{y}}{\dot{x}}\right)^2}\frac{1}{\sqrt{\dot{x}^2+\dot{y}^2}} = \frac{\dot{x}\ddot{y}-\dot{y}\ddot{x}}{(\dot{x}^2+\dot{y}^2)^{3/2}}.$$

The center of curvature of C at $P = (x, y)$ is the center of the osculating circle that is tangent to C at P. Suppose $P' = (\xi, \eta)$ is its coordinate. Then PP' is normal to C (the normal (n_x, n_y) is $(-\dot{y}, \dot{x})/\sqrt{\dot{x}^2+\dot{y}^2}$) and the radius of the osculating circle is $1/\kappa$. Thus, the coordinate of the center

[2] Courant and John's book, Vol. I, pp. 428.

of curvature is

$$\begin{aligned}\xi &= x + \frac{1}{\kappa} n_x = x - \dot{y}\frac{\dot{x}^2 + \dot{y}^2}{\dot{x}\ddot{y} - \dot{y}\ddot{x}},\\ \eta &= y + \frac{1}{\kappa} n_y = y + \dot{x}\frac{\dot{x}^2 + \dot{y}^2}{\dot{x}\ddot{y} - \dot{y}\ddot{x}}.\end{aligned}$$

When $(x(\theta), y(\theta))$ is given by the cycloid equation (7.11),

$$x = \pi + \theta + \sin\theta,\ \ y = -1 - \cos\theta,\ -\pi \le \theta \le \pi,$$

we find that its evolute

$$\xi = \pi + \theta - \sin\theta,\ \ \eta = 1 + \cos\theta, \tag{7.12}$$

is also a cycloid.

2. The evolute of C is the envelope of its normals. We want to find the tangent of the evolute E and show it is identical to the normal of C. To see this, we use arc length s as a parameter on C. With this, the normal $(n_x, n_y) = (-y', x')$ and the curvature $\kappa = x'y'' - y'x''$, where $'$ is d/ds. The evolute is

$$\xi = x - \rho y',\ \ \eta = y + \rho x', \tag{7.13}$$

where $\rho = 1/\kappa$. Thus, the evolute E is also parametrized by s. Since $x'^2 + y'^2 = 1$, we differentiate it in s to get $x'x'' + y'y'' = 0$. This together with $\kappa = x'y'' - y'x''$ yield

$$x'' = -y'/\rho,\ = y'' = x'/\rho.$$

Differentiating (7.13) in s, we can get the tangent of the evolute E:

$$\xi' = x' - \rho y'' - \rho' y' = -\rho' y',\ \eta' = y' + \rho x'' + \rho' x' = \rho' x', \tag{7.14}$$

Therefore,

$$\xi' x' + \eta' y' = 0.$$

This means that the tangent (ξ', η') of the evolute at the center of curvature is parallel to the normal direction $(-y', x')$ of the curve C. Since both of them pass through (ξ, η), they are coincide. In other words, *the normal to the curve C is tangent to the evolute E at the center of curvature.*

3. The end point of the thread P lies on the cycloid C. We show that the radius of curvature plus the length of portion on E where the thread is attched to is 4. To see this, we denote the acr length on the evolute E by σ. The evolute E, as parametrized by the arc length s of C is given by (7.13). Its arc length σ satisfies

$$\left(\frac{d\sigma}{ds}\right)^2 = \xi'^2 + \eta'^2 = (-\rho' y')^2 + (\rho' x')^2 = \rho'^2$$

Here, we have used (7.14). Hence, $\sigma'^2 = \rho'^2$. We take $s = 0$ at $\theta = \pi$ ($(x, y) = (\pi, -2)$). We choose $s > 0$ when $\theta > \pi$. We take $\sigma(0) = 0$ which corresponds to $(\xi, \eta) = (\pi, 2)$. We call this point A (the cusp of the cycloid E). We also choose $\sigma(s) > 0$ for $s > 0$. Notice that $\rho'(s) < 0$. From these normalization, we have

$$\sigma'(s) = -\rho'(s).$$

Now, as the mass moves along C to a point P on C, the center of curvature of C at P is Q which is on the evolute E. We claim that

$$\text{length of the arc } AQ \text{ on } E + \text{the length of the straight line } PQ \ = 4.$$

To see that, the first part above is

$$\int_0^s \sigma' \, ds = -\int_0^s \rho' \, ds = \rho(0) - \rho(s).$$

The second part is simply the radius of curvature $\rho(s)$. Hence the above sum is $\rho(0) = 4$.

Homework.

1. Given a family of curves $\Gamma_\lambda : \{(x(t, \lambda), y(t, \lambda))|t \in \mathbb{R}\}$, a curve E is said to be the *envelop* of Γ_λ if

   (a) For each λ, Γ_λ is tangent to E. Let us denote the tangent point by P_λ¿

   (b) The envelop E is made of P_λ with $\lambda \in \mathbb{R}$.

   Now consider the family of curves to be the normal of a cycliod C, namely

   $$\Gamma_\theta = (x(\theta) + tn_x(\theta), y(\theta) + tn_y(\theta)),$$

   where $(x(\theta), y(\theta))$ is given by (7.11) and (n_x, n_y) is its normal. Using this definition of envelop, show that the envelop of Γ_θ is the cycloid given by (7.12).

7.5 The orbits of planets and stars

7.5.1 Centrally directed force and conservation of angular momentum

The motion of planets or stars can be viewed as a particle moving under a centrally directed field of force:

$$\mathbf{F} = F(r)\hat{\mathbf{e}}_r,$$

where r is the distance from the star to the center, $\mathbf{r}$ is the position vector from the center to the star and

$$\hat{\mathbf{e}}_r = \frac{\mathbf{r}}{r},$$

is the unit director. The equation of motion of the star is

$$\ddot{\mathbf{r}} = F(r)\hat{\mathbf{e}}_r.$$

Define the angular momentum $\mathbf{L} = \mathbf{r} \times \dot{\mathbf{r}}$. We find

$$\frac{d\mathbf{L}}{dt} = \dot{\mathbf{r}} \times \dot{\mathbf{r}} + \mathbf{r} \times \ddot{\mathbf{r}} = F(r)\mathbf{r} \times \hat{\mathbf{e}}_r = 0.$$

Hence , $\mathbf{L}$ is a constant. A function in the state space $(\mathbf{r}, \dot{\mathbf{r}})$ is called an integral if it is unchanged along any orbits. *The integrals can be used to reduce number of unknowns of the system.* The conservation of angular momentum provides us three integrals. Let us write $\mathbf{L} = L\mathbf{n}$ where $L = |\mathbf{L}|$ and $\mathbf{n}$ is a unit vector. The position vector $\mathbf{r}$ and the velocity $\dot{\mathbf{r}}$ always lie on the plane which is perpendicular to $\mathbf{n}$. This plane is called the orbital plane. We use polar coordinates (r, θ) on this plane. Thus, by using the integrals $\mathbf{n}$, which has two parameters, we can reduce the number of unknowns from 6 to 4, that is, from $(\mathbf{r}, \dot{\mathbf{r}})$ to $(r, \theta, \dot{r}, \dot{\theta})$. To find the equation of motion on this plane, we express

$$\mathbf{r} = r\hat{\mathbf{e}}_r = r(\cos\theta, \sin\theta).$$

Define

$$\hat{\mathbf{e}}_\theta := (-\sin\theta, \cos\theta)$$

be the unit vector perpendicular to $\hat{\mathbf{e}}_r$. Then a particle motion on a plane with trajectory $\mathbf{r}(t)$ has the following velocity

$$\dot{\mathbf{r}} = \dot{r}\hat{\mathbf{e}}_r + r\dot{\hat{\mathbf{e}}}_r = \dot{r}\hat{\mathbf{e}}_r + r\dot{\theta}\hat{\mathbf{e}}_\theta.$$

where $\dot{r}$ is the radial speed and $r\dot{\theta}$ is the circular speed. Here, we have used

$$\dot{\hat{\mathbf{e}}}_r = \frac{d}{dt}(\cos\theta, \sin\theta) = \dot{\theta}\hat{\mathbf{e}}_\theta.$$

The acceleration is

$$\begin{aligned} \ddot{\mathbf{r}} &= \ddot{r}\hat{\mathbf{e}}_r + \dot{r}\dot{\hat{\mathbf{e}}}_r + \dot{r}\dot{\theta}\hat{\mathbf{e}}_\theta + r\ddot{\theta}\hat{\mathbf{e}}_\theta + r\dot{\theta}\dot{\hat{\mathbf{e}}}_\theta \\ &= (\ddot{r} - r\dot{\theta}^2)\hat{\mathbf{e}}_r + (2\dot{r}\dot{\theta} + r\ddot{\theta})\hat{\mathbf{e}}_\theta. \end{aligned}$$

Here, we have used $\dot{\hat{\mathbf{e}}}_\theta = -\hat{\mathbf{e}}_r$. In this formula, $\ddot{r}$ is the radial acceleration, and $-r\dot{\theta}^2$ is the centripetal acceleration. The term

$$r(2\dot{r}\dot{\theta} + r\ddot{\theta}) = \frac{d}{dt}(r^2\dot{\theta})$$

is the change of angular momentum. Indeed, the angular momentum is

$$\mathbf{L} = \mathbf{r} \times \dot{\mathbf{r}} = r\hat{\mathbf{e}}_r \times (\dot{r}\hat{\mathbf{e}}_r + r\dot{\theta}\hat{\mathbf{e}}_\theta) = r^2\dot{\theta}\mathbf{n}.$$

The equation of motion $\ddot{\mathbf{r}} = F(r)\hat{\mathbf{e}}_r$ gives

$$\ddot{r} - r\dot{\theta}^2 = F(r), \tag{7.15}$$

$$\frac{d}{dt}(r^2\dot{\theta}) = 0. \tag{7.16}$$

These are the two second-order equations for the unknowns $(r, \theta, \dot{r}, \dot{\theta})$. The θ equation (7.16) can be integrated and gives the conservation of angular momentum

$$r^2\dot{\theta} = \text{ constant } = L. \tag{7.17}$$

If we prescribe an L, the trajectory lies on the set

$$\{(r, \theta, \dot{r}, \dot{\theta}) \,|\, \dot{\theta} = L/r^2\}.$$

We may project this set to the $(r, \theta, \dot{r})$-space and our unknowns now are reduced to $(r, \theta, \dot{r})$. The equations of motion in this space are (7.15) and (7.17).

The integral L can be used to eliminate $\dot{\theta}$ from the first equation. We get

$$\ddot{r} = F(r) + \frac{L^2}{r^3}, \tag{7.18}$$

where the second term on the right-hand side is the centrifugal force. Notice that this equation is independent of θ. Thus, given initial data $(r_0, \theta_0, \dot{r}_0)$ at time $t = 0$, we can find $r(t)$ and $\dot{r}(t)$ from (7.18) by using $(r_0, \dot{r}_0)$ only. We can then use $r^2\dot{\theta} = L$ to find $\theta(t)$:

$$\theta(t) = \theta_0 + \int_0^t \frac{L}{r(t)^2}\, dt.$$

The equation (7.18) can be solved by the energy method. We multiply (7.18) by $\dot{r}$ on both sides to obtain

$$\frac{d}{dt}\left(\frac{1}{2}\dot{r}^2 + \Phi(r) + \frac{1}{2}\frac{L^2}{r^2}\right) = 0,$$

where Φ with $\Phi'(r) = -F(r)$ is the potential. We obtain the law of conservation of energy:

$$\frac{1}{2}\dot{r}^2 + \Phi(r) + \frac{1}{2}\frac{L^2}{r^2} = \text{ constant } = E. \tag{7.19}$$

This energy is another integral. A prescribed energy E defines a surface in the $(r, \theta, \dot{r})$-space. Since the energy $\frac{1}{2}\dot{r}^2 + \Phi(r) + \frac{1}{2}\frac{L^2}{r^2}$ is *independent of* θ (a consequence of centrally forcing), this energy surface is a *cylinder* $C_E \times \mathbb{R}_\theta$, where C_E is the curve defined by (7.19) on the phase plane r-$\dot{r}$.

The equation of motion with a prescribed energy E is

$$\frac{dr}{dt} = \pm\sqrt{2(E - \Phi(r)) - \frac{L^2}{r^2}}. \tag{7.20}$$

It is symmetric about the r-axis. Let us suppose that r_1 and r_2 ($r_1 < r_2$) are two roots of the right-hand side of the above equation:

$$2(E - \Phi(r)) - \frac{L^2}{r^2} = 0$$

and no other root in between. Then the curve defined by (7.20) is a closed curve connecting $(r_1, 0)$ and $(r_2, 0)$. The radial period is defined to be the time the particle travels from $(r_1, 0)$ to $(r_2, 0)$ and back. That is,

$$T_r = 2\int_{r_1}^{r_2} \frac{dr}{\sqrt{2(E - \Phi(r)) - L^2/r^2}}.$$

Next, we shall represent this orbit on the orbital plane (r, θ). From the conservation of angular momentum

$$\frac{d\theta}{dt} = \frac{L}{r^2} \neq 0,$$

we can invert the function $\theta(t)$ and use θ as our independent variable instead of the time variable t. The chain rule gives

$$\frac{d}{dt} = \frac{L}{r^2}\frac{d}{d\theta}.$$

The equation of motion now reads

$$\frac{L}{r^2}\frac{d}{d\theta}\left(\frac{L}{r^2}\frac{dr}{d\theta}\right) - \frac{L^2}{r^3} = F(r). \tag{7.21}$$

The energy equation (7.20) becomes

$$\frac{dr}{d\theta} = \pm\frac{r^2}{L}\sqrt{2(E - \Phi(r)) - \frac{L^2}{r^2}}. \tag{7.22}$$

We can integrate this equation by separation of variable to obtain the trajectory $r = r(\theta)$ in the orbital plane. Sometimes, it is convinient to introduce $u = 1/r$ to simplify the equation (7.21):

$$\frac{d^2u}{d\theta^2} + u = -\frac{F\left(\frac{1}{u}\right)}{L^2u^2}. \tag{7.23}$$

Multiplying $du/d\theta$ on both sides, we get the conservation of energy in u variable:

$$\frac{1}{2}\left(\frac{du}{d\theta}\right)^2 + \frac{u^2}{2} + \frac{\Phi}{L^2} = \frac{E}{L^2}. \tag{7.24}$$

Next, we check the variation of θ as r changes for a radial period. The roots of the right-hand side of (7.22) are equilibria. From (7.20) and (7.22), we see that $dr/d\theta = 0$ if and only if $dr/dt = 0$. Hence these roots are exactly r_1 and r_2 in (7.20). The orbit $r = r(\theta)$ defined by (7.20) must lie between its two extremals where $dr/d\theta = 0$. That is, the orbit $r = r(\theta)$ must lie between the inner circle $r \equiv r_1$ and the outer circle $r \equiv r_2$. The inner radius r_1 is called the pericenter distance, whereas r_2 the apocenter distance.

As the particle travels from pericenter to apocenter and back (i.e. one radial period T_r), the azimuthal angle θ increases by an amount

$$\begin{aligned}\Delta\theta &= 2\int_{r_1}^{r_2}\frac{d\theta}{dr}dr = 2\int_{r_1}^{r_2}\frac{L}{r^2}\frac{dt}{dr}\,dr \\ &= 2L\int_{r_1}^{r_2}\frac{dr}{r^2\sqrt{2(E - \Phi(r)) - L^2/r^2}}.\end{aligned}$$

The azimuthal period is defined as the time that θ varies 2π:

$$T_\theta := \frac{2\pi}{\Delta\theta} T_r.$$

In general, $2\pi/\Delta\theta$ is not a rational number. Hence, the orbit may not be closed.

Below, we see some concrete examples. We shall find the trajectory of the motion $r = r(\theta)$.

Quadratic potential

The potential generated by a homogeneous sphere has the form $\Phi(r) = \frac{1}{2}\Omega^2 r^2$, where Ω is a constant. The force in Cartesian coordinate is $\mathbf{F} = -\Omega^2(x, y)$. Hence the equation of motion is

$$\ddot{x} = -\Omega^2 x, \ \ddot{y} = -\Omega^2 y.$$

We notice that the x and y components are decoupled. Its solution is

$$x(t) = a\cos(\Omega t + \theta_x), \ y(t) = b\cos(\Omega t + \theta_y). \tag{7.25}$$

where a, b and θ_x, θ_y are constants. The orbits are ellipses.

The energy equation is

$$\frac{1}{2}\dot{r}^2 + \frac{\Omega^2}{2} r^2 + \frac{1}{2}\frac{L^2}{r^2} = E.$$

Its contour curves are bounded and symmetric about r and $\dot{r}$ axis. The solution is

$$\dot{r} = \pm\sqrt{2E - \Omega^2 r^2 - \frac{L^2}{r^2}}.$$

The trajectory intersect $\dot{r} = 0$ at r_1 and r_2, where r_i satisfies $2E - \Omega^2 r^2 - \frac{L^2}{r^2}$. This yields

$$r_i^2 = \frac{E \pm \sqrt{E^2 - \Omega^2 L^2}}{\Omega^2}$$

There are two real roots when $E^2 > \Omega^2 L^2$. The above elliptical orbit moves between between r_1 and r_2. From the solution being an ellipse, we can also get that $T_r = T_\theta$.

Homework.

1. Show that the trajectory defined by (7.25) is an ellipse.

2. * Find the integral

$$\Delta\theta := \int_{r_1}^{r_2} \frac{2L}{r^2} \frac{dr}{\sqrt{2E - \Omega^2 r^2 - \frac{L^2}{r^2}}}.$$

Kepler potential

The Kepler force is $F(r) = -GM/r^2$, where M is the center mass, G the gravitational constant. The potential is $\Phi(r) = -GM/r$. From (7.23),

$$\frac{d^2u}{d\theta^2} + u = \frac{GM}{L^2}.$$

This yields

$$u = C\cos(\theta - \theta_0) + \frac{GM}{L^2}$$

where C and θ_0 are constants. By plugging this solution into the energy equation (7.24), we obtain

$$\frac{1}{2}C^2\sin^2(\theta-\theta_0)+\frac{1}{2}C^2\cos^2(\theta-\theta_0)+C\cos(\theta-\theta_0)\cdot\frac{GM}{L^2}+\frac{G^2M^2}{2L^4}-\frac{GM}{L^2}C\cos(\theta-\theta_0)=\frac{E}{L^2}.$$

This yields

$$C = \frac{\sqrt{2E - G^2M^2/L^2}}{L}.$$

We may assume $\theta_0 = 0$. Define

$$e = \frac{CL^2}{GM},\ a = \frac{L^2}{GM(1-e^2)},$$

the eccentricity and the semi-major axis, respectively. The trajectory reads

$$r = \frac{a(1-e^2)}{1+e\cos\theta}. \tag{7.26}$$

This is an ellipse. The pericenter distance $r_1 = a(1-e)$, whereas the apocenter distance $r_2 = a(1+e)$. The periods are

$$T_r = T_\theta = 2\pi\sqrt{\frac{a^3}{GM}}. \tag{7.27}$$

Homework.

1. Prove (7.27).

A perturbation of Kepler potential

Let us consider the potential

$$\Phi(r) = -GM\left(\frac{1}{r} + \frac{a}{r^2}\right).$$

This potential can be viewed as a perturbation of the Kepler potential. The far field is dominated by the Kepler potential. However, in the near field, the force is attractive (but stronger) when $a > 0$ and becomes repulsive when $a < 0$.

The equation for this potential in the r-θ plane is

$$\frac{d^2u}{d\theta^2} + \left(1 - \frac{2GMa}{L^2}\right) u = \frac{GM}{L^2},$$

where $u = 1/r$. Its general solution is

$$\frac{1}{r} = u = C\cos\left(\frac{\theta - \theta_0}{K}\right) + \frac{GMK^2}{L^2},$$

where

$$K = \left(1 - \frac{2GMa}{L^2}\right)^{-1/2}.$$

The constant $K > 1$ for $a > 0$ and $0 < K < 1$ for $a < 0$. The constant C is related to the energy E by

$$E = \frac{1}{2}\frac{C^2L^2}{K^2} - \frac{1}{2}\left(\frac{GMK}{L}\right)^2.$$

The pericenter and apocenter distances are respectively

$$r_1 = \left(\frac{GMK^2}{L^2} + C\right)^{-1},\ r_2 = \left(\frac{GMK^2}{L^2} - C\right)^{-1}.$$

The trajectory in u-θ plane is

$$u = \frac{u_1 + u_2}{2} + \left(\frac{u_1 - u_2}{2}\right)\cos\left(\frac{\theta - \theta_0}{K}\right).$$

Here, $u_1 = 1/r_1$ and $u_2 = 1/r_2$. To plot the trajectory on u-θ plane, we may assume $\theta_0 = 0$. If K is rational, then the orbit is closed. For instance, when $K = 1$, the trajectory is an ellipse. When $K = 3/2$, the particle starts from $(u_1, 0)$, travels to $(u_2, 3/2\pi)$, then back to $(u_1, 3\pi)$, then to $(u_2, (3 + 3/2)\pi)$, finally return to $(r_1, 6\pi)$.

Reference. James Binney and Scott Tremaine, Galactic Dynamics, Princeton University Press, 1987.

Homeworks

1. Consider the Duffing's equation

$$\ddot{s} = -y'(s),\ y(s) = -\delta s^2/2 + s^4/4.$$

   (a) Find the equilibria.

   (b) Plot the level curve of the energy E on the phase plane s-s'.

   (c) Find the period T as a function of E and δ.

(d) Analyze the stability of the equilibria.

2. Consider the equation

$$\ddot{x} = -V'(x),\ V(x) = -\frac{x^2}{2} + \frac{x^3}{3}.$$

(a) Find the equilibria.

(b) Plot the level curve of the energy E on the phase plane s-s'.

(c) Find the period T as a function of E.

(d) Analyze the stability of the equilibria.

(e) There is a special orbit, called the homoclinic orbit, which starts from the orgin, goes around a circle, then comes back to the orgin. Find this orbit on the phase plane and try to find its analytic form.

3. Consider the Kepler problem.

(a) Plot the level curve of E on the phase plane r-$\dot{r}$.

(b) Plot the level curve of E on the r-r' plane, where r' denotes for $dr/d\theta$.

7.6 General Hamiltonian flows

The above Hamiltonian formulation holds for quite general Lagrange mechanics. Consider the action

$$S = \int L(\mathbf{x}, \dot{\mathbf{x}})\, dt,$$

where $L : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. The least action principle gives the Euler-Lagrange equation:

$$\frac{d}{dt} L_{\mathbf{v}}(\mathbf{x}, \dot{\mathbf{x}}) = L_{\mathbf{x}}(\mathbf{x}, \dot{\mathbf{x}}).$$

In mechanical application, $L(\mathbf{x}, \mathbf{v})$ is usually a *convex* function in $\mathbf{v}$. We define the map

$$\mathbf{p} = L_{\mathbf{v}}(\mathbf{x}, \mathbf{v})$$

from $\mathbb{R}^n \to \mathbb{R}^n$. This mapping is 1-1 and has a unique inverse due to the convexity of $L(\mathbf{x}, \cdot)$. We multiply the Euler-Lagrange equation by $\dot{\mathbf{x}}$,

$$\begin{aligned} 0 &= (\frac{d}{dt} L_{\mathbf{v}}) \cdot \dot{\mathbf{x}} - L_{\mathbf{x}} \cdot \dot{\mathbf{x}} \\ &= \frac{d}{dt}(L_{\mathbf{v}} \cdot \dot{\mathbf{x}} - L) \end{aligned}$$

Therefore, we define

$$H(\mathbf{x}, \mathbf{p}) = L_{\mathbf{v}}(\mathbf{x}, \mathbf{v}) \cdot \mathbf{v} - L(\mathbf{x}, \mathbf{v}), \tag{7.28}$$

where $\mathbf{v} := \mathbf{v}(\mathbf{x}, \mathbf{p})$ is the inverse function of $\mathbf{p} = L_{\mathbf{v}}(\mathbf{x}, \mathbf{v})$. This inversion can be expressed in terms of H. Namely,

$$\mathbf{v} = H_{\mathbf{p}}(\mathbf{x}, \mathbf{p}).$$

To see this, we express

$$H(\mathbf{x}, \mathbf{p}) = \mathbf{p} \cdot \mathbf{v}(\mathbf{x}, \mathbf{p}) - L(\mathbf{x}, \mathbf{v}(\mathbf{x}, \mathbf{p})).$$

We differentiate it in $\mathbf{p}$ and get

$$H_{\mathbf{p}} = \mathbf{v} + \mathbf{p} \cdot \mathbf{v}_{\mathbf{p}} - L_{\mathbf{v}}\mathbf{v}_{\mathbf{p}} = \mathbf{v}.$$

We can also compute $H_{\mathbf{x}}$:

$$H_{\mathbf{x}} = \mathbf{p} \cdot \mathbf{v}_{\mathbf{x}}(\mathbf{x}, \mathbf{p}) - L_{\mathbf{x}}(\mathbf{x}, \mathbf{v}(\mathbf{x}, \mathbf{p})) - L_{\mathbf{v}}(\mathbf{x}, \mathbf{v}(\mathbf{x}, \mathbf{p}))\mathbf{v}_{\mathbf{x}}(\mathbf{x}, \mathbf{p}) = -L_{\mathbf{x}}(\mathbf{x}, \mathbf{v}).$$

Thus, the Euler-Lagrange equation

$$\frac{d}{dt}L_{\mathbf{v}} = L_{\mathbf{x}} \tag{7.29}$$

now can be expressed as

$$\dot{\mathbf{p}} = -H_{\mathbf{x}}.$$

Let us summary the above discussion below.

From Euler-Lagrange equation to Hamiltonian equation We start from the minimal action principle to get the Euler-Lagrange equation

$$\frac{d}{dt}L_{\mathbf{v}}(\mathbf{x}, \dot{\mathbf{x}}) = L_{\mathbf{x}}(\mathbf{x}, \dot{\mathbf{x}}).$$

From this, we define the mapping

$$\mathbf{p} := L_{\mathbf{v}}(\mathbf{x}, \mathbf{v}) \text{ and its inverse mapping } \mathbf{v} = \mathbf{v}(\mathbf{x}, \mathbf{p}),$$

and the Hamiltonian

$$H(\mathbf{x}, \mathbf{p}) := \mathbf{p} \cdot \mathbf{v}(\mathbf{x}, \mathbf{p}) - L(\mathbf{x}, \mathbf{v}(\mathbf{x}, \mathbf{p})).$$

We then get

$$\mathbf{v}(\mathbf{x}, \mathbf{p}) = H_{\mathbf{p}}(\mathbf{x}, \mathbf{p}) \text{ and } \dot{\mathbf{p}} = H_{\mathbf{x}}(\mathbf{x}, \mathbf{p}).$$

Now, we claim that if $\mathbf{x}(\cdot)$ is a solution of (7.29) then

$$\begin{cases} \dot{\mathbf{x}} = H_{\mathbf{p}}(\mathbf{x}, \mathbf{p}) \\ \dot{\mathbf{p}} = -H_{\mathbf{x}}(\mathbf{x}, \mathbf{p}) \end{cases} \tag{7.30}$$

with $\mathbf{p}(t) := L_{\mathbf{v}}(\mathbf{x}(t), \dot{\mathbf{x}}(t))$.

From Hamiltonian equation to Euler-Lagrange equation If $(\mathbf{x}(t), \mathbf{p}(t))$ satisfies (7.30), then define

$$\mathbf{v} = H_{\mathbf{p}}(\mathbf{x}, \mathbf{p}) \text{ and its inverse map } \mathbf{p}(\mathbf{x}, \mathbf{v}),$$

and the Lagrangian

$$L(\mathbf{x}, \mathbf{v}) = \mathbf{v} \cdot \mathbf{p}(\mathbf{x}, \mathbf{v}) - H(\mathbf{x}, \mathbf{p}(\mathbf{x}, \mathbf{v})),$$

Then $\mathbf{x}(\cdot)$ satisfies the Euler-Lagrange equation (7.29).

7.6.1 Noether Theorem

Project

1. Write a model for double pendulum. Solve it numerically, analyze it.

Chapter 8

General Theory for ODE in $\mathbb{R}^n$

8.1 Well-postness

8.1.1 Local existence

In this section, we develop general theory for the initial value problem

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)), \tag{8.1}$$

$$\mathbf{y}(t_0) = \mathbf{y}_0. \tag{8.2}$$

This includes, the existence, uniqueness, continuous dependence of the initial data. In the next section, we will develop the general stability theory. This includes the linear stability analysis and method of Lyapunov function.

We assume that $\mathbf{f} : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous. We are interested in existence of solutions in a neighborhood of $(t_0, \mathbf{y}_0)$. Let us choose such a neighborhood, say $J = [t_0 - \tau_0, t_0 + \tau_0]$ and $V = \{\mathbf{y} || \mathbf{y} - \mathbf{y}_0| \leq R\}$. Let us denote $\max\{|\mathbf{f}(s, \mathbf{y})| | (s, \mathbf{y}) \in J \times V\}$ by M.

Definition 8.1. *We say that $\mathbf{f}(s, \mathbf{y})$ is Lipschitz continuous in $\mathbf{y}$ in a neighborhood $J \times V$ if there exists a constant L such that*

$$|\mathbf{f}(s, \mathbf{y}_1) - \mathbf{f}(s, \mathbf{y}_2)| \leq L|\mathbf{y}_1 - \mathbf{y}_2|$$

for any $\mathbf{y}_1, \mathbf{y}_2 \in V$ and any $s \in J$.

If $\mathbf{f}(s, \mathbf{y})$ is continuously differentiable in $\mathbf{y}$ on $J \times V$, then by the mean value theorem, it is also Lipschitz continuous in $\mathbf{y}$.

Theorem 8.1 (Local Existence, Cauchy-Peano theory)**.** *Consider the initial value problem (8.1), (8.2). Suppose $\mathbf{f}(t, \mathbf{y})$ is continuous in $(t, \mathbf{y})$ and Lipschitz continuous in $\mathbf{y}$ in a neighborhood of $(t_0, \mathbf{y}_0)$, then the initial value problem (8.1) and (8.2) has a solution $\mathbf{y}(\cdot)$ in $[t_0 - \delta, t_0 + \delta]$ for some $\delta > 0$.*

Proof. We partition the existence proof into following steps.

1. Convert (8.1) (8.2) into an equivalent integral equation. We can integrate (8.1) in t and obtain

$$\mathbf{y}(t) = \mathbf{y}_0 + \int_{t_0}^{t} \mathbf{f}(s, \mathbf{y}(s))\, ds. \tag{8.3}$$

This is an integral equation for $\mathbf{y}(\cdot)$. We claim that *the initial value problem (8.1) (8.2) is equivalent to the integral equation (8.3).*
We have seen the derivation from (8.1) and (8.2) to (8.3). Conversely, if $\mathbf{y}(\cdot)$ is continuous and satisfies (8.3), then $\mathbf{f}(\cdot, \mathbf{y}(\cdot))$ is continuous. Hence, $\int_{t_0}^{t} \mathbf{f}(s, \mathbf{y}(s))\, ds$ is differentiable. By the Fundamental Theorem of Calculus, we get $\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t))$. Hence, $\mathbf{y}(\cdot)$ is differentiable and satisfies (8.1). At $t = t_0$, the integral part of (8.3) is zero. Hence $\mathbf{y}(t_0) = \mathbf{y}_0$.

2. We shall use method of contraction map to solve this integral equation in the function space $C(I)$, the space of all continuous functions on interval I. First, let me introduce the function space $C(I)$:

$$C(I) := \{\mathbf{y} |\, \mathbf{y} : I \to \mathbb{R}^n \text{ is continuous}\}.$$

Here, $I = [t_0 - \delta, t_0 + \delta]$ is an interval of existence. The parameter $\delta \le \tau_0$ will be chosen later. In $C(I)$, we define a norm

$$\|\mathbf{y}\| = \max_{t \in I} |\mathbf{y}(t)|.$$

It is a fact that, with this norm, every Cauchy sequence $\{\mathbf{y}^n\}$ in $C(I)$ converges to $\mathbf{y} \in C(I)$.

3. We perform Picard iteration to generate approximate solutions: define

$$\mathbf{y}^0(t) \equiv \mathbf{y}_0$$

$$\mathbf{y}^{n+1}(t) = \mathbf{\Phi}(\mathbf{y}^n)(t) := \mathbf{y}_0 + \int_{t_0}^{t} \mathbf{f}(s, \mathbf{y}^n(s))\, ds,\ n \ge 1. \tag{8.4}$$

We will show that $\{\mathbf{y}^n\}$ is a Cauchy in $C(I)$. But, first, we need to show that $\Phi(\mathbf{y})(t)$ stay in V for $|t - t_0|$ small enough so that the Lipschitz condition of $\mathbf{f}$ can be applied.

4. Let us consider the closed ball

$$X := \{\mathbf{y} \in C(I) | \|\mathbf{y} - \mathbf{y}_0\| \le R\} \subset C(I).$$

We claim that if $\mathbf{y} \in X$, then $\Phi(\mathbf{y}) \in X$, provided $\delta \le \frac{R}{M}$. This is because

$$\|\Phi(\mathbf{y}) - \mathbf{y}_0\| = \left| \int_0^t \mathbf{f}(s, \mathbf{y}(s))\, ds \right| \le \int_0^t |\mathbf{f}(s, \mathbf{y}(s))|\, ds \le Mt \le \delta.$$

5. We claim that the sequence $\{\mathbf{y}^n\}$ is a Cauchy sequence in $C(I)$, provided δ is small enough. From (8.4), we have

$$\begin{aligned} \|\mathbf{y}^{n+1} - \mathbf{y}^n\| &= \|\mathbf{\Phi}(\mathbf{y}^n) - \mathbf{\Phi}(\mathbf{y}^{n-1})\| \le \int_{t_0}^{t} |\mathbf{f}(s, \mathbf{y}^n(s)) - \mathbf{f}(s, \mathbf{y}^{n-1}(s))|\, ds \\ &\le \int_{t_0}^{t} L|\mathbf{y}^n(s) - \mathbf{y}^{n-1}(s)|\, ds \le \tau L \|\mathbf{y}^n - \mathbf{y}^{n-1}\| \end{aligned}$$

Here, L is the Lipschitz constant of $\mathbf{f}$ in $J \times U$. We choose a constant $\rho < 1$ and choose δ such that

$$\delta = \min\{\frac{\rho}{L}, \frac{R}{M}\}. \tag{8.5}$$

With this δ, $\mathbf{y}^n \in X$ and

$$\|\mathbf{y}^m - \mathbf{y}^n\| \leq \sum_{k=n}^{m-1} \|\mathbf{y}^{k+1} - \mathbf{y}^k\| \leq \sum_{n}^{m-1} \rho^k < \epsilon,$$

provided $n < m$ are large enough.

6. By the completeness of $C(I)$, $\mathbf{y}^n$ converges to a function $\mathbf{y} \in C(I)$. This convergence above is called *uniform convergence*. It means

$$\lim_{n\to\infty} \max_{s\in I} |\mathbf{y}_n(s) - \mathbf{y}(s)| = 0.$$

This implies that $\mathbf{y}_n(s) \to \mathbf{y}(s)$ for every $s \in I$. This also yields that, for every $s \in I$, $\lim_{n\to\infty} \mathbf{f}(s, \mathbf{y}^n(s)) = \mathbf{f}(s, \mathbf{y}(s))$, because f is continuous in $\mathbf{y}$. By the continuity of integration, we then get

$$\int_{t_0}^{t} \mathbf{f}(s, \mathbf{y}^n(s))\, ds \to \int_{t_0}^{t} \mathbf{f}(s, \mathbf{y}(s))\, ds$$

for any $t \in I$. By taking limit $n \to \infty$ in (8.4), we get that $\mathbf{y}(\cdot)$ satisfies the integral equation (8.3).

□

8.1.2 Uniqueness

Theorem 8.2. *If $\mathbf{f}(s, \mathbf{y})$ is Lipschitz continuous in $\mathbf{y}$ in a neighborhood of $(t_0, \mathbf{y}_0)$, then the initial value problem*

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)), \quad \mathbf{y}(0) = \mathbf{y}_0$$

has a unique solution in the region where the solution exists.

Proof. Suppose $\mathbf{y}_1(\cdot)$ and $\mathbf{y}_2(\cdot)$ are two solutions. Then Let $\eta(t) := |\mathbf{y}_2(t) - \mathbf{y}_1(t)|$. [1] We have

$$\begin{aligned} \eta'(t) &\leq |(\mathbf{y}_2(t) - \mathbf{y}_1(t))'| \leq |\mathbf{f}(t, \mathbf{y}_2(t)) - \mathbf{f}(t, \mathbf{y}_1(t))| \\ &\leq L|\mathbf{y}_2(t) - \mathbf{y}_1(t)| = L\eta(t) \end{aligned}$$

[1] The norm here can be any norm in $\mathbb{R}^n$. What we need is the triangle inequality which gives

$$|\mathbf{y}_1(t) - \mathbf{y}_2(t)|' \leq |\mathbf{y}_1'(t) - \mathbf{y}_2'(t)|.$$

The $|\mathbf{y}|_2 = \sqrt{y_1^2 + \cdots + y_n^2}$ has this property.

We get the following differential inequality

$$\eta'(t) - L\eta(t) \leq 0.$$

Multiplying e^{-Lt} on both sides, we get

$$\left(e^{-Lt}\eta(t)\right)' \leq 0.$$

Hence

$$e^{-Lt}\eta(t) \leq \eta(0).$$

But $\eta(0) = 0$ (because $\mathbf{y}_1(0) = \mathbf{y}_2(0) = \mathbf{y}_0$) and $\eta(t) = |\mathbf{y}_1(t) - \mathbf{y}_2(t)| \geq 0$, we conclude that $\eta(t) \equiv 0$. □

If f does not satisfies the Lipschitz condition, then a counter example does exist. Typical counter example is

$$y'(t) = 2\sqrt{y},\ y(0) = 0.$$

Any function has the form

$$y(t) = \begin{cases} 0 & t < c \\ (t-c)^2 & t \geq c \end{cases}$$

with arbitrary $c \geq 0$ is a solution.

Homework Let $|\mathbf{y}| := \sqrt{y_1^2 + \cdots + y_n^2}$. Let $\mathbf{y} : \mathbb{R} \to \mathbb{R}^n$ be a smooth function. Show that

$$|\mathbf{y}|' \leq |\mathbf{y}'|.$$

8.1.3 Continuous dependence on initial data

Let us denote the solution to the ODE

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \quad \mathbf{y}(t_0) = \xi$$

by $\mathbf{y}(t, \xi)$. We shall show that the solution continuously depends on its initial data ξ. If $\mathbf{f}$ is twice differentiable in $\mathbf{y}$, then $\mathbf{y}(\cdot, \xi)$ is also differentiable in ξ.

Theorem 8.3. *Under the same assumption of f in the local existence theorem above, the solution $\mathbf{y}(t; \xi)$ of the ODE: $\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$, $\mathbf{y}(t_0, \xi) = \xi$ is a continuous function in ξ in a neighborhood of $\mathbf{y}_0$. That is, the solution $\mathbf{y}(\cdot, \xi)$ continuously depends on its initial data ξ.*

Proof. The proof is a simple modification of the proof of the local existence theorem.

1. Let us define $I = [t_0 - \delta, t_0 + \delta]$ (δ is to be determined later), $U = \overline{B_{R/2}(\mathbf{y}_0)}$ and consider

   $$C(I \times U) := \{\mathbf{y}(\cdot) | \, \mathbf{y} : I \times U \to \mathbb{R}^n \text{ is continuous}\},$$

   with the norm

   $$\|\mathbf{y}_1 - \mathbf{y}_2\| := \max_{(t,\xi) \in I \times U} |\mathbf{y}_1(t, \xi) - \mathbf{y}_2(t, \xi)|$$

   $$X = \{\mathbf{y} \in C(I \times U) | \|\mathbf{y} - \mathbf{y}_0\| \leq R/2\}.$$

   Here, $\mathbf{y}_0$ denotes for both the constant and the constant function in t with value $\mathbf{y}_0$.

2. We define the Picard iteration to be:

$$\mathbf{y}^{n+1}(t,\xi) := \mathbf{\Phi}(\mathbf{y}^n)(t,\xi) := \xi + \int_{t_0}^{t} \mathbf{f}(s,\mathbf{y}^n(s,\xi))\,ds,$$

$$\mathbf{y}^0(t,\xi) \equiv \xi,$$

for $(t,\xi) \in I \times U$.

3. We claim that if $\mathbf{y} \in X$, then $|\Phi(\mathbf{y})(t,\xi) - \mathbf{y}_0| \le R$. This is because

$$|\mathbf{\Phi}(\mathbf{y}) - \mathbf{y}_0| \le |\xi - \mathbf{y}_0| + \int_{t_0}^{t} |\mathbf{f}(s,\mathbf{y}(s,\xi))|\,ds \le \frac{R}{2} + M\delta \le R$$

provided $\delta \le R/(2M)$.

4. The sequence $\{\mathbf{y}^n\}$ is a Cauchy sequence in $C(I \times U)$ provided δ is small. From (8.4), we have

$$\begin{aligned} \|\mathbf{y}^{n+1} - \mathbf{y}^n\| &= \|\mathbf{\Phi}(\mathbf{y}^n) - \mathbf{\Phi}(\mathbf{y}^{n-1})\| \le \int_{t_0}^{t} |\mathbf{f}(s,\mathbf{y}^n(s)) - \mathbf{f}(s,\mathbf{y}^{n-1}(s))|\,ds \\ &\le \int_{t_0}^{t} L|\mathbf{y}^n(s) - \mathbf{y}^{n-1}(s)|\,ds \le \tau L \|\mathbf{y}^n - \mathbf{y}^{n-1}\| \end{aligned}$$

We choose a constant $\rho < 1$ and choose δ such that

$$\delta = \min\{\frac{\rho}{L}, \frac{R}{2M}\}. \tag{8.6}$$

With this δ, $\mathbf{y}^n \in X$ and

$$\|\mathbf{y}^m - \mathbf{y}^n\| \le \sum_{k=n}^{m-1} \|\mathbf{y}^{k+1} - \mathbf{y}^k\| \le \sum_{n}^{m-1} \rho^k < \epsilon,$$

provided $n < m$ are large enough.

5. The sequence $\{\mathbf{y}^n\}$ converges to a function $\mathbf{y} \in C(I \times U)$ and satisfies

$$\mathbf{y}(t,\xi) = \xi + \int_{t_0}^{t} \mathbf{f}(s,\mathbf{y}(s,\xi))\,ds$$

which solves the ODE: $\mathbf{y}' = \mathbf{f}(t,\mathbf{y})$, $\mathbf{y}(t_0,\xi) = \xi$. Furthermore, $\mathbf{y}(t,\xi)$ is continuous in ξ.

□

Remark. Given a function $\mathbf{y} \in C(I \times U)$. For each fixed $\xi \in U$, $\mathbf{y}(\cdot, \xi) \in C(I)$ for every $\xi \in U$. Thus, $\mathbf{y}$ can be viewed as a function $\mathbf{z} : U \to C(I)$ defined by $\mathbf{z}(\xi) = \mathbf{y}(\cdot, \xi) \in C(I)$. This function is indeed a continuous function from U to $C(I)$. This means that if we define

$$C(U, C(I)) := \{\mathbf{w} : U \to C(I) \text{ is continuous.}\}$$

equipped with the norm

$$\|\mathbf{w}\| := \max_{\xi \in U} \|\mathbf{w}(\xi)\|,$$

then $\mathbf{z} \in C(U, C(I))$. This can be proven by the following arguments. Because $\mathbf{y} \in C(I \times U)$, we have for any $\epsilon > 0$ small, there exists $\delta_1 > 0$ such that

$$|\mathbf{y}(t, \xi_1) - \mathbf{y}(t, \xi_2)| < \epsilon$$

for all $t \in I$ and $\xi_1, \xi_2 \in U$ with $|\xi_1 - \xi_2| < \delta_1$. This δ_1 is independent of t because of the uniform continuity of $\mathbf{y}$ on $I \times U$. We can take maximum in $t \in I$, then obtain

$$\|\mathbf{z}(\xi_1) - \mathbf{z}(\xi_2)\| = \|\mathbf{y}(\cdot, \xi_1) - \mathbf{y}(\cdot, \xi_2)\| := \max_{t \in I} |\mathbf{y}(t, \xi_1) - \mathbf{y}(t, \xi_2)| \leq \epsilon.$$

Thus, $\mathbf{z} \in C(U, C(I))$.

Conversely, given a function $\mathbf{z} \in C(U, C(I))$, $\mathbf{z}(\xi) \in C(I)$ for any $\xi \in U$. Thus, we can associate it with a function $\mathbf{y} : I \times U \to \mathbb{R}^n$ defined by $\mathbf{y}(t, \xi) := \mathbf{z}(\xi)(t)$. This function is indeed uniformly continuous on $I \times U$. Thus, $C(I \times U)$ is the same as the space $C(U, C(I))$.

Homework

1. Let I be closed interval and $U \in \mathbb{R}^n$ be a closed ball. Prove that if $\mathbf{z} \in C(U, C(I))$, then the function $\mathbf{y}(t, \xi) := \mathbf{z}(\xi)(t)$ is uniformly continuous on $I \times U$.

2. *The continuity of $\mathbf{y}(t, \xi)$ in ξ can become differentiable if the Jacobian $[\partial \mathbf{f}/\partial \mathbf{y}](t, \mathbf{y})$ is Lipschitz continuous in $\mathbf{y}$. Prove such a result and show that along a solution $\mathbf{y}(\cdot, \xi)$, the Jacobian $[\partial \mathbf{y}(\cdot, \xi)/\partial \xi]$, which is an $n \times n$ matrix, satisfies the matrix ODE:

$$\frac{d}{dt}\left[\frac{\partial \mathbf{y}}{\partial \xi}(t, \xi)\right] = \left[\frac{\partial \mathbf{f}}{\partial \mathbf{y}}(t, \mathbf{y}(t, \xi))\right]\left[\frac{\partial \mathbf{y}}{\partial \xi}\right].$$

3. Show that $det\left[\frac{\partial \mathbf{y}}{\partial \xi}(t, \xi)\right] \neq 0$ if $det\left[\frac{\partial \mathbf{y}}{\partial \xi}(0, \xi)\right] \neq 0$

8.1.4 A priori estimate and global existence

The global existence results are usually followed from so-called a priori estimate plus the local existence result. Let us recall the *a priori estimate* for scalar equations.

Examples of a priori estimates

1. In logistic equation $y' = ry(1 - y/K)$, if a solution $y(\cdot)$ with $0 < y(0) < K$ exists, then it always satisfies $0 < y(t) < K$ for all t. This is because $y(t) \equiv 0$ and $y(t) \equiv K$ are two equilibrium solutions and no solution can cross equilibrium (uniqueness theorem). Such kind of estimate is called an a priori estimate.

2. In spring-mass model $m\ddot{x} - kx = 0$, if the solution exists, then it always satisfies

$$\frac{1}{2}\dot{x}^2 + kx^2 = E$$

for some constant $E > 0$. This automatically gives boundedness of $(x(t), \dot{x}(t))$, as long as it exists. The estimate is called a priori estimate.

Global existence theorem

Theorem 8.4. *Consider* $\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$. *If a solution* $\mathbf{y}(t)$ *stays bounded as long as it exists, then such a solution exists for all* $t \in \mathbb{R}$.

Proof. Suppose a solution exists in $[0, T)$ and cannot be extended. By the assumption of boundedness, the limit $\mathbf{y}(T-)$ exists. This is because $\mathbf{y}(\cdot)$ is bounded, hence $\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t))$ is bounded and continuous for $t \in [0, T)$. Hence the limit

$$\mathbf{y}(T-) = \lim_{t \to T-} \mathbf{y}(0) + \int_0^t \mathbf{y}'(s)\, ds$$

exists. We can extend $\mathbf{y}(\cdot)$ from T with the $\mathbf{y}(T+) = \mathbf{y}(T-)$. By the local existence theorem, the solution can be extended for a short time. Now, we have a solution on two sides of T with the same data $\mathbf{y}(T-)$, we still need to show that it satisfies the equation at $t = T$. To see this, on the right-hand side

$$\lim_{t \to T+} \mathbf{y}'(t) = \lim_{t \to T+} \mathbf{f}(t, \mathbf{y}(t)) = \mathbf{f}(T, \mathbf{y}(T-)).$$

On the left-hand side, we also have

$$\lim_{t \to T-} \mathbf{y}'(t) = \lim_{t \to T-} \mathbf{f}(t, \mathbf{y}(t)) = \mathbf{f}(T, \mathbf{y}(T-)).$$

Therefore $\mathbf{y}'(t)$ is continuous at T and $\mathbf{y}'(T) = \mathbf{f}(T, \mathbf{y}(T))$. Hence we get the extended solution also satisfies the equation at T. This is a contradiction. □

Below, we give several examples of a priori estimates.

Example 1 A vector field $\mathbf{f}(t, \mathbf{y})$ is said to grow at most linearly as $|\mathbf{y}| \to \infty$ if there exist some positive constants a, b such that

$$|\mathbf{f}(t, \mathbf{y})| \leq a|\mathbf{y}| + b \tag{8.7}$$

whenever $|\mathbf{y}|$ is large enough.

Theorem 8.5. *If $\mathbf{f}(t, \mathbf{y})$ is smooth and grows at most linearly as $|\mathbf{y}| \to \infty$, then all solutions of ODE $\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$ can be extended to $t = \infty$.*

Proof. Suppose a solution exists in $[0, T)$, we give a priori estimate for this solution. From the grow condition of $\mathbf{f}$, we have

$$|\mathbf{y}(t)|' \leq |\mathbf{y}'(t)| \leq a|\mathbf{y}(t)| + b.$$

Multiplying e^{-at} on both sides, we get

$$\left(e^{-at}|\mathbf{y}(t)|\right)' \leq e^{-at}b.$$

Integrating t from 0 to T, we obtain

$$e^{-aT}|\mathbf{y}(T)| - |\mathbf{y}(0)| \leq \int_0^T e^{-at}b\, dt = \frac{b}{a}\left(1 - e^{aT}\right).$$

Hence

$$|\mathbf{y}(T)| \leq |\mathbf{y}(0)|e^{aT} + \frac{b}{a}e^{aT}.$$

Such an estimate is called *a priori estimate* of solutions. It means that as long as solution exists, it satisfies the above estimate. □

Remarks.

1. We can replace the growth condition by

$$|\mathbf{f}(t, \mathbf{y})| \leq a(t)|\mathbf{y}| + b(t) \tag{8.8}$$

   where $a(t)$ and $b(t)$ are two positive functions and locally integrable, which means

$$\int_I a(t)\, dt, \int_I b(t)\, dt < \infty$$

   for any bounded interval I.

2. In the proofs of the uniqueness theorem and the global existence theorem, we use so called the Gronwall inequality, which is important in the estimate of solutions of ODE.

Lemma 8.1 (Gronwall inequality)**.** *If*

$$\eta'(t) \leq a(t)\eta(t) + b(t) \tag{8.9}$$

then

$$\eta(t) \leq e^{\int_0^t a(s)\, ds}\eta(0) + \int_0^t e^{\int_s^t a(\tau)\, d\tau}b(s)\, ds \tag{8.10}$$

Proof. Let $A(t) = \int_0^t a(s)\,ds$. We multiply (8.9) by the integration factor $e^{-A(t)}$ to get

$$e^{-A(t)}\eta'(t) - a(t)e^{-A(t)}\eta(t) \le e^{-A(t)}b(t).$$

This gives

$$\left(e^{-A(t)}\eta(t)\right)' \le e^{-A(t)}b(t).$$

We rename the independent variable as s then integrate this inequality in s from 0 to t. We get

$$e^{-A(t)}\eta(t) \le \eta(0) + \int_0^t e^{-A(s)}b(s)\,ds.$$

Multiply both sides by $e^{A(t)}$, we get (8.10). □

Gronwall inequality can also be used to show that the continuous dependence of solution to its initial data.

Homework

1. **Gronwall inequality in integral form** Suppose $\eta(t)$ satisfies

$$\eta(t) \le \eta(0) + \int_0^t L\eta(s) + b(s)\,ds$$

Show that $eta(t)$ satisfies

$$e^{-Lt}\eta(t) \le \eta(0) + \int_0^t e^{-L(t-s)}b(s)\,ds.$$

Hint: Let $\zeta(t) := \int_0^t \eta(s)\,ds$. Then $\zeta(t)$ satisfies

$$\zeta' = \zeta'(0) + L\zeta + B(t), \quad B(t) = \int_0^t b(s)\,ds.$$

Use the differential form of the Gronwall inequality.

2. Generalize the above Gronwall inequality to

$$\eta(t) \le \eta(0) + \int_0^t a(s)\eta(s) + b(s)\,ds.$$

Example 2: Lyapunov functional and a priori estimate

Theorem 8.6. *Consider the ODE in* $\mathbb{R}^n$*:*

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}),\ \mathbf{y}(0) = \mathbf{y}_0.$$

Suppose there exists a function Φ *such that*

(i) $\nabla\Phi(\mathbf{y}) \cdot \mathbf{f}(\mathbf{y}) \le 0$, *and*

(ii) $\Phi(\mathbf{y}) \to \infty$ *as* $|\mathbf{y}| \to \infty$.

Then the solution exists on $[0, \infty)$.

Proof. Consider $\Phi(\mathbf{y}(t))$. It is a non-increasing function because

$$\frac{d}{dt}\Phi(\mathbf{y}(t)) = \nabla\Phi(\mathbf{y}(t)) \cdot \mathbf{f}(\mathbf{y}(t)) \le 0$$

Thus,

$$\Phi(\mathbf{y}(t)) \le \Phi(\mathbf{y}(0))$$

Since $\Phi(\mathbf{y}) \to \infty$ as $\mathbf{y} \to \infty$, the set

$$\{\mathbf{y} | \Phi(\mathbf{y}) \le \Phi(\mathbf{y}_0)\}$$

is a bounded set. If the maximal existence of interval is $[0, T)$ with $T < \infty$, then $\mathbf{y}(\cdot)$ is bounded in $[0, T)$ and can be extended to T. By the local existence of ODE, we can always extend $\mathbf{y}(\cdot)$ to $T + \epsilon$. This is a contradiction. Hence $T = \infty$. □

As an example, let us consider a damping system

$$\ddot{\mathbf{x}} + \gamma\dot{\mathbf{x}} = -V'(\mathbf{x})$$

where V is a trap potential, which means that $V(\mathbf{x}) \to \infty$ as $|\mathbf{x}| \to \infty$. By multiplying $\dot{\mathbf{x}}$ both sides, we obtain

$$\frac{dE}{dt} = -\gamma|\dot{\mathbf{x}}|^2 \le 0$$

Here,

$$E(t) := \frac{1}{2}|\dot{\mathbf{x}}|^2 + V(\mathbf{x})$$

is the energy. The term $\gamma|\dot{\mathbf{x}}|^2$ is called the energy dissipation rate. We integrate the above equation from 0 to t, drop the dissipation term to get

$$E(t) \le E(0), \text{ for all } t > 0.$$

This gives a priori estimate of solution

$$\frac{1}{2}|\dot{\mathbf{x}}(t)|^2 + V(\mathbf{x}(t)) \le E(0).$$

This implies both $\dot{\mathbf{x}}(t)$ and $\mathbf{x}(t)$ are bounded, because of the property of V.

Homeworks

1. Suppose $\eta(\cdot)$ satisfies
$$\eta' \le a\eta + b\eta^2, \quad \eta \ge 0,$$
where a, b are two positive constants. Show that $\eta(t)$ is bounded for $t \ge 0$ if $\eta(0)$ is small enough.

2. Consider the equation
$$\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{B}(\mathbf{y}, \mathbf{y}), \quad \mathbf{y}(0) = \mathbf{y}_0.$$
Here, $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{A}$ is an $n \times n$ matrix, $\mathbf{B} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is a bilinear function. Show that the solution $\mathbf{y}(t)$ exists for all $t \ge 0$ if $|\mathbf{y}(0)|$ is small enough.

8.2 Supplementary

8.2.1 Uniform continuity

Pointwise continuity. The concept of continuity is a local concept. Namely, $\mathbf{y}$ is continuous at t_0 means that for any $\epsilon > 0$ there exists $\delta > 0$ such that $|\mathbf{y}(t) - \mathbf{y}(t_0)| < \epsilon$ as $|t - t_0| < \delta$. The continuity property of $\mathbf{y}$ at t_0 is measured by the relation $\delta(\epsilon)$. The locality here means that δ also depends on t_0. This can be read by the example $y = 1/t$ for $t_0 \sim 0$. For any ϵ, in order to have $|1/t - 1/t_0| < \epsilon$, we can choose $\delta \approx \epsilon t_0^2$ (Check by yourself). Thus, the *continuity property* of $y(t)$ for t_0 near 0 and 1 is different. The ratio ϵ/δ is of the same magnitude of $\mathbf{y}'(t_0)$, in the case when $\mathbf{y}(\cdot)$ is differentiable.

Uniform continuity

Theorem 8.7. *When a function* $\mathbf{y}$ *is continuous on a* bounded closed *interval* I*, the above* local *continuity becomes* uniform*. Namely, for any* $\epsilon > 0$*, there exists a* $\delta > 0$ *such that* $|\mathbf{y}(t_1) - \mathbf{y}(t_2)| < \epsilon$ *whenever* $|t_1 - t_2| < \delta$.

Proof. For any $\epsilon > 0$, any $s \in I$, there exists $\delta(\epsilon, s) > 0$ such that $|\mathbf{y}(t) - \mathbf{y}(s)| < \epsilon$ whenever $|t - s| < \delta(\epsilon, s)$. Let us consider the open intervals $U(s, \delta(\epsilon, s)) := (s - \delta(\epsilon, s), s + \delta(\epsilon, s))$. The union $\cup_{s\in I} U(s, \delta(\epsilon, s))$ contain I. Since I is closed and bounded, by so called the finite covering lemma, there exist finite many $U(s_i, \delta(\epsilon, s_i)), i = 1, ..., n$ such that $I \subset \cup_{i=1}^n U(s_i, \delta(\epsilon, s_i))$. Then we choose
$$\delta := \min_{i=1}^{n} \delta(\epsilon, s_i)$$
then the distances between any pair s_i and s_j must be less than δ. For any $t_1, t_2 \in I$ with $|t_1 - t_2| < \delta$, Suppose $t_1 \in U(s_k, \delta(\epsilon, s_k))$ and $t_2 \in U(s_l, \delta(\epsilon, s_l))$, then we must have $|s_k - s_l| < \delta$.
$$|\mathbf{y}(t_1) - \mathbf{y}(t_2)| \le |\mathbf{y}(t_1) - \mathbf{y}(s_k)| + |\mathbf{y}(s_k) - \mathbf{y}(s_l)| + |\mathbf{y}(s_l) - \mathbf{y}(t_2)| < 3\epsilon.$$
This completes the proof. □

The key of the proof is the finite covering lemma. It says that a local property can be uniform through out the whole interval I. This is a key step from local to global.

8.2.2 $C(I)$ is a normed linear space

If this distance is zero, it implies $\mathbf{y}_1 \equiv \mathbf{y}_2$ in I. Also,

$$\|a\mathbf{y}\| = |a|\|\mathbf{y}\|$$

for any scalar a. Moreover, we have

$$\|\mathbf{y}_1 + \mathbf{y}_2\| \le \|\mathbf{y}_1\| + \|\mathbf{y}_2\|.$$

If we replace $\mathbf{y}_2$ by $-\mathbf{y}_2$, it says that the distance between the two functions is less than $\|\mathbf{y}_1\|$ and $\|\mathbf{y}_2\|$. This is exactly the triangular inequality. To show this inequality, we notice that

$$|\mathbf{y}_1(t)| \le \|\mathbf{y}_1\|, \ |\mathbf{y}_2(t)| \le \|\mathbf{y}_2\|, \text{ for all } t \in I$$

Hence,

$$|\mathbf{y}_1(t) + \mathbf{y}_2(t)| \le |\mathbf{y}_1(t)| + |\mathbf{y}_2(t)| \le \|\mathbf{y}_1\| + \|\mathbf{y}_2\|.$$

By taking maximal value on the left-hand side for $t \in I$, we obtain

$$\|\mathbf{y}_1 + \mathbf{y}_2\| \le \|\mathbf{y}_1\| + \|\mathbf{y}_2\|.$$

The function space $C(I)$ with the norm $\|\cdot\|$ is an example of normed linear space.

8.2.3 $C(I)$ is a complete space

A complete normed linear space is called a Banach space.

Definition 8.2. *A sequence $\{\mathbf{y}^n\}$ is called a Cauchy sequence if for any $\epsilon > 0$, there exists an N such that for any $m, n \ge N$, we have*

$$\|\mathbf{y}^n - \mathbf{y}^m\| < \epsilon.$$

Theorem 8.8. *Let $\{\mathbf{y}^n\}$ be a Cauchy sequence in $C(I)$. Then there exist $\mathbf{y} \in C(I)$ such that*

$$\|\mathbf{y}^n - \mathbf{y}\| \to 0 \text{ as } n \to \infty.$$

To prove this theorem, we notice that for each $t \in I$, $\{\mathbf{y}^n(t)\}$ is a Cauchy sequence in $\mathbb{R}$. Hence, the limit $\lim_{n\to\infty} \mathbf{y}^n(t)$ exists. We define

$$\mathbf{y}(t) = \lim_{n\to\infty} \mathbf{y}^n(t) \text{ for each } t \in I.$$

We need to show that $\mathbf{y}$ is continuous and $\|\mathbf{y}^n - \mathbf{y}\| \to 0$. To see $\mathbf{y}$ is continuous, let $t_1, t_2 \in I$. At these two points, $\lim_n \mathbf{y}^n(t_i) = \mathbf{y}(t_i)$, $i = 1, 2$. This means that for any $\epsilon > 0$, there exists an $N > 0$ such that

$$|\mathbf{y}^n(t_i) - \mathbf{y}(t_i)| < \epsilon, \ i = 1, 2, \text{ for all } n \ge N.$$

With this, we can estimate $|\mathbf{y}(t_1) - \mathbf{y}(t_2)|$ through the help of $\mathbf{y}^n$ with $n \geq N$. Namely,

$$\begin{aligned} |\mathbf{y}(t_1) - \mathbf{y}(t_2)| &\leq |\mathbf{y}(t_1) - \mathbf{y}^n(t_1)| + |\mathbf{y}^n(t_1) - \mathbf{y}^n(t_2)| + |\mathbf{y}^n(t_2) - \mathbf{y}(t_2)| \\ &\leq 2\epsilon + |\mathbf{y}^n(t_1) - \mathbf{y}^n(t_2)| \leq 3\epsilon \end{aligned}$$

In the last step, we have used the uniform continuity of $\mathbf{y}^n$ on I. Hence, $\mathbf{y}$ is continuous in I.

Also, from the Cauchy property of $\mathbf{y}^n$ in $C(I)$, we have for any $\epsilon > 0$, there exists an $N > 0$ such that for all $n, m > N$, we have

$$\|\mathbf{y}^n - \mathbf{y}^m\| < \epsilon$$

But this implies that for all $t \in I$, we have

$$|\mathbf{y}^n(t) - \mathbf{y}^m(t)| < \epsilon$$

Now, we fix n and let $m \to \infty$. This yields

$$|\mathbf{y}^n(t) - \mathbf{y}(t)| \leq \epsilon$$

and this holds for $n > N$. Now we take maximum in $t \in I$. This yields

$$\|\mathbf{y}^n - \mathbf{y}\| \leq \epsilon$$

Thus, we have shown $\lim \mathbf{y}^n = \mathbf{y}$ in $C(I)$.

Chapter 9

Stability Analysis

9.1 Introduction

In dynamical systems, there are important special solutions such as equilibria, periodic solutions, etc. It is important to understand their stability. In this chapter, we shall discuss the stability of equilibria and periodic solutions. The theory consists of local theory and global theory. It includes

- Local stability theory of equilibria
- Global stability theory – Lyapunov theory
- Stability of periodic orbits on the plane.

We shall introduce definitions and examples to guide us to develop theory.

Definition 9.1. *An equilibrium $\bar{\mathbf{y}}$ of the ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ is said to be stable if for any $\epsilon > 0$, there exists a $\delta > 0$ such that for any solution $\mathbf{y}(\cdot)$ with $|\mathbf{y}(0) - \bar{\mathbf{y}}| < \delta$, we have $|\mathbf{y}(t) - \bar{\mathbf{y}}| < \varepsilon$.*

Definition 9.2. *An equilibrium $\bar{\mathbf{y}}$ of the ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ is said to be* asymptotically stable *if it is stable, and in addition, there exists a $\delta > 0$ such that any solution $\mathbf{y}(\cdot)$ with $|\mathbf{y}(0) - \bar{\mathbf{y}}| < \delta$ satisfies $\mathbf{y}(t) \to \bar{\mathbf{y}}$ as $t \to \infty$.*

Definition 9.3. *An equilibrium $\bar{\mathbf{y}}$ of the ODE $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ is said to be* exponentially stable *if there exist an $\alpha > 0$ and a $\delta > 0$ such that any solution $\mathbf{y}(\cdot)$ with $|\mathbf{y}(0) - \bar{\mathbf{y}}| < \delta$ satisfies $\mathbf{y}(t) - \bar{\mathbf{y}} = O(e^{-\alpha t})$ as $t \to \infty$.*

Examples.

- For linear systems, centers are stable whereas sinks and spiral sinks are asymptotically stable.
- For hamiltonian system, the minimum of a hamiltonian H is a stable center. The saddles are unstable.
- For gradient systems, the sinks are stable while the sources are unstable.

We shall discuss dissipative systems below.

9.2 Damping systems

In this section, we consider dissipative nonlinear oscillators. The dissipation is due to friction. The friction force is usually a function of velocity. Let us call it $b(\dot{y})$. In general, we consider

$$\ddot{y} = -V'(y) + b(\dot{y}), \tag{9.1}$$

where 0 is the minimum of V, having the lowest potential. The friction force has the property:

$$b(\dot{y}) \cdot \dot{y} < 0 \text{ and } b(0) = 0.$$

This means that the direction of the frictional force is in the opposite direction of the velocity and the friction force is zero if the particle is at rest. Here are some concrete examples of damping.

- Simple pendulum with damping The equation for simple pendulum is

  $$ml\ddot{\theta} = -mg\sin\theta.$$

  A simple damping force is protortional to the angular speed $\beta\dot{\theta}$, provided the damping comes from the friction at the fixed point. Here, $\beta > 0$. Thus the model for simple pendulum with friction reads

  $$ml\ddot{\theta} = -\beta\dot{\theta} - mg\sin\theta. \tag{9.2}$$

- An active shock absorber In the mass-spring model, the friction force may depend on the velocity nonlinearly, say $\beta(v)$, say $\beta(v) = v^4$. Then the corresponding oscillation is nonlinear:

  $$m\ddot{y} = -\beta(\dot{y}) - ky,\ \beta(v) = v^4, \tag{9.3}$$

- V is a general satisfying $V(y) \to \infty$ as $|y| \to \infty$.

We can rewrite this damped oscillation system in the following form

$$\begin{cases} \dot{x} &= p \\ \dot{p} &= -V'(x) + b(p) \end{cases}$$

Let us define the Hamilton (or the energy) $H(x,p) = 1/2p^2 + V(x)$. You can check that along any trajectory $(x(t), v(t))$,

$$\frac{d}{dt}(H(x(t),p(t)) = H_x\dot{x} + H_p\dot{p} = -V'(x)p + p(-V'(x) - b) = bp < 0.$$

The Hamiltonian $H(x(t),p(t))$ decreases along any trajectories until $p = 0$. Such a perturbation is called a dissipative perturbation. As a result, we can see that $(0,0)$ becomes asymptotic stable. Indeed, we shall show in the section of Liapunov function that $(x(t),p(t)) \to (0,0)$ as $t \to \infty$ for any trajectories. Here, we just show that, from the linear analysis, the center becomes a spiral sink

for a Hamiltonian system with dissipative perturbation. We shall assume $b'(0) \neq 0$. The variational matrix at $(0,0)$ is

$$\begin{pmatrix} 0 & 1 \\ -H_{xx} & -H_{xp} + b'(0) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -V''(0) & -b'(0) \end{pmatrix}$$

Its eigenvalues are

$$\lambda_{\pm} = b'(0) \pm i\sqrt{V''(0)}$$

Now, the force $b(p)$ is a friction which means that $b(p)p < 0$. But $b(0) = 0$. We get that

$$0 > b(p)p \sim b'(0)p \cdot p$$

Thus, if $b'(0) \neq 0$, then $b'(0) < 0$. Hence, $(0,0)$ becomes a spiral sink.

Let us study the stability of the following damped system:

$$\ddot{y} = -V'(y) + b(\dot{y}), \quad b(\dot{y}) \cdot \dot{y} < 0. \tag{9.4}$$

The following theorem give a sufficient condition for global stability of the equilibrium.

Theorem 9.1. *Consider the system (9.4). Suppose $V(y) \to \infty$ as $|y| \to \infty$ and $V(y)$ has only one minimum $\bar{y}$. Then any solution y satisfies*

$$y(t) \to \bar{y} \text{ and } \dot{y}(t) \to 0 \text{ as } t \to \infty.$$

Proof. Without loss of generality, we may also assume $\bar{y} = 0$ and $V(0) = 0$. Otherwise, we may just replace y by $y - \bar{y}$ and $V(y)$ by $V(y) - V(\bar{y})$, which does not alter $F(y)$ in the original problem.

We use energy method: multiplying $\dot{y}$ on both sides of (9.4), we obtain

$$\dot{y}\ddot{y} = -V'(y)\dot{y} - \dot{y}\dot{y}$$

$$\frac{dE}{dt} = -\dot{y}^2, \tag{9.5}$$

where

$$E(y, \dot{y}) := \frac{\dot{y}^2}{2} + V(y). \tag{9.6}$$

The strategy is to prove (i) $E(t) \to 0$, and (ii) $E(y, \dot{y}) = 0$ if and only if $(y, \dot{y}) = (0,0)$, and (iii) $(y(t), \dot{y}(t)) \to (0,0)$. We divide the proof into the following steps.

Step 1. From (9.5), $E(t) := E(y(t), \dot{y}(t))$ is a decreasing function along any trajectort $(y(t), \dot{y}(t))$. Further, it has lower bound, namely, $E(y, \dot{y}) \geq 0$. we get $E(t) \searrow \alpha$ as $t \to \infty$ for some number α.

Step 2. Let us call the limiting set of $(y(t), \dot{y}(t))$ by Ω^+. That is

$$\Omega^+ = \{(y, \dot{y}) | \exists t_n, t_n \to \infty \; s.t. \; (y(t_n), \dot{y}(t_n)) \to (y, \dot{y})\}.$$

Such a set is called an ω-limit set. We claim that any trajectory $(\tilde{y}(\cdot), \dot{\tilde{y}}(\cdot))$ with initial data $(\tilde{y}(0), \dot{\tilde{y}}(0)) \in \Omega^+$ lies on Ω^+ forever. The proof of this claim relies on the continuity theorem on the initial data. Namely, the solution of an ODE depends on its initial data continuously. Let us accept this fact. Suppose $(\tilde{y}(0), \dot{\tilde{y}}(0)) \in \Omega^+$, we want to prove that for any fixed $s > 0$, $(\tilde{y}(s), \dot{\tilde{y}}(s)) \in \Omega^+$. Given fixed $s > 0$, by the continuous dependence of initial data, we have for any $\epsilon > 0$, there exists a $\delta > 0$ such that if $|(y_1, \dot{y}_1) - (\tilde{y}(0), \dot{\tilde{y}}(0))| < \delta$, then the solution $y(\cdot)$ with initial data $(y_1, \dot{y}_1)$ is in an ϵ neighborhood of $\tilde{y}(s)$. Now, since $(\tilde{y}(0), \dot{\tilde{y}}(0)) \in \Omega^+$, with this $\delta > 0$, there exist t_n such that $|(y(t_n), \dot{y}(t_n)) - (\tilde{y}(0), \dot{\tilde{y}}(0))| < \delta$. Let us consider two solutions, one has initial data $(y(t_n), \dot{y}(t_n))$, the other has initial data $(\tilde{y}(0), \dot{\tilde{y}}(0))$. By the continuity dependence of the initial data, we get $(y(t_n + s), \dot{y}(t_n + s)) - (\tilde{y}(s), \dot{\tilde{y}}(s))| < \epsilon$. This yields that $\forall \epsilon > 0$, there exists an n such that $|(y(t_n + s), \dot{y}(t_n + s)) - (\tilde{y}(s), \dot{\tilde{y}}(s))| < \epsilon$. Thus, $(\tilde{y}(s), \dot{\tilde{y}}(s)) \in \Omega^+$.

Step 3. We claim that, for any $(\tilde{y}(\cdot), \dot{\tilde{y}}(\cdot))$ in Ω^+, the corresponding energy $E(\tilde{y}(s), \dot{\tilde{y}}(s)) = \alpha$ for any $s \geq 0$. This is because (1) for any fixed s, there exist $t_n \to \infty$ such that $(y(t_n + s), \dot{y}(t_n + s)) \to (\tilde{y}(s), \dot{\tilde{y}}(s))$ as $n \to \infty$, and (2), from step 1, $E(y(t), \dot{y}(t)) \searrow \alpha$ as $t \to \infty$. Thus, we get $E(y(t_n + s), \dot{y}(t_n + s)) \to \alpha$ for any s. This implies

$$\frac{d}{ds} E(\tilde{y}(s), \dot{\tilde{y}}(s)) = 0.$$

On the other hand, $\frac{d}{ds} E(\tilde{y}(s), \dot{\tilde{y}}(s)) = -\dot{\tilde{y}}^2(s)$. Hence, we get $\dot{\tilde{y}}(s) \equiv 0$. This again implies $\tilde{y}(s) \equiv \hat{y}$ for some constant $\hat{y}$. Thus, $(\hat{y}, 0)$ is an equilibrium state of the damping oscillation system (9.4). However, the only equilibrium state for (9.4) is $(0, 0)$ because V has a unique minimum and thus the only zero of $F := -V'$ is 0. This implies

$$E(\tilde{y}(s), \dot{\tilde{y}}(s)) = \alpha = 0.$$

We conclude that

$$E(y(t), \dot{y}(t)) \to \alpha = 0 \text{ as } t \to \infty.$$

Step 4. From step 3,

$$E(y(t), \dot{y}(t)) = \frac{1}{2} \dot{y}(t)^2 + V(y(t)) \to 0 \text{ as } t \to \infty.$$

and $V(y) \geq 0$, we get

$$\dot{y}(t) \to 0 \text{ and } V(y(t)) \to 0, \text{ as } t \to \infty.$$

Since 0 is the unique minimum of V, we get that $V(y) \to 0$ forces $y \to 0$. □

The above method to show global stability is called the Lyapunov method. The energy function E above is called a Lyapunov function. Thus, *the effect of damping (dissipation) is a loss of energy*.

In the active shock absorber:

$$m\ddot{y} = -\beta(\dot{y})\dot{y} - ky,\ \beta(v) = v^4,$$

the equilibrium state is $(0, 0)$. From Lyapunov method, we see that this equilibrium is globally stable.

For the simple pendulum, we see that $V(\theta) = -g/l \cos\theta$ has infinite many minima: $\theta = 2n\pi$. The function $E(y, \dot{y})$ has local minima $(2n\pi, 0)$. The local minimum $(2n\pi, 0)$ sits inside the *basin*

$$B_n = \{(y, \dot{y}) \,|\, E(y, \dot{y}) < g/l\}.$$

The equilibrium $(2n\pi, 0)$ is the only minimum of E in the basin B_n. Suppose a solution starts from a state $(y(0), \dot{y}(0)) \in B_n$, then by using the Lyapunov method, we see that $(y(t), \dot{y}(t)) \to (2n\pi, 0)$ as $t \to \infty$.

What will happen if $E(0) \geq g/l$ initially? From the loss of energy we have $E(t)$ will eventually go below g/l. Thus, the trajectory will fall into some basin B_n for some n and finally goes to $(2n\pi, 0)$ as $t \to \infty$.

Homeworks

1. Plot the phase portrait for the damped simple pendulum (9.2).

2. Consider a simple pendulum of length l with mass m at one end and the other end is attached to a vibrator. The motion of the vibrator is given by $(x_0(t), y_0(t)$. Let the angle of the pendulum to the verticle axis (in counterclockwise direction) is $\theta(t)$.

   (a) Show that the position of the mass m at time t is $(x(t), y(t)) = (x_0(t)+l\sin\theta(t), y_0(t)-\cos\theta(t))$.

   (b) Find the velocity and acceleration of m.

   (c) Suppose the mass is in the uniform gravitational field $(0, -mg)$. Use the Newton's law to derive the equation of motion of m.

   (d) Suppose $(x_0(t), y_0(t))$ is given by $(0, \alpha\sin(\omega_0 t))$. Can you solve this equation?

3. B-D, pp. 502: 22

9.3 Local stability

Theorem 9.2. *Consider the nonlinear equation*

$$\mathbf{y}' = \mathbf{f}(\mathbf{y})$$

Suppose $\bar{\mathbf{y}}$ is an equilibrium of the nonlinear equation, i.e. $\mathbf{f}(\bar{\mathbf{y}}) = \mathbf{0}$. If $\bar{\mathbf{y}}$ is an exponentially stable equilibrium for the linearized ODE:

$$\mathbf{y}' = \mathbf{f}'(\bar{\mathbf{y}})(\mathbf{y} - \bar{\mathbf{y}}),$$

that is

$$Re(\lambda(\mathbf{f}'(\bar{\mathbf{y}}))) < 0.$$

then $\bar{\mathbf{y}}$ is also an exponentially stable equilibrium of the nonlinear squation.

Proof. 1. We start from

$$\mathbf{u}' = \mathbf{A}\mathbf{u} + \mathbf{g}(\mathbf{u}),$$

We want to show that $\mathbf{u}(t) \to 0$ at rate $O(e^{-\alpha' t})$ for any $0 < \alpha' < \alpha$. Let us define $\mathbf{v}(t) = e^{\alpha' t}\mathbf{u}(t)$. We want to show that $\mathbf{v}(t)$ remains bounded for all $t \geq 0$. To show this a priori estimate, let T be any positive number, assume that $\mathbf{v}(t)$ exists on $[0, T]$. Let us define

$$M(T) := \max_{t\in[0,T]} |\mathbf{v}(t)|$$

We want to show that $M(T)$ remains bounded by a constant induependent of T. If so, then we can always extend $\mathbf{v}$ beyond T (so does $\mathbf{u}$), then the solution $\mathbf{v}(t)$ exists for all $t \geq 0$ and remains bounded. The boundedness of $\mathbf{v}$ gives the exponential convergence of $\mathbf{u}(\cdot)$.

2. Let us denote $\mathbf{A} + \alpha'\mathbf{I}$ by $\mathbf{A}'$. The function $\mathbf{v}$ satisfies

$$\begin{aligned}\dot{\mathbf{v}} &= (\mathbf{A} + \alpha'\mathbf{I})\mathbf{v} + e^{\alpha' t}\mathbf{g}(e^{-\alpha t}\mathbf{v}) \\ &= \mathbf{A}'\mathbf{v} + e^{(\alpha-2\alpha)t}O(|\mathbf{v}|^2)\end{aligned}$$

The eigenvalues of $\mathbf{A}'$ satisfy

$$Re(\lambda(\mathbf{A}')) = Re(\lambda(A)) + \alpha' < 0.$$

We write this perturbed equation in integral form:

$$\mathbf{v}(t) = e^{\mathbf{A}'t}\mathbf{v}(0) + \int_0^t e^{\mathbf{A}'(t-s)}e^{(\alpha'-2\alpha)s}O(|\mathbf{v}(s)|^2)\,ds$$

Taking maximal on the right-hand side, we get

$$\begin{aligned}|\mathbf{v}(t)| &\leq |\mathbf{v}(0)| + \int_0^t e^{-\epsilon(t-s)}e^{-\alpha s}M(T)^2\,ds \\ &\leq |\mathbf{v}(0)| + \frac{C}{\epsilon}M(T)^2.\end{aligned}$$

Here, we have used $O(|\mathbf{v}|^2) \leq C|\mathbf{v}|^2$ for $\mathbf{v}$ in a bounded set. Taking maximum in t on the left hand side, we get

$$M(T) \leq |\mathbf{v}(0)| + \frac{C}{\epsilon}M(T)^2.$$

Thus, there exists a $\delta > 0$ such that if $|\mathbf{v}(0)| \leq \delta$, then the above inequality always holds. This completes the proof.

□

9.4 Lyapunov function

We recall that when the perturbation of a hamiltonian system is dissipative, we observe that the hamiltonian H decreases along any trajectory and eventually reaches a minimum of H. If there is only one minimum of H, then this minimum must be globally asymptotically stable. That is, every trajectory tends to this minimum as $t \to \infty$. So, the key idea here is that the globally asymptotic stability of an equilibrium is resulted from the the decreasing of H. This idea can be generalized to general systems. The dissipation is measured by so called the Liapunov function Φ, which decreases along trajectories. More precisely, let consider the general system

$$\begin{cases} \dot{x} &= f(x,y) \\ \dot{y} &= g(x,y) \end{cases} \tag{9.7}$$

Suppose $(0,0)$ is an equilibrium of this system. We have the following definition.

Definition 9.4. *A C^1-function $\Phi(x,y)$ is called a Liapunov function for (9.7) if*

(i) $\Phi(0,0) = 0$, $\Phi(x,y) > 0$ *for* $(x,y) \neq (0,0)$.

(ii) $\Phi(x,y) \to \infty$ *as* $|(x,y)| \to \infty$.

(iii) $\dot{\Phi} := \Phi_x(x,y)f(x,y) + \Phi_y(x,y)g(x,y) < 0$ *for* $(x,y) \neq (0,0)$.

Remark

- Condition (i) says that $(0,0)$ is the only isolated minimum of Φ.
- Condition (ii) says that the region $\Phi(x,y) \leq E$ is always bounded.
- Condition (iii) implies that along any trajectory

$$\frac{d\Phi(x(t),y(t))}{dt} < 0, \tag{9.8}$$

  unless it reaches the equilibrium $(0,0)$. Thus, $\Phi(x(t),y(t))$ is a decreasing function.

Theorem 9.3. *Consider the system (9.7). Suppose $(0,0)$ is its equilibrium. Suppose the system possesses a Liapunov function Φ, then $(0,0)$ is globally and asymptotically stable. That is, for any trajectory, we have*

$$\lim_{t\to\infty}(x(t),y(t)) = (0,0).$$

Proof. We shall use the extremal value theorem to prove this theorem. The extremal value theorem states that
a continuous function in a bounded and closed domain in $\mathbb{R}^n$ attains its extremal value.
Along any trajectory $(x(t),y(t))$, we have that $\Phi(x(t),y(t))$ is decreasing (condition (iii)) and bounded below (condition (i)). Hence it has a limit as t tends to infinity. Suppose $\lim_{t\to\infty}\Phi(x(t),y(t)) = m > 0$. Then the orbit $(x(t),y(t)), t \in (0,\infty)$ is confined in the region $S := \{(x,y)|m \leq \Phi(x,y) \leq \Phi(x(0),y(0))\}$. From condition (ii), this region is bounded and closed. Hence $d\Phi(x(t),y(t))/dt$

can attain a maximum in this region (by the extremal value theorem). Let us call it α. From (9.8), we have $\alpha < 0$. But this implies

$$\Phi(x(t), y(t)) = \int_0^t \frac{d\Phi(x(t), y(t))}{dt}\, dt \leq \alpha t \to -\infty \text{ as } t \to \infty.$$

This is a contradiction. Hence $\lim_{t\to\infty} \Phi(x(t), y(t)) = 0$.

Next, we show $(x(t), y(t)) \to (0, 0)$ as $t \to \infty$. Let $\rho(t) = x(t)^2 + y(t)^2$. Suppose $\rho(t)$ does not tend to 0. This means that there exists a sequence t_n with $t_n \to \infty$ such that $\rho(t_n) \geq \rho_0 > 0$. Then the region

$$R := \{(x, y) | x^2 + y^2 \geq \rho_0 \text{ and } \Phi(x, y) \leq \Phi(x(0), y(0))\}$$

is bounded and closed. Hence, by the extremal value theorem again that Φ attains a minimum in this region. Since $\Phi > 0$ in this region, we have

$$\min_R \Phi(x, y) \geq \beta > 0.$$

and because $(x(t_n), y(t_n)) \in R$, we obtain

$$\min_{t_n} \Phi(x(t_n), y(t_n)) \geq \beta > 0.$$

This contradicts to $\lim_{t\to\infty} \Phi(x(t), y(t)) = 0$. Hence, $x^2(t) + y^2(t) \to 0$ as $t \to \infty$. Thus, we obtain that the global minimum $(0, 0)$ is asymptotically stable. □

If the Lyapunov function Φ satisfies additional conditions:

(iv) The condition (iii) is replaced by $\dot{\Phi}(x, y) \leq -\alpha\Phi(x, y)$ *for some positive constant* α *for all* (x, y)*,*

(v) $\Phi \in C^2$ *and* $\Phi(x, y) \geq c(x^2 + y^2)$ *in a neighborhood of* $(0, 0)$*.*

Theorem 9.4. *Under the assumptions (i),(ii),(iv), (v), the state* $(0, 0)$ *is asymptotically stable and any solution* $|(x(t), y(t))| = O(e^{-\alpha t})$ *as* $t \to \infty$*.*

If the Lyapunov only satisfies the following weaker condition, then we can only have stability result, not asymptotic stability.

Definition 9.5. *A* C^1*-function* $\Phi(x, y)$ *is called a (weak) Liapunov function for (9.7) if*

(i) $\Phi(0, 0) = 0$*,* $\Phi(x, y) \geq 0$ *for* $(x, y) \neq (0, 0)$*.*

(ii) $\Phi(x, y) \to \infty$ *as* $|(x, y)| \to \infty$*.*

(iii) $\Phi_x(x, y)f(x, y) + \Phi_y(x, y)g(x, y) \leq 0$ *for* $(x, y) \neq (0, 0)$*.*

Theorem 9.5. *Consider the system (9.7). Suppose* $(0, 0)$ *is its equilibrium. Suppose the system possesses a weak Liapunov function* Φ*, then* $(0, 0)$ *is stable.*

Example. Damped simple pendulum.

$$\ddot{\theta} = \frac{g}{l}\sin\theta - b\dot{\theta}$$

Here, $b > 0$ is the damping coefficient. In the form of first order equation, it reads

$$\begin{cases} \dot{x} &= y \\ \dot{y} &= \frac{g}{l}\sin x - by \end{cases}$$

We take

$$\Phi(x,y) = \frac{1}{2}y^2 + \frac{g}{l}(1-\cos x).$$

Then

$$\dot{\Phi} := \Phi_x f + \Phi_y g = \frac{g}{l}\sin(x)y + y(\frac{g}{l}\sin x - by) = -by^2 < 0.$$

We see that $\dot{\Phi}(x,y) = 0$ if and only if $y = 0$. This is weaker than $\dot{\Phi}(x,y) = 0$ if and only if $(x,y) = (0,0)$. So, it only satisfies condition of the *weak* Lyapunov function. Thus, we can only get a stability result, not asymptotic stability result. However, suppose we consider the linear problem, say the spring-mass system with a linear damper. We know the solutions decay to $(0,0)$ state exponentially fast from explicit solution formula. Such result cannot be obtained via the weak Lyapunov function. There are two ways to solve this. One is we modify the Lyapunov function. The other is we provide another linear stability theory based on perturbation theory.

*** Lyapunov function for Linear Stable System** Consider the linear system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}.$$

Suppose $\mathbf{A}$ is a stable matrix. That is, $Re(\lambda(\mathbf{A})) \le -\alpha$ for some $\alpha > 0$. We want to construct a Lyapunov function of the form $\Phi(\mathbf{x}) = \mathbf{x}^T\mathbf{P}\mathbf{x}$ such that (i) $\mathbf{P} > 0$ and (ii) $\dot{\Phi} < 0$. We have

$$\dot{\Phi} = \dot{\mathbf{x}}^T\mathbf{P}\mathbf{x} + \mathbf{x}^T\mathbf{P}\dot{x} = \mathbf{x}^T\mathbf{A}^T\mathbf{P}\mathbf{x} + \mathbf{x}^T\mathbf{P}\mathbf{A}\mathbf{x} = \mathbf{x}^T(\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A})\mathbf{x}.$$

This means that

$$\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{Q} < \mathbf{0}.$$

Theorem 9.6. *Suppose $Re(\lambda(\mathbf{A})) \le -\alpha$ for some $\alpha > 0$. For any $\mathbf{Q} > \mathbf{0}$, there exists a $\mathbf{P} > \mathbf{0}$ such that*

$$\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{Q} < \mathbf{0}.$$

Proof. Define

$$\mathbf{P} = \int_0^\infty e^{\mathbf{A}^T t}\mathbf{Q}e^{\mathbf{A}t}\,dt.$$

Then

$$\begin{aligned}\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A} &= \int_0^\infty \mathbf{A}^T e^{\mathbf{A}^T t}\mathbf{Q}e^{\mathbf{A}t} + e^{\mathbf{A}^T t}\mathbf{Q}e^{\mathbf{A}t}\mathbf{A}\,dt \\ &= \int_0^\infty \frac{d}{dt}\left(e^{\mathbf{A}^T t}\mathbf{Q}e^{\mathbf{A}t}\right)\,dt = -\mathbf{Q}.\end{aligned}$$

□

Remark. We claim that when $Re(\lambda(\mathbf{A})) \leq -\alpha$, then there exists a $\mathbf{Q}$ such that

$$\mathbf{x}^T\mathbf{Q}\mathbf{x} \geq \alpha \mathbf{x}^T\mathbf{P}\mathbf{x}.$$

This is equivalent to

$$\dot{\Phi} \leq -\alpha\Phi,$$

and leads to

$$\Phi(\mathbf{x}(t)) \leq \Phi(\mathbf{x}(0))e^{-\alpha y}.$$

This gives exponential convergence of all solutions to $(0, 0)$. We leave its proof as an exercise.

Homework For the damped spring-mass system

$$m\ddot{y} + \gamma\dot{y} + ky = 0.$$

Find a Lyapunov function Φ that gives exponential convergence result.

Project Study the damper of Taipei 101.

9.5 Poincaré-Bendixson Theorem

We still focus on two-dimensional systems

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}), \mathbf{y}(0) = \mathbf{y}_0 \tag{9.9}$$

where $\mathbf{y} \in \mathbb{R}^2$. As we mentioned, our goal is to characteristized the whole orbital structure. We have seen the basic solutions are the equilibria. The second class are the orbits connecting these equilibria. In particular, we introduce the separatrices and the homoclinic orbits. We have seen in the damped pendulum that solutions enclosed in separatrices go to a sink time asymptotically. In this section, we shall see the case that the solution may go to an periodic solution. In other words, the solution goes to another separatrix. The van de Pol oscillator and the predator-prey system are two important examples.

Basic notions of dynamical systems We first introduce some basic notions.

- **positive orbits and negative orbits**. Let us denote by $\phi(t, \mathbf{y}_0)$ the solution to the problem (9.9). The orbit $\gamma^+(\mathbf{y}) = \{\phi(t, \mathbf{y})|t \geq 0\}$ is the positive orbit through $\mathbf{y}$. Similarly, $\gamma^-(\mathbf{y}) = \{\phi(t, \mathbf{y})|t \leq 0\}$ and $\gamma(\mathbf{y}) = \{\phi(t, \mathbf{y})| - \infty < t < \infty\}$ are the negative orbit and the orbit through $\mathbf{y}$.

- **periodic orbits** If $\phi(T, \mathbf{y}) = \mathbf{y}$ and $\phi(t, \mathbf{y}) \neq \mathbf{y}$ for all $0 < t < T$, we say $\{\phi(t, \mathbf{y})|0 \leq t < T\}$ a periodic orbit.

- **ω-limit sets** A point p is called an ω (resp. α) point of $\mathbf{y}$ if there exists a sequence $\{t_n\}$, $t_n \to \infty$ (resp. $-\infty$) such that $p = \lim_{n\to\infty} \phi(t_n, \mathbf{y})$. The collection of all ω (resp. α) limit point of $\mathbf{y}$ is called the ω (resp. α) limit set of $\mathbf{y}$ and is denoted by $\omega(\mathbf{y})$ (resp. $\alpha(\mathbf{y})$). One can show that
$$\omega(\mathbf{y}) = \bigcap_{t\geq 0} \overline{\bigcup_{s\geq t} \phi(s, \mathbf{y})}$$
Thus, $\omega(\mathbf{y})$ represents where the positive $\gamma^+(\mathbf{y})$ ends up.

- **Invariant sets** A set S is called positive (resp. negative) invariant under ϕ if $\phi(t, S) \subset S$ for all $t \geq 0$ (resp. $t \leq 0$). A set S is called invariant if S is both positive invariant and negative invariant. It is easy to see that equilibria and periodic orbits are invariant set. The closure of an invariant set is invariant.

Theorem 9.7. *The sets $\omega(\mathbf{y})$ and $\alpha(\mathbf{y})$ are invariant.*

Proof. The proof is based on the continuous dependence of the initial data. Suppose $p \in \omega$. Thus, there exists $t_n \to \infty$ such that $p = \lim_{n\to\infty} \phi(t_n, \mathbf{y})$. Consider two solutions: $\phi(s, p)$ and $\phi(s + t_n, \mathbf{y}) = \phi(s, \phi(t_n, \mathbf{y}))$, for any $s > 0$. The initial data are closed to each other when n is enough. Thus, by the continuous dependence of the initial data, we get $\phi(s, p)$ is closed to $\phi(s + t_n, \mathbf{y})$. □

Here are some examples of ω-limit sets and periodic solutions.

Example 1 Consider
$$\begin{cases} \dot{x} &= x + y - x(x^2 + y^2) \\ \dot{y} &= -x + y - y(x^2 + y^2). \end{cases} \tag{9.10}$$

1. The state $(0, 0)$ is a spiral source. The $(0, 0)$ state is an equilibrium state. The corresponding linearized equation near $(0, 0)$ is
$$\begin{cases} \dot{x} &= x + y \\ \dot{y} &= -x + y \end{cases}$$
whose characteristic roots are $\lambda = 1 \pm i$. Thus, $(0, 0)$ is a spiral source.

2. We express this equation in polar coordinate: $x = r\cos\theta$, $y = r\sin\theta$. We multiply the first equation by x, the second equation by y, then add, we get
$$x\dot{x} + y\dot{y} = x^2 + y^2 - (x^2 + y^2)^2.$$
That is
$$\dot{r} = r - r^3.$$

3. If we multiply the first equation by y, the second equation by x, then subtract, we get
$$y\dot{x} - x\dot{y} = x^2 + y^2.$$
In polar coordinate, this is
$$\dot{\theta} = -1.$$

4. The solution with initial data $(r(0), \theta(0)) = (r_0, \theta_0)$ is

$$r = \frac{1}{1 + ((1/r_0^2) - 1)e^{-2t}}$$
$$\theta = -t + \theta_0.$$

We see that

- Solutions with $0 < r_0 < 1$ converge to $r = 1$.
- Solutions with $r_0 > 1$ also converge to $r = 1$.
- $r = 1$ is an ω-limit set. It is a periodic solution.

Theorem 9.8 (Poincaré-Bendixson)**.** *If $\gamma^+(\mathbf{y})$ is contained in a bounded closed subset in $\mathbb{R}^2$ and $\omega(\mathbf{y}) \neq \emptyset$ and does not contain any critical points (i.e. where $\mathbf{f}(\mathbf{y}) = 0$), then $\omega(\mathbf{y})$ is a periodic orbit.*

Theorem 9.9. *The Poincaré-Bendixson theorem states that: if $\gamma^+(\mathbf{y})$ remains bounded, then one of the follows must be true*

- *$\omega(\mathbf{y})$ is a periodic orbit,*
- *$\omega(\mathbf{y})$ is a critical point.*
- *$\omega(\mathbf{y})$ consists of one or more critical points joined by solution paths (i.e. homoclinic or heteroclinic orbits).*

In all these cases, the $\omega(\mathbf{y})$-limit set is stable in the sense that $\gamma^+(\mathbf{y}) \to \omega(\mathbf{y})$.

Example 2 Consider the Hamiltonian system

$$\dot{x} = H_y$$
$$\dot{y} = -H_x$$
$$H(x, y) = \frac{1}{2}y^2 - \frac{1}{2}x^2 + \frac{1}{4}x^4.$$

The orbits

$$\{(x, y) | H(x, y) = 0\}$$

consists of a critical critical point $(0, 0)$ and two homoclinic orbits.

Now, consider a perturbation of this system by

$$\dot{x} = H_y - \mu H H_x$$
$$\dot{y} = -H_x - \mu H H_y$$

Multiplying first equation by H_x, second equation by H_y, then add. We get

$$\dot{H} = H_x\dot{x} + H_y\dot{y} = -\mu H(H_x^2 + H_y^2).$$

In the case $\mu > 0$, we see that $H(x(t), y(t)) \to 0$ along any path. Thus, the set $\Omega = \{(x, y) | H(x, y) = 0\}$ is the ω-limit set.

Homework Plot the orbits for this perturbed Hamiltonian systems.

Example 3. van der Pol oscillator Recall the van der Pol equation (3.14) for triode oscillator:

$$L\frac{d^2I}{dt^2} + \mu(I^2 - 1)\frac{dI}{dt} + \frac{I}{C} = 0. \tag{9.11}$$

where $\mu > 0$, I is the current. Let us write it as a 2×2 system

$$\begin{cases} \dot{x} = y \\ \dot{y} = -x + \mu(1 - x^2)y \end{cases} \tag{9.12}$$

We shall show this system has a periodic orbit for any μ.

1. When $\mu = 0$, the orbits are circles. They are periodic orbits. The case $\mu < 0$ can be transformed to the case of $\mu > 0$ by reverting t to $-t$. Thus, we shall only consider the case $\mu > 0$.

2. The state $(0, 0)$ is the only equilibrium of this system. Near $(0, 0)$, the linearized equation is

$$\begin{aligned} \dot{x} &= y \\ \dot{y} &= -x + \mu y. \end{aligned}$$

   The eigenvalues of this linearized system is $\lambda = (\mu \pm \sqrt{\mu^2 - 1})/2$, which has positive real part. Therefore, $(0, 0)$ is unstable. It can not lie in the ω-limit set.

3. We shall construct a Jordan curve (i.e. a simple closed curve) $\mathcal{C}$ on the plane encircle $(0, 0)$ such that no orbit can leave $\mathcal{C}$. Then by the Poincaré-Bendixson theorem, there exists a periodic orbit in the interior of the Jordan curve $\mathcal{C}$. I shall refer the proof to a Note of F. Bonetto, which proof was originally from [Yeh 86]. The idea is that $\mathcal{C}$ is composed of piecewise arcs. Each arc is the orbit of a simple ODE. On which, it is easy to show (by taking the cross product of two vector fields) that the flow of (9.12) goes inward.

Liénard equation The Liénard equation has the form

$$\ddot{x} + f(x)\dot{x} + g(x) = 0,$$

where g is an odd function and f is an even function. It can be changed to a 2 system through the transform

$$x = x, \quad y = \dot{x} + F(x), \quad F(x) = \int_0^x f(x)\,dx.$$

The new system becomes

$$\begin{cases} \dot{x} = y - F(x) \\ \dot{y} = -g(x). \end{cases} \tag{9.13}$$

called the Liénard system. The van der Pol oscillator $\ddot{x} + \mu(x^2 - 1)\dot{x} + x = 0$ is a special case of the Lienard equation. The corresponding Liénard system is

$$\begin{cases} \dot{x} = y - \mu(\frac{1}{3}x^3 - x) \\ \dot{y} = -x. \end{cases}$$

The Liénard equation possess a periodic solution under the following assumptions

- g is odd and F is odd;
- $g(x) > 0$ for $x > 0$;
- $F(x) \to \infty$ as $x \to \infty$;
- $F(x)$ has exactly one positive p, $F(x) < 0$ for $0 < x < p$ and $F(x) > 0$ and monotone for $x > p$.

The proof of such result is left for a project to you to complete.

Homeworks

1. B-D pp. 556, 13, 15, 16, 17.

Chapter 10

Numerical Methods for Ordinary Differential Equations

10.1 Design of numerical schemes

We shall solve the initial value problem

$$y' = f(t, y),\ y(0) = y_0. \tag{10.1}$$

numerically. It is to to approximate the solution $y(\cdot)$ by discrete values $y^n \sim y(t^n)$ at discrete times $t^0 = 0 < t^1 < \cdots t^n]$. For simplicity, we take uniform step size h and define $t^k = kh$. At time t, we expect that as the mesh size $h \to 0$, the discrete value y^n tends to $y(t)$, where $nh = t$.

A numerical scheme is to produce the discrete values y^n from the initial data y_0. It is usually designed as an iterative procedure. Namely, given y^n, we want to find y^{n+1} which is an approximation of $y(t^{n+1})$. Such design procedure can be based on approximation of integration, or on Taylor expansion. Let us explain below.

Integral Approximation Approach By integrating the ODE from t^n to t^{n+1}, we get

$$y(t^{n+1}) = y(t^n) + \int_{t^n}^{t^{n+1}} f(t, y(t))\, dt$$

So the strategy is to approximate the integral by a numerical integral

$$hF_h(t^n, y^n) \approx \int_{t^n}^{t^{n+1}} f(t, y(t))\, dt.$$

Below, we give several popular schemes

- Forward Euler method

$$y^{n+1} = y^n + hf(t^n, y^n).$$

- Backward Euler method,
$$y^{n+1} = y^n + hf(t^{n+1}, y^{n+1}).$$

- Runge-Kutta method, based on mid point rule
$$y^{n+1} = y^n + hf(t^{n+1/2}, y^n + \frac{h}{2}f(t^n, y^n))$$
Here, we approximate
$$\int_{t^n}^{t^{n+1}} f(t, y(t))\, dt \approx hf(t^{n+1/2}, y(t^{n+1/2}))$$
then approximate
$$y(t^{n+1/2}) \approx y(t^n) + \frac{h}{2}f(t^n, y(t^n))$$

- Second-order Runge-Kutta method (RK2): based on trapezoidal rule
$$\begin{aligned} y_1 &= y^n + hf(t^n, y^n), \\ y^{n+1} &= y^n + \frac{1}{2}h(f(t^n, y^n) + f(t^{n+1}, y_1)) \\ &= \frac{1}{2}(y_1 + (y^n + hf(t^{n+1}, y_1)) \end{aligned}$$

Finite Difference Approach Alternatively, we can also approximate the ODE by finite difference methods

- Forward Euler: we approximate $y'(t^n)$ by forward finite differencing:
$$y'(t^n) \approx \frac{y(t^{n+1}) - y(t^n)}{h}.$$
Then $y'(t^n) = f(t^n, y(t^n))$ is approximated by
$$\frac{y^{n+1} - y^n}{h} = f(t^n, y^n).$$

- Backward Euler method: we approximate $y'(t^{n+1})$ by forward finite differencing:
$$y'(t^{n+1}) \approx \frac{y(t^{n+1}) - y(t^n)}{h}$$
in the equation $y'(t^{n+1}) = f(t^{n+1}, y(t^{n+1}))$

- Mid point method. We approximate

$$y'(t^{n+1/2}) = f(t^{n+1/2}, y(t^{n+1/2}))$$

by

$$\frac{y^{n+1} - y^n}{h} = f(t^{n+1/2}, y^{n+1/2})$$

where

$$y^{n+1/2} = y^n + \frac{h}{2} f(t^n, y^n).$$

- RK2: We approximate

$$y'(t^{n+1/2}) = f(t^{n+1/2}, y(t^{n+1/2}))$$

by

$$\frac{y^{n+1} - y^n}{h} = \frac{1}{2}\left(f(t^n, y^n) + f(t^{n+1}, \bar{y}^{n+1})\right)$$

where

$$\bar{y}^{n+1} = y^n + hf(t^n, y^n).$$

10.2 Truncation error and orders of accuracy

Truncation error We would like to estimate the error e^n, which is defined to be

$$e^n = y(t^n) - y^n.$$

In order to find the equation that e^n satisfies, we plug the true solution into the finite difference equation

$$y(t^{n+1}) - y(t^n) = hF_h(t^n, y(t^n)) + h\tau(h). \tag{10.2}$$

The remaining term $\tau(h)$ is called the truncation error.

Definition 10.1. *The truncation error for the numerical scheme*

$$\frac{y^{n+1} - y^n}{h} - F_h(t^n, y^n) = 0 \tag{10.3}$$

is defined to be

$$\tau(h) := \frac{y(t^{n+1}) - y(t^n)}{h} - F_h(t^n, y(t^n))$$

where $y(t)$ is a smooth true solution for $y' = f(t, y)$.

- Forward Euler: For instance, in the forward Euler method, by Taylor expansion,

$$\tau(h) := \frac{y(t^{n+1}) - y(t^n)}{h} - f(t^n, y^n) = \frac{y(t^{n+1}) - y(t^n)}{h} - y'(t^n) = O(h).$$

- RK2: We use trapezoidal rule

$$\int_{t^n}^{t^{n+1}} f(s, y(s))\, ds = \frac{1}{2}h(f(t^n, y(t^n)) + f(t^{n+1}, y(t^{n+1})) + O(h^3).$$

We do not have $y(t^{n+1})$, yet we can use y_1 obtained by the forward Euler to approximate $y(t^{n+1})$. That is, $y_1 = y(t^n) + hf(t^n, y(t^n))$. From (10.2), $|y_1 - y(t^{n+1})| = O(h^2)$. Hence,

$$f(t^{n+1}, y_1) = f(t^{n+1}, y(t^{n+1})) + O(h^2).$$

This yields

$$y(t^{n+1}) = y(t^n) + \frac{1}{2}h(f(t^n, y^n) + f(t^{n+1}, y_1)) + O(h^3).$$

Alternatively, we can use Taylor expansion. The numerical field of RK2 is

$$F_h(t, y) = \frac{1}{2}\left(f(t, y) + f(t + h, y + hf(t, y))\right).$$

The truncation error is defined to be

$$\tau(h) := \frac{y(t^{n+1}) - y(t^n)}{h} - F_h(t^n, y(t^n)).$$

We expand the above equation about t^n: (we abbreviate $y(t^n)$ by y^n in the calculation)

$$\frac{y(t^{n+1}) - y(t^n)}{h} = y'(t^n) + \frac{1}{2}hy''(t^n) + O(h^2).$$

$$\begin{aligned} F_h(t^n, y(t^n)) &:= \frac{1}{2}\left(f(t^n, y^n) + f(t^n + h, y^n + hf(t^n, y^n))\right) \\ &= f(t^n, y^n) + \frac{h}{2}\left(f_t(t^n, y^n) + f_y(y^n)f(t^n, y^n)\right) + O(h^2) \\ &= f(t^n, y^n) + \frac{h}{2}\left(f_t(t^n, y^n) + f_y(y^n)y'(t^n)\right) + O(h^2) \\ &= f(t^n, y^n) + \frac{h}{2}y''(t^n) + O(h^2) \end{aligned}$$

By subtracting these two equations, we get $\tau(h) = O(h^2)$.

Order of accuracy

Definition 10.2. *The numerical scheme (10.3) for (10.1) is said of order p if any smooth solution $y(\cdot)$ of (10.1) satisfies*

$$\tau(h) = O(h^p) \tag{10.4}$$

Thus, forward Euler is first order while RK2 is second order. The quantity

$$\epsilon^n(h) := y(t^{n+1}) - y(t^n) - hF_h(t^n, y(t^n))$$

is called the truncation error of the scheme (10.3).

We can estimate the true error $|y(t^n) - y^n|$ in terms of truncation errors. From

$$\begin{aligned} y(t^{n+1}) &= y(t^n) + hF_h(t^n, y(t^n)) + \epsilon^n \\ y^{n+1} &= y^n + hF_h(t^n, y^n) \end{aligned}$$

Subtracting two equations, we get

$$y(t^{n+1}) - y^{n+1} = (y(t^n) - y^n) + h(F(t^n, y(t^n)) - F(t^n, y^n)) + \epsilon^n$$

Let us denote the true error by $e^n := |y(t^n) - y^n|$ It satisfies

$$e^{n+1} \leq e^n + hLe^n + |\epsilon^n| \leq e^n + hLe^n + Mh^{p+1}.$$

Here we have used the assumption

$$|e^n| \leq Mh^{p+1}$$

for order p schemes. This is a finite difference inequality. We can derive a discrete Gronwall inequality as below. We have

$$\begin{aligned} e^n &\leq (1+hL)e^{n-1} + Mh^{p+1} \\ &\leq (1+hL)^2 e^{n-2} + ((1+hL)+1)Mh^{p+1} \\ &\vdots \\ &\leq (1+hL)^n e^0 + \left(\sum_{k=0}^{n-1}(1+hL)^k\right) Mh^{p+1} \\ &\leq (1+hL)^n e^0 + \frac{(1+hL)^n}{hL} Mh^{p+1} \\ &\leq (1+hL)^n e^0 + \frac{(1+hL)^n}{L} Mh^p \end{aligned}$$

Now, we fix $nh = t$, this means that we want to find the true error at t as $h \to 0$. With t fixed, we have

$$(1+nh)^n = \left((1+hL)^{1/hL}\right)^{Lt} \leq e^{Lt}.$$

Since the initial error $e^0 = 0$, the true error at t is

$$e^n \leq Me^{Lt}h^p.$$

We conclude this analysis by the following theorem.

Theorem 10.1. *If the numerical scheme (10.3) is of order p, then the true error at a fixed time is of order $O(h^p)$.*

10.3 High-order schemes

We list a fourth order Runge-Kutta method (RK4). Basically, we use Simpson rule for integration

$$\int_{t^n}^{t^{n+1}} f(t, y(t))\, dt \approx h\left(f(t^n, y(t^n)) + 4f(t^{n+1/2}, y(t^{n+1/2})) + f(t^{n+1}, y(t^{n+1})\right).$$

The RK4 can be expressed as

$$\begin{aligned} k_1 &= f(t, y) \\ k_2 &= f(t + h/2, y + hk_1/2) \\ k_3 &= f(t + h/2, y + hk_2/2) \\ k_4 &= f(t + h, y + hk_3) \end{aligned}$$

and

$$F(t, y) = \frac{k_1 + 2(k_2 + k_3) + k_4}{6}.$$

One can check that the truncation error by Taylor expansion is $O(h^5)$. Hence the RK4 is a fourth order scheme.