Equations of Evolution

Richard Palais

NCTS Workshop on Differential Equations, Surface Theory, and Mathematical Visualization

February–March, 2003

Part I

The General Theory of Initial Value Problems

Predicting the Future

What we are going to study in this course can be looked at as a generalization of the following interesting and important question:

Suppose that we know the wind velocity at every point of space and at each moment of time.

Suppose also that at a particular moment t_0 we see a tiny puff of smoke pass by, with its center located at a point x_0.

Can we then predict the position $x(t)$ of the smoke at times t close to t_0?

Mathematical Model Building

Translating this kind of vague question concerning the real world into a precisely stated question about rigorously defined mathematical objects is called building a mathematical model, or simply "model building".

We will see that in the present case it leads us to what is called an initial value problem (IVP) for a time-dependent ordinary differential equation (ODE).

Let's see how this works.

Space and Time

In our model, we will represent "time" by $\mathbf{R}$.

It is natural to represent "space" by $\mathbf{R}^3$, but nothing essential changes if we generalize to the case that space is $\mathbf{R}^k$, for some positive integer k, or even by an arbitrary finite dimensional real vector space V.

Moreover—as we shall soon see—this extra generality proves **very** useful, so in what follows we identify space with a finite dimensional orthogonal vector space V, that you may think of $\mathbf{R}^3$ if that helps your intuition.

Later, we will also consider cases where V is an infinite dimensional vector space and we will see that this will lead us to consider initial value problems for partial differential equations (PDE).)

Wind is a Time-Dependent Vector Field

Now the wind velocity at a point x of V at time t is a vector in V that we will denote by $f(x,t)$.

A function $f : V \times \mathbf{R} \to V$ is called a *time-dependent vector field on* V, so saying that we know the wind velocity at all points of space and all instants of time just amounts to saying that we are given such a function f.

(We will always assume that f is at least continuous, but to prove theorems we will actually assume more than that.)

The Path $x(t)$ of the "Puff of Smoke"

We select a smoke particle that at time t_0 (which we will call the *initial time*) is located at the center of the puff of smoke, and we identify the position $x(t)$ of this particle at time t with the position of the puff at time t.

We will assume that the position $x(t)$ is defined for all times t sufficiently close to t_0, so we can think of $t \mapsto x(t)$ as a function defined in some open interval $I \subseteq \mathbf{R}$ containing t_0 and having values in V.

Note that by definition, $x(t_0) = x_0$, and we will call x_0 the *initial position*.

Going With the Flow—the ODE

The characteristic property of a smoke particle is that it "goes with the flow", i.e., its velocity at any time t is the same as the wind velocity at its position, $x(t)$, at time t.

Now the velocity of the smoke particle (at time t) is just the tangent vector to its path $x(t)$ at this time—namely $x'(t)$.

This means that the path $x(t)$ satisfies the relation $x'(t) = f(x(t), t)$!

Such a relation is called a *time-dependent ordinary differential equation*.

Initial Value Problems

Definition. Let V be a finite dimensional real vector space and let $f : V \times \mathbf{R} \to V$ be a time-dependent vector field on V. Given an initial time $t_0 \in \mathbf{R}$ and an initial position x_0 in V, we associate a so-called Initial Value Problem (IVP)

$$\frac{dx}{dt} = f(x,t) \qquad \text{(The ODE)}$$

$$x(t_0) = x_0 \qquad \text{(The Initial Condition)}$$

and we define a solution to this IVP to be any differentiable path $x : I \to V$ defined in some interval I containing t_0 and satisfying $x(t_0) = x_0$ and $x'(t) = f(x(t),t)$ for all $t \in I$.

Our original problem can be paraphrased in terms of this model as: Can we always solve such an IVP?

Questions Concerning the IVP

Here are a few of the interesting questions associated to the IVP that we will consider.

• Given an interval I containing t_0, when can we be sure there exists a solution of the IVP defined in I?

• If a solution does exist in I, is it necessarily unique?

• What can we say about the dependence of a solution on the initial time and initial position.

• Assuming that a solution does exist, can we find efficient algorithms for actually computing it numerically with any desired accuracy?

• What are good methods for visualizing solutions in low dimensions.

A Reformulation of the IVP

Let J be a closed and bounded interval of real numbers containing t_0 and let $C(J, V)$ denote the space of all continuous maps of J into V.

We associate to a time-dependent vector field f on V and $x_0 \in V$ a mapping F of $C(J, V)$ to itself as follows: if $x : J \to V$ is continuous, $F(x) : J \to V$ is defined by:

$$F(x)(t) := x_0 + \int_{t_0}^{t} f(x(s), s)\, ds.$$

Proposition. *A mapping $x : J \to V$ solves the IVP $\frac{dx}{dt} = f(x, t)$, $x(t_0) = x_0$ if and only if x is a fixed point of F, i.e., if and only if $x(t) = x_0 + \int_{t_0}^{t} f(x(s), s)\, ds$.*

Proof. Trivial.

Successive Approximations

Suppose that X is a metric space and F is a continuous map of X to itself. The "Method of Successive Approximations" is a technique for locating fixed points of F. It works as follows. Define $F^n : X \to X$ by composing F with itself n times. If x is any element of X we call the sequence $\{F^n(x)\}$ the sequence of successive approximations defined by x.

Proposition. *If a sequence of successive approximations $\{F^n(x)\}$ converges to a limit p, then p is a fixed point of F.*

Proof. Since F is continuous, $F(p) = F(\lim F^n(x)) = \lim F^{n+1}(x) = p$.

Solving IVPs by Successive Approximations

As above, let $f : V \times \mathbf{R} \to V$ be a time-dependent vector field, J a closed, bounded interval, and define $F : C(J, V) \to C(J, V)$ by:

$$F(x)(t) := x_0 + \int_{t_0}^{t} f(x(s), s)\, ds.$$

Let's try to use Sucessive Approximations to solve IVPs for a couple of special classes of vector fields, f. Perhaps the simplest kind of time-dependent vector fields is one that is constant in space, i.e., of the form $f(x, t) = \phi(t)$, In this case the mapping F is clearly the constant map with value $\Phi \in C(J, V)$ defined by $\Phi(t) = x_0 + \int_{t_0}^{t} \phi(s)\, ds$, i.e., the antiderivative of ϕ.

This is clearly a fixed point of F and also the solution of the IVP, and we see that Successive approximations works in this case.

Linear ODE

Let us denote by $L(V)$ the vector space of linear maps of V to itself. Any $T \in L(V)$, defines a time-independent vector field: $f(x,t) = Tx$.

Recall how the exponential of T is defined using a power series, namely $\sum_{k=0}^{\infty} \frac{1}{k!} T^k$. The series is clearly absolutely convergent and so defines an element $\exp(T)$ in $L(V)$.

By absolute convergence, we can differentiate the series for $\exp(tT)$ termwise, and we see that $\frac{d}{dt} \exp(tT) = T \exp(tT)$.

It follows that the solution of the IVP for the vector field f with initial data t_0, x_0 is $\exp((t-t_0)T)x_0$.

Linear ODE by Successive Approximation

Let's try to solve the linear ODE $\frac{dx}{dt} = Tx$ with the initial condition $x(t_0) = x_0$ by successive approximation, choosing as the initial approximation the costant curve $x_1(t) = x_0$ for all t in J.
The next approximation is:
$x_2(t) := x_0 + \int_{t_0}^{t} T(x_1(s))\, ds = (I + (t - t_0)T)(x_0),$
and an easy induction shows that the $n+1$-st successive approximation is:
$x_{n+1}(t) := x_0 + \int_{t_0}^{t} T(x_n(s))\, ds = (\sum_{k=0}^{n} \frac{(t-t_0)^k}{k!} T^k)(x_0).$
Since this converges to $\exp((t - t_0)T)(x_0)$, the solution of the IVP, we see that the method of successive approximations solves the IVP for the case of linear ODE also.

Locally Lipschitz Vector Fields

There is a simple answer to the existence question for an IVP. It turns out that just as long as the time-dependent vector field f is continuous, a solution will always exist on some interval containing t_0.

However to get a satisfactory theory for the IVP, a somewhat more stringent condition than just continuity is required.

Definition. A time-dependent vector field on V, $f : V \times \mathbf{R} \to V$ is called **locally Lipschitz** if for each $(x_0, t_0) \in V \times \mathbf{R}$ there exists a $K > 0$ such that $\|f(x_1, t) - f(x_2, t)\| < K \|x_1 - x_2\|$ for all x_1 and x_2 sufficiently close to x_0 and all t sufficiently close to t_0.

Existence and Uniqueness Theorem for Locally Lipschitz Vector Fields

Theorem. *Let V be a finite dimensional orthogonal vector space and let $f : V \times \mathbf{R} \to V$ be a time-dependent locally Lipschitz vector field in V. Given any $T \in \mathbf{R}$ and $P \in V$ there exist positive real numbers δ and ϵ such that if $|T - t_0| < \delta$ and $\|P - x_0\| < \delta$ then the IVP:*

$\frac{dx}{dt} = f(x, t)$
$x(t_0) = x_0$
has a unique solution $x_{x_0,t_0}(t)$ on the interval $I = (t_0 - \epsilon, t_0 + \epsilon)$. Moreover, this solution is contiuously diferrerentiable in t and is Lipschitz in the initial data x_0 and t_0.

We will sketch the proof below after a few remarks.

C^1 Implies Locally Lipschitz

Let V and W be orthogonal vector spaces, U a convex open set in V, $f : U \to W$ a C^1 map, and Df_p, the differential of f at $p \in U$.

If $p, q \in U$ and $\sigma(t) = p + t(q - p)$ is the line joining them, the so-called "finite difference formula" says: $f(q) - f(p) = \int_0^1 Df_{\sigma(t)}(q-p)\, dt$, and it follows that:

$$\|f(q) - f(p)\| \leq \left(\int_0^1 \left\|Df_{\sigma(t)}\right\| \, dt\right) \|(q - p)\|,$$

Consequently, the supremum of $\|Df_p\|$ for p in U is a Lipschitz constant for f. (In fact, the smallest one.)

In particular it follows that a C^1 time-dependent vector field is locally Lipschitz and so satisfies the Local Existence and Uniqueness Theorem.

A Counter-Example

Exercise. Show that continuity of V is **not** sufficient to guarantee uniqueness for an IVP.

Hint: the classic example (with $n = 1$) is the initial value problem $\frac{dx}{dt} = \sqrt{x}$, and $x(0) = 0$.

Show that for each $T > 0$, we get a distinct solution $x_{_T}(t)$ of this IVP by defining $x_{_T}(t) = 0$ for $t < T$ and $x_{_T}(t) = \frac{1}{4}(t - T)^2$ for $t \geq T$.

Contracting Maps

A mapping F of a metric space X to itself is called a contracting map (or a contraction) if it satisfies a Lipschitz condition with constant $K < 1$.

Fundamental Contraction Inequality.
If $F : X \to X$ is a contraction, and if $K < 1$ is a Lipschitz constant for F, then for all x_1 and x_2 in X,

$$\rho(x_1, x_2) \leq \frac{1}{1-K}\Big(\rho(x_1, F(x_1)) + \rho(x_2, F(x_2))\Big).$$

Proof. Exercise.

Corollary. *A contraction mapping can have at most one fixed point.*

Proof. Assuming that x_1 and x_2 are fixed points we deduce immediately that $\rho(x_1.x_2)$ must be zero.

The Banach Contraction Principle

Let $F : X \to X$ and F^n its n-fold composition with itself. If F satisfies a Lipschitz condition with constant K, by an easy induction F^n satisfies a Lipschitz condition with constant K^n, so by the Fundamental Contraction Mapping Inequality, if $K < 1$ then

$$\rho(F^n(x), F^m(x)) \leq \frac{K^n + K^m}{1 - K}\Big(\rho(x, F(x))\Big).$$

In particular the successive approximation sequence $\{F^n(x)\}$ is a Cauchy sequence. Hence:

Banach Contraction Principle. *If X is a complete metric space and if $F : X \to X$ is a contraction mapping, then F has a unique fixed point p in X and for any $x \in X$ the successive approximation sequence $\{F^n(x)\}$ converges to p.*

A Stopping Rule

When do we stop iterating and accept the current approximation? Suppose an "error" of ϵ is acceptable, and we start our iteration at $x \in X$. The Fundamental Inequality, with $x_1 = f^N(x)$ and $x_2 = p$ gives:

$$\begin{aligned}\rho(f^N(x), p) &\leq \tfrac{1}{1-K}\rho(f^N(x), f^N(f(x))) \\ &\leq \tfrac{K^N}{1-K}\rho(x, f(x)).\end{aligned}$$

To insure $\rho(f^N(x), p) \leq \epsilon$, we must choose N so large that $\frac{K^N}{1-K}\rho(x, f(x)) < \epsilon$. We can compute $\rho(x, f(x))$ after the first iteration and then find N by solving the above inequality for N:

Stopping Rule. If $N > \frac{\log(\epsilon)+\log(1-K)-\log(\rho(x,f(x)))}{\log(K)}$ then $\rho(f^N(x), p) < \epsilon$.

Speed of Convergence

Suppose we take $\epsilon = 10^{-m}$ in our stopping rule inequality. What we see is that the growth of N with m is a constant plus $m/|\log(K)|$, or in other words, to get one more decimal digit of precision we have to do (roughly) $1/|\log(K)|$ more iteration steps. So if we need N iterative steps to get m decimal digits of precision, then we need another N to double the precision to $2m$ digits.

Existence and Uniqueness Proof

The Existence and Uniqueness Theorem will follow from the Banach Contraction Principle if we can show that for $\|v_0\| < \epsilon$, $F : C(J, V) \to C(J, V)$ maps X into itself and has K as a Lipschitz bound.

If $\sigma \in X$ then

$$\begin{aligned}\|F(\sigma)(t) - p\| &\leq \|v_0 - p\| + \int_0^t \|f(\sigma(s), s)\| \, ds \\ &\leq \epsilon + \delta B \leq 2\epsilon,\end{aligned}$$

so F maps X to itself.

Existence and Uniqueness Proof (Cont.)

And if $\sigma, \tau \in X$ then

$\|f(\sigma(t), t) - f(\tau(t), t)\| \leq M \|\sigma(t) - \tau(t)\|$, so

$$\begin{aligned}
\|F(\sigma)(t) - F(\tau)(t)\| &\leq \int_0^t \|f(\sigma(s), s) - f(\tau(s), s)\| \, ds \\
&\leq \int_0^t M \|\sigma(s) - \tau(s)\| \, ds \\
&\leq \int_0^t M \rho(\sigma, \tau) \, ds \\
&\leq \delta M \rho(\sigma, \tau) \leq K \rho(\sigma, \tau),
\end{aligned}$$

and it follows that $\rho(F(\sigma), F(\tau) \leq K \rho(\sigma, \tau)$. ∎

Maximal Solutions of the IVP

The Existence and Uniqueness Theorem is a central result in the theory of ODE with a great many important consequences. We next consider one easy corollary. A solution $\sigma : J \to V$ of the IVP is called the **maximal solution** for given initial data t_0 and x_0 if any other solution $x : I \to V$ with the same initial data, is a restriction of σ to a subinterval I of J.

Proposition. *If $f : V \times \mathbf{R} \to V$ is a locally Lipschitz time-dependent vector field, then for any initial data t_0 and x_0, the maximal solution of the IVP exists.*

Proof. Exercise. Hint:If $x_1 : I_1 \to V$ and $x_2 : I_2 \to V$ are solutions of the IVP with the same initial data, show that the set of $t \in \mathbf{R}$ with $x_1(t) = x_2(t)$ is a non-empty open and closed subset of $I_1 \cap I_2$.

Maximal Solutions (Cont.)

Exercise. Show that if $\sigma : (a, b) \to V$ is a maximal solution of an IVP, then either $b = \infty$ or $\|\sigma(t)\| \to \infty$ as $t \to b$. Similarly, either $a = -\infty$ or $\|\sigma(t)\| \to \infty$ as $t \to a$. Hint: If $b < \infty$ and $\|\sigma(t)\| \not\to \infty$ as $t \to b$, there is a sequence $\{t_n\}$ converging to b with$\{\sigma(t_n)\}$ converging to $p \in V$. Use the Existence and Uniqueness Theorem with $t_0 = b$ and $x_0 = p$ to show that the solution σ could be extended to $(a, b + \epsilon)$ with $\epsilon > 0$, contradicting maximality of σ.

Exercise. Suppose the vector field f is bounded, or more generally satisfies $\int_1^\infty \frac{dr}{B(r)} = \infty$ where $B(r) = \sup_{\|x\|<r} \|f(x, t)\|$. Show that each maximal solution is defined on all of $\mathbf{R}$. Hint: How long does it take a solution to get outside a ball of radius R?

Global Existence vs. Finite Time Blowup

If for a particular initial condition the maximal solution is defined on the entire real line then we say we have global existence for that initial condition, otherwise we say that there is finite-time blowup.

Exercise. On $\mathbf{R}$, consider the time-independent ODE $\frac{dx}{dt} = x^2$ with the initial condition $x(0) = x_0$. Show that in this case the maximal solution is $x(t) = \frac{x_0}{1-x_0 t}$ with the interval of definition is $(-\infty, \frac{1}{x_0})$ if $x_0 > 0$ and $(\frac{1}{x_0}, \infty)$ if $x_0 < 0$—in other words we have finite-time blowup at time $T = \frac{1}{x_0}$.

Autonomous vs. Non-Autonomous ODE

A time-independent vector field f on V is also called **autonomous**. An obvious and characteristic property of autonomous ODEs $\frac{dx}{dt} = f(x)$ is that if $x(t)$ is a solution defined on (a, b) then $x(t + c)$ is a solution defined on $(a - c, b - c)$. In particular, if the maximal solution for the initial condition $x(0) = p$ is $\sigma_p : (a, b) \to V$ then the maximal solution for the initial condition $x(t_0) = p$ is just $\sigma_p(t - t_0)$, defined on $(a + t_0, b + t_0)$.

Exercise. We call f **complete** if σ_p has domain $\mathbf{R}$ for all $p \in V$. In this case we define the map $\phi_t : V \to V$ for each $t \in \mathbf{R}$ by $\phi_t(p) = \sigma_p(t)$. Show that $t \mapsto \phi_t$ is a homomorphism of $\mathbf{R}$ into the group of diffeomorphisms of V (i.e., $\phi_{t_1+t_2} = \phi_{t_1} \circ \phi_{t_2}$).

Reduction Theorems

We remarked earlier that even if one is interested only in solving the IVP for time-dependent vector fields in $\mathbf{R}^3$, there are still good reasons to consider the problem in more general vector spaces.

We illustrate this by showing how to:

1) reduce an IVP for a non-autonomous vector field in V to an IVP for a time-independent vector field in $V \times \mathbf{R}$, and

2) reduce an IVP for higher order ODE in V to an IVP for a vector field in a product of copies of V.

The proofs will be left as exercises.

Time-Dependent → Time-Independent

As we have seen, autonomous ODE have a number of simplifying features, and this makes the following reduction quite useful.

Exercise. Let $f : V \times \mathbf{R} \to V$ be a time-dependent vector field in V, and define an associated time independent vector field $\tilde{f}$ in $V \times \mathbf{R}$ by $\tilde{f}(x, z) = (f(x, z), 1)$. Show that $y(t) = (x(t), z(t))$ is a solution of the differential equation $\frac{dy}{dt} = \tilde{f}(y)$ if and only if $z(t) = t + c$ and $x(t)$ is a solution of $\frac{dx}{dt} = f(x, t + c)$. Deduce that if $y(t) = (x(t), z(t))$ solves the IVP $\frac{dy}{dt} = \tilde{f}(y)$, $y(t_0) = (x_0, t_0)$, then $x(t)$ is a solution of the IVP $\frac{dx}{dt} = f(x, t)$, $x(t_0) = x_0$.

Second Order → First Order

A curve $x(t)$ in V is a solution of the second ODE $\frac{d^2x}{dt^2} = f(x, \frac{dx}{dt}, t)$ in V if $x''(t) = f(x(t), x'(t), t)$. (Here of course f is a map $V \times V \times \mathbf{R} \to V$.)

Define an associated time-dependent vector field $\tilde{f}$ on $V \times V$ by $\tilde{f}(x, v, t) = (v, f(x, v, t))$—so the associated first order ODE in $V \times V$ is $\frac{dx}{dt} = v$, $\frac{dv}{dt} = f(x, v, t)$.

Exercise. Define the IVP for for the above second order ODE, and analyze the relation of this IVP and the IVP for the time-dependent vector field $\tilde{f}$ on $V \times V$. Use this to formulate and prove an Existence and Uniqueness Theorem for second order ODE. Now, generalize this to m-th order ODE in V

Gronwall's Inequality.

The following estimate plays a very important role in ODE theory.

Gronwall's Inequality. *Let $u : [0, T) \to [0, \infty)$ be a continuous, non-negative, real-valued function and assume that $u(t) \leq U(t) := C + K \int_0^t u(s)\, ds$ for certain constants $C \geq 0$ and $K > 0$. Then $u(t) \leq Ce^{Kt}$.*

Exercise. Prove Gronwall's Inequality. Hint: Since $u \leq U$, it is enough to show that $U(t) \leq Ce^{Kt}$, or equivalently that $e^{-Kt}U(t) \leq C$, and since $U(0) = C$, it will suffice to show that $e^{-Kt}U(t)$ is non-increasing, i.e., that $(e^{-Kt}U(t))' \leq 0$. But, since $(e^{-Kt}U(t))' = e^{-Kt}(U'(t) - KU)$ and $U' = Ku$, this just says that $Ke^{-Kt}(u - U) \leq 0$.

Continuity w.r.t Initial Conditions.

Theorem. *If f is a C^1 vector field on V and $\sigma_p(t)$ the maximal solution curve of $\frac{dx}{dt} = f(x)$ with initial condition p, then as q tends to p, $\sigma_q(t)$ approaches $\sigma_p(t)$, uniformly for t in a bounded interval I.*

Proof. Since $\sigma_p(t) = p + \int_0^t f(\sigma_p(s), s)\, ds$,
$\|\sigma_p(t) - \sigma_q(t)\| \leq$
$\|p - q\| + \int_0^t \|f(\sigma_p(s), s) - f(\sigma_q(s), s)\|\, ds.$
But on any bounded set (so on some neighborhood of $\sigma_p(I) \times I$), f satisfies a Lipschitz condition:
$\|f(x,t) - f(y,t)\| \leq K\|x - y\|$, so $\|\sigma_p(t) - \sigma_q(t)\| \leq \|p - q\| + K\int_{t_0}^t \|\sigma_p(s) - \sigma_q(s)\|\, ds$, and by Gronwall's Inequality, $\|\sigma_p(t) - \sigma_q(t)\| \leq \|p - q\|\, e^{Kt}$. ∎

We prove in Appendix A that if f is C^r then $(p,t) \mapsto \sigma_p(t)$ is a C^r map.

The IVP for Inhomogeneous Linear ODE

For a linear ODE $\frac{dx}{dt} = Ax$ with initial condition $x(t_0) = x_0$ we saw that the solution is $\exp((t-t_0)A)x_0$. If $g : \mathbf{R} \to V$ is a smooth function, then we can add it to the right hand side of the ODE, getting a so-called inhomogeneous linear ODE, and it turns out that the IVP for such equations can be solved in a fairly explicit form by a formula that for historical reasons is called "The Variation of Parameters Formula".

Exercise. Show (by direct verification) that the solution of the IVP for the inhomogeneous linear differential equation $\frac{dx}{dt} = Ax + g(t)$ with initial condition $x(0) = x_0$ is given by:

$$x(t) = \exp(tA)x_0 + \int_0^t \exp((t-s)A)g(s)\,ds.$$

Existence of a Periodic Orbit

Exercise. Assume that the linear operator A is what is called "asymptotically stable"—namely that all of its eigenvalues have negative real part—and also that the forcing term $g(t)$ is periodic with period $T > 0$. Show that there is a point $p \in V$ for which the solution $x(t)$ with initial value $x(0) = p$ is periodic with period T.

Hint: Since x is given by the above variation of parameters formula, the condition that it be periodic of period T is that $p = e^{TA}p + \int_0^T e^{(T-s)A} g(s)\, ds$, or that $p = (I - e^{TA})^{-1} \int_0^T e^{(T-s)A} g(s)\, ds$. Why is the operator $(I - e^{tA})$ invertible?

Part II

Numerical Solutions of Initial Value Problems

Numerical Algorithms for Solving IVPs

Very few initial value problems admit explicit closed form solutions, so in general we must construct solutions numerically with the aid of a computer. But what algorithm should we use?

The natural first guess is successive approximations. But while that is a powerful theoretical tool for studying general properties of solutions (in particular, existence and uniqueness), it is not an efficient method for constructing numerical solutions.

In fact there is no one simple answer to this question, for there is no one algorithm that is "best" in all situations. Below we will look at just two methods from the numerical analyst's extensive toolbox for solving initial value problems, Euler and Runge-Kutta.

The General Approach

In what follows we will suppose that f is a C^1 time-dependent vector field on $\mathbf{R}^n$, and given t_0 in $\mathbf{R}$ and x_0 in $\mathbf{R}^n$ we will denote by $x(t)$ or $\sigma(f, x_0, t_0, t)$ the maximal solution of the IVP $\frac{dx}{dt} = f(x, t)$ with initial condition $x(t_0) = x_0$.

The goal in the numerical integration of ODE is to devise efficient methods for approximating the solution $x(t)$ on an interval $I = [t_0, T]$.

The basic strategy is to interpolate N equally spaced gridpoints $t_1, \ldots t_N$ in the interval I (defined by $t_k := t_0 + k\Delta t$ with $\Delta t = \frac{T - t_0}{N}$), and use some algorithm to define values $x_1, \ldots, x_N$ in $\mathbf{R}^n$, in such a way that when N is large each x_k is close to the corresponding $x(t_k)$.

The General Approach (Cont.)

The quantity $\max_{1 \leq k \leq N} \|x_k - x(t_k)\|$ is called the **global error** of the algorithm, and if it converges to zero as N tends to infinity (for every choice of f, t_0, x_0, and T), then we say that we have a **convergent algorithm**.

Even if a algorithm is convergent, that does not necessarily mean that it will provide an adequate method for solving initial value problems in all situations; other considerations such as stability and rate of convergence are important. However, if an algorithm is not at least convergent, that is sufficient reason to reject it as a tool for solving IVPs numerically.

Stepping Methods

A common way to construct the algorithm that produces the values $x_1, \ldots, x_N$ uses a recursion based on a so-called **stepping procedure**, namely a function, $\Sigma(f, x_0, t_0, \Delta t)$, having as inputs:

1) a time-dependent vector field f on $\mathbf{R}^n$,
2) an initial condition x_0 in $\mathbf{R}^n$,
3) an initial time t_0 in $\mathbf{R}$, and
4) a "time-step" Δt in $\mathbf{R}$,
and with output a point of $\mathbf{R}^n$ that approximates $\sigma(f, x_0, t_0, t_0 + \Delta t)$ well when Δt is small.

More precisely, the so-called **local truncation error**, defined by $\|\sigma(f, x_0, t_0, t_0 + \Delta t) - \Sigma(f, x_0, t_0, \Delta t)\|$, should approach zero at least quadratically in the time-step Δt.

Stepping Methods (Cont.)

Given such a stepping procedure, the approximations x_k of the $x(t_k)$ are defined recursively by $x_{k+1} := \Sigma(f, x_k, t_k, \Delta t)$. Numerical integration methods that follow this general pattern are referred to as **finite difference methods**.

There are two main sources contributing to the global error, $\|x_k - x(t_k)\|$. At each step there will be an additional local truncation error, and after the first step, there will be an error because the recursion uses $\Sigma(f, x_k, t_k, \Delta t)$ rather than $\Sigma(f, x(t_k), t_k, \Delta t)$. In practice there is a third source of error, namely round-off error from using floating-point arithmetic, We will ignore this, pretending that our computers do precise real arithmetic, but there are situations where it is important to take round-off error into consideration.

Euler's Method

Euler's Method is defined by the particularly simple and natural stepping procedure: :

Euler Step: $\Sigma^E(f, x_0, t_0, \Delta t) := x_0 + \Delta t\, f(x_0, t_0)$.

It is not hard to see why this is a good choice. If as above we denote $\sigma(f, x_0, t_0, t)$ by $x(t)$, then by Taylor's Theorem:

$$\begin{aligned} x(t_0 + \Delta t) &= x(t_0) + \Delta t\, x'(t_0) + O(\Delta t^2) \\ &= x_0 + \Delta t\, f(x_0, t_0) + O(\Delta t^2) \\ &= \Sigma^E(f, x_0, t_0, \Delta t) + O(\Delta t^2), \end{aligned}$$

so that $\|\sigma(f, x_0, t_0, t_0 + \Delta t) - \Sigma(f, x_0, t_0, \Delta t)\|$, the local truncation error for Euler's Method, does go to zero quadratically in Δt.

Euler's Method (Cont.)

When we partition $[0, T]$ into N equal parts, $\Delta t = \frac{T-t_0}{N}$, each step in the recursion for computing x_k will contribute a local truncation error that is $O(\Delta t^2) = O(\frac{1}{N^2})$, and since there are N steps in the recursion, this suggests that the global error will be $O(\frac{1}{N})$, and hence will go to zero as N tends to infinity. Thus we expect Euler's Method to be a convergent algorithm. We will give a rigorous argument below.

Exercise. Show that Euler's Method applied to the initial value problem $\frac{dx}{dt} = x$ with $x(0) = 1$ gives $\lim_{N\to\infty}(1 + \frac{t}{N})^N = e^t$.

Error Estimate for Euler Method

Assume the vector field f satisfies the local Lipschitz bound $\|f(p,t) - f(q,t)\| \leq L\,\|p - q\|$. We use an argument of Hermann Karcher to estimate the error in Euler's method.

Recall that Euler's Method approximates $x(t)$ at points $t_n := t_{n-1} + \Delta T = t_0 + n\,\Delta T$, where $\Delta T = \frac{T - t_0}{N}$, and the approximations $e(t_n)$ are defined inductively by $e(t_0) := x_0$, and $e(t_{n+1}) := e(t_n) + \Delta T\, f(e(t_n), t_n)$.

We interpolate the discrete approximations $e(t_n)$ with a piecewise-linear curve $e(t)$ defined on $[t_n, t_{n+1}]$ by $e(t) := e(t_n) + (t - t_n)\, f(e(t_n), t_n)$. Think of this as considering the time difference $\Delta t := t - t_n$ as a variable, allowing us to estimate the difference or "error", $\mathrm{Err}(t) := \|x(t) - e(t)\|$ by a Gronwall-like argument.

Error Estimate for Euler Method (Cont.)

We first estimate the error for a single time-step, i.e., on the interval $t_0 \le t \le t_0 + \Delta T$. From the definition of e, $\dot{e}(t) = f(x_0, t_0)$, $t_0 \le t \le t_0 + \Delta T$, so $\ddot{e} = 0$. It follows that $\dot{x}(t) - \dot{e}(t) = f(x(t), t) - f(x_0, t_0) = f(x(t), t) - f(e(t), t) + f(e(t), t) - f(x_0, t_0)$, so

$$\|\dot{x}(t) - \dot{e}(t)\| \le L \|x(t) - e(t)\| + \|f(e(t), t) - f(x_0, t_0)\|.$$

The second term on the right is not yet in a form to apply a Gronwall argument. However, if we define $K := \max_{t_0 \le t \le t_0 + \Delta t} \left\| \frac{d}{dt}(f(e(t), t) - f(x_0, t_0) \right\|$ then we obtain the differential inequality

$$\|\dot{x}(t) - \dot{e}(t)\| \le L \|\mathrm{Err}(t)\| + K(t - t_0).$$

Error Estimate for Euler Method (Cont.)

Since
$\mathrm{Err}(t) = \left\| \int_{t_0}^{t} (\dot{x}(t) - \dot{e}(t))\, dt \right\| \leq \int_{t_0}^{t} \|\dot{x}(t) - \dot{e}(t)\|\, dt$,
we see that $\|\mathrm{Err}(t)\| \leq \psi(t)$, where ψ is the differentiable function:

$$\psi(t) := \|\mathrm{Err}(t_0)\| + L \int_{t_0}^{t} \|\mathrm{Err}(t)\|\, dt + K \int_{t_0}^{t} (t - t_0)\, dt.$$

Since $\dot{\psi} = L\,\|\mathrm{Err}(t)\| + K(t - t_0)$ and $\|\mathrm{Err}(t)\| \leq \psi$, we have the differential inequality $\dot{\psi} \leq L\psi + K(t - t_0)$ that we now use for a Gronwall argument.

Error Estimate for Euler Method (Cont.)

Compute the derivative of the function

$$\left(\psi + \frac{K}{L^2} + \frac{K}{L}(t - t_0)\right) \cdot e^{-L\cdot(t-t_0)}.$$

$$\frac{d}{dt}\left(\left(\psi + \frac{K}{L^2} + \frac{K}{L}(t - t_0)\right) \cdot e^{-L\cdot(t-t_0)}\right) =$$

$$\left(\dot{\psi} + \frac{K}{L} - L\left(\psi + \frac{K}{L^2} + \frac{K}{L}(t - t_0)\right)\right) \cdot e^{-L\cdot(t-t_0)}.$$

By the differential inequality for ψ, this function has a non-positive derivative, so all of its values are less than its "initial" value at t_0, namely $\mathrm{Err}(t_0) + \frac{K}{L^2}$, and solving for ψ we obtain the desired error estimate:

$$\mathrm{Err}(t) \leq \psi(t) \leq \left(\mathrm{Err}(t_0) + \frac{K}{2}(t - t_0)^2\right) e^{L\cdot(t-t_0)}$$

for $t_0 \leq t \leq t_0 + \Delta T$.

Error Estimate for Euler Method (Cont.)

To iterate this estimate we define the starting point for the second time step as $x_1 := e(t_0 + \Delta T)$, so that we have the initial error bound

$$|x(t_0 + \Delta t) - x_1| \leq$$

$$\mathrm{Err}(t_1) := \left(\mathrm{Err}(t_0) + \frac{K}{2}(t - t_0)^2\right) e^{L \cdot \Delta T},$$

and after the second time step we have

$$\mathrm{Err}(t) \leq \psi(t) \leq \left(\mathrm{Err}(t_1) + \frac{K}{2}(t - t_1)^2\right) e^{L \cdot (t - t_1)}$$

for $t_1 \leq t \leq t_1 + \Delta T$.

Error Estimate for Euler Method (Cont.)

To reach the fixed time T one needs N time steps of size $\Delta t := (T - t_0)/N$ and the N-fold iteration of the error estimate gives (replace $N \cdot \Delta T$ by $(T - t_0)$, recall $\mathrm{Err}(t_0) = 0$ and use the sum of the geometric series):

$$\begin{aligned}
\mathrm{Err}(t) &\leq \mathrm{Err}(t_0) \cdot e^{L(T-t_0)} + \frac{K}{2}\Delta T^2 \cdot \sum_{k=1}^{N} e^{kL\cdot\Delta T} \\
&\leq \frac{K}{2}\Delta T^2 \cdot (e^{L(T-t_0+\Delta T)} - 1)/(e^{L\Delta t} - 1) \\
&\leq \frac{K}{2L}\Delta T \cdot (e^{L(T-t_0+\Delta T)} - 1).
\end{aligned}$$

This proves in particular that when $\Delta T \to 0$ the iterated Euler curves converge uniformly to the exact solution, or in other words that Euler's Method is a convergent algorithm.

Runge-Kutta

Despite what we have just proved, perhaps the only positive thing that can be said about the Euler method for solving an IVP is that it is intuitive and easy to program. Beyond that there is little to recommend it as a practical method for solving real-world problems. It requires very small time steps to get reasonable accuracy, making it very slow, and in fact it is rarely used except for pedagogical purposes.

A general purpose finite difference method for solving IVPs that is the most useful (and the only other one that we will consider) is Runge-Kutta, or more precisely the fourth order Runge-Kutta Method, as there is a whole family of Runge-Kutta methods. It is in fact one of the most implemented an useful pieces of numerical software for any purpose.

Runge-Kutta Step

The stepping procedure for fourth order Runge-Kutta is a **lot** less transparent than that for Euler. It is given by the following formula:

Runge-Kutta Step

$\Sigma^{RK^4}(f, x_0, t_0, \Delta t) := x_0 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$,
where:
$k_1 = \Delta t\, f(x_0, t_0)$
$k_2 = \Delta t\, f(x_0 + \frac{1}{2}k_1, t_0 + \frac{\Delta t}{2})$
$k_3 = \Delta t\, f(x_0 + \frac{1}{2}k_2, t_0 + \frac{\Delta t}{2})$
$k_4 = \Delta t\, f(x_0 + k_3, t_0 + \Delta t)$

Runge-Kutta Pseudo Code

```
function RungeKutta4(f,x0,t0,h,N):vector;
f: function(v:vector;s:real):vector;
x0:vector ;t0,h:real; N:integer;
var  j:integer; t:real; x,k1,k2,k3,k4:vector;
begin
   t := t0;    x := x0;
   for j := 1 to N do
      begin
         k1 := h f(x,t);
         k2 := h f(x+ k1/2,t+h/2);
         k3 := h f(x+ k2/2,t+h/2);
         k4 := h f(x+ k3,t+h);
         x := x + (k1+2(k2+k3)+k4)/6;
         t := t + h;
      end;
   RungeKutta4 := x;
end;
```

Runge-Kutta (Cont.)

It is of course a fair question to ask where such a strange formula comes from. If you are familiar with Simpson's Rule for evaluating the definite integral of a function $\phi(t)$, then the above should not look unreasonable, and indeed if $f(x,t) = \phi(t)$ then recall that the solution of the IVP reduces to the integral of ϕ and in this case the Runge-Kutta formula reduces precisely to Simpson's Rule. And like Simpson's Rule, Runge-Kutta is fourth order, meaning that the local truncation error goes to zero as the fifth power of the step-size, and the global error as the fourth power. So if for a fixed step-size we have attained an accuracy of 0.1, then with one-tenth the step-size (and so ten times the number of steps and ten times the time) we can expect an accuracy of 0.00001, whereas with the Euler method, ten times the time would only increase accuracy from 0.1 to 0.01.

Part III

Classical Mechanics Lattice Models and Wave Equations

Newton's Equations

The second order ODE that arise in practice frequently have a special form that we will call **Generalized Newton's Equations**. We will define this class of equations below

We start with an orthogonal vector space $\mathcal{C}$ that we call **configuration space** and define $T\mathcal{C}$ to be $\mathcal{C} \times \mathcal{C}$, and $\Pi : T\mathcal{C} \to \mathcal{C}$ to be the projection $(x, \dot{x}) \mapsto x$. (If you know about such things you can think of $\mathcal{C}$ as a Riemannian manifold and $T\mathcal{C}$ as its tangent bundle.)

We next define four real-valued functions on $T\mathcal{C}$, the Kinetic energy $K(x, \dot{x}) := \frac{1}{2} \|\dot{x}\|^2$, the potential energy $U(x, \dot{x}) := U(x)$ where $U(x)$ is a smooth function on $\mathcal{C}$ (also called the potential energy), the Lagrangian $\mathcal{L} := K - U$, and the Hamiltonian (or total energy) $H := K + U$.

Newton's Equations (Cont.)

If $(x_1, \ldots, x_n)$ are local coordinates in $\mathcal{C}$, we define associated "canonical coordinates" $(q_1, \ldots, q_n, \dot{q}_1, \ldots \dot{q}_n)$ in $T\mathcal{C}$ by $q_i(x, \dot{x}) := x_i(x)$, and $\dot{q}_i(x, \dot{x}) := dx_i(\dot{x})$.

In other words, $q_i(x, \dot{x})$ is the i-th coordinate of the projection $\Pi(x, \dot{x})$, and $\dot{q}_i(x, \dot{x})$ is the i-th component of the vector $\dot{x}$ in the coordinate system.

Exercise. Show that expressed in canonical coordinates, the Kinetic Energy has the form

$$K(q, \dot{q}) = \sum_{i,j} \tfrac{1}{2} g_{ij}(q) \dot{q}_i \dot{q}_j$$

where $g_{ij}(q)$ is a positive definite symmetric matrix. Hint: In orthonormal coordinates y_i this is clear, with $g_{ij} = \delta_{ij}$. If the transformation of coordinates is given by $y_i = \phi_i(x)$ then $dy_i = \sum_j \frac{\partial \phi_i}{\partial x_j} dx_j$, and show that $g_{ij} = \sum_k \frac{\partial \phi_k}{\partial x_i} \frac{\partial \phi_k}{\partial x_j}$.

Newton's Equations (Cont.)

If $\sigma : [a, b] \to \mathcal{C}$ is a smooth curve on $\mathcal{C}$, then we call a curve $\tilde{\sigma} : [a, b] \to T\mathcal{C}$ a **lifting** of σ if $\Pi \circ \tilde{\sigma} = \sigma$. This is just the same thing as a vector field defined along σ. We define the canonical lifting of σ to be $\dot{\sigma}(t) := (\sigma(t), \sigma'(t)$, where as usual $\sigma'(t)$ is the tangent to σ at t.

Exercise. Given canonical coordinates $q_i, \dot{q}_i$ as above, and a curve $\tilde{\sigma} : [a, b] \to T\mathcal{C}$, define $q_i(t) := q_i(\tilde{\sigma}(t))$ and $\dot{q}_i(t) = \dot{q}_i(\tilde{\sigma}(t))$. Show that a necessary and sufficient condition for $\tilde{\sigma}$ to be the canonical lifting of its projection is that the $q_i(t)$ and $\dot{q}_i(t)$ satisfy the system of first order ODE $\frac{dq_i(t)}{dt} = \dot{q}_i(t)$.

Newton's Equations (Cont.)

Given canonical coordinates $q_i, \dot{q}_i$ as above, We define functions p_i on $T\mathcal{C}$ called the **canonical momenta** associated to the these coordinates by $p_i := \frac{\partial \mathcal{L}}{\partial \dot{q}_i}$. Since $\mathcal{L}(q, \dot{q}) = K(q, \dot{q}) - U(q) = \sum_{i,j} \frac{1}{2} g_{ij}(q) \dot{q}_i \dot{q}_j - U(q)$, $p_i = \sum_i g_{ij}(q) \dot{q}_j$. Since the matrix $g_{ij}(q)$ is positive definite and so non-singular, and since the q_i and $\dot{q}_i$ are local coordinates in $T\mathcal{C}$, it follows that the q_i and p_i are likewise local coordinates. We now define the generalized Newton's Equations to be the first order system of equations:

$$\frac{dq_i}{dt} = \dot{q}_i$$
$$\frac{dp_i}{dt} = \frac{\partial \mathcal{L}}{\partial q_i}$$

Newton's Equations (Cont.)

Exercise. Let G_{ij} be the matrix-valued function on $T\mathcal{C}$ which at each point is the matrix inverse to g_{ij} at that point. Show that the Generalized Newton's Equations on $T\mathcal{C}$ are equivalent to the system of n second order ODE on $\mathcal{C}$:

$$\frac{d^2x_k}{dt^2} = \sum_{i=1}^{n} G_{ki} \left(\frac{\partial \mathcal{L}}{\partial q_i} - \sum_{j=1}^{n} \left(\frac{\partial^2 \mathcal{L}}{\partial \dot{q}_i \partial q_j} \right) \frac{dx_j}{dt} \right).$$

The matrix elements G_{ki}, and the partial derivatives of $\mathcal{L}$ are all evaluated at $(q_i(t), \dot{q}_i(t)) = \left(x(t), \frac{dx_i(t)}{dt} \right)$ so this latter equation has exactly the form of a system of n second order ODE for the n functions $x_i(t)$.

Newton's Equations Example

Consider a particle of mass m moving in $\mathbf{R}^3$ under a force that is given as usual by $F = -\nabla U$, and let the x_i be the usual Cartesian coordinates of the particle. Then its kinetic energy is $K(x, \dot{x}) = \frac{m}{2}(\dot{q}_1^2 + \dot{q}_2^2 + \dot{q}_3^2)$, its potential energy is $U(x) = U(q_1, q_2, q_3)$, $p_i = m\dot{q}_i$, and the generalized Newton's Equations are the usual Newton's Equations $\frac{dq_i}{dt} = \dot{q}_i$ and $m\frac{d\dot{q}_i}{dt} = -\frac{\partial U}{\partial q_i}$.

Exercise. Generalize the above to a sytem of n particles in $\mathbf{R}^3$ with masses $m_1, \ldots, m_n$ interacting under forces derived from a potential.

Lagrange's Equations

If we replace the momentum p_i by its definition $\frac{\partial \mathcal{L}}{\partial \dot{q}_i}$ and note then Newton's Equations $\frac{dp_i}{dt} = \frac{\partial \mathcal{L}}{\partial q_i}$ becomes the so-called Lagrange (or Euler-Lagrange) equations:

$$\frac{\partial \mathcal{L}}{\partial q_i} - \frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}_i}\right) = 0.$$

In this form these equations have a very remarkable and important interpretation that we consider next.

If $\sigma : [a, b] \to \mathcal{C}$ is a smooth path in $\mathcal{C}$, we define its **action**, $\mathcal{A}(\sigma)$, by $\mathcal{A}(\sigma) := \int_a^b \mathcal{L}(\dot{\sigma}(t))\, dt$.

Lagrange's Equations (Cont.)

Now suppose $x_s : [a, b] \to \mathcal{C}$ is a *variation* of x—i.e., a smooth 1-parameter family of paths in $\mathcal{C}$ defined for s near 0 and such that $x_0 = x$—and define $\delta x(t) := (\frac{\partial}{\partial s})_{s=0} x_s(t)$. By differentiating under the integral sign, using $\frac{\partial}{\partial t}\frac{\partial}{\partial s} = \frac{\partial}{\partial s}\frac{\partial}{\partial t}$, and integrating by parts, it is easy to see that:

$$\left.\frac{d\mathcal{A}(x_s)}{ds}\right|_{s=0} = \sum_i \int_a^b \left(\frac{\partial \mathcal{L}}{\partial q_i} - \frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}_i}\right)\right) \cdot \delta x_i(t)\, dt + \sum_i \frac{\partial \mathcal{L}}{\partial \dot{q}_i} \cdot \delta x_i(t) \Bigg]_a^b .$$

Exercise. Fill in the details!

Lagrange's Equations (Cont.)

We will come back to the second term later when we discuss Noether's Theorem, but for now let us assume that the variation $x_s(t)$ vanishes identically in s for $t = a$ and $t = b$, so that clearly $\delta x(t)$ also vanishes at the endpoints a and b and hence this second term vanishes. We will call x an *extremal* of the action functional if $\frac{d}{ds}\big|_{s=0}\mathcal{A}(x_s) = 0$ for all variations x_s vanishing at the endpoints a and b, and it follows easily from the above formula that a necessary and sufficient condition for x to be an extremal of the action is that it satisfy the Euler-Lagrange equations.

Hamilton's Equations

Since the q_i together with their conjugate momenta p_i form a local coordinate system in $T\mathcal{C}$, it is natural to try to write the Generalized Newton's Equations as a system of first order ODE. When we do so there is a major surprise—the resulting equations, have an unexpectedly simple and symmetric form.

Since K is a homogeneous quadratic form, Euler's Theorem says that $\sum_i p_i \dot{q}_i = 2K$, and hence $H = \sum_i p_i \dot{q}_i - \mathcal{L}$. It follows that $dH = \sum_i (\dot{q}_i dp_i + p_i d\dot{q}_i - \frac{\partial \mathcal{L}}{\partial q_i} dq_i - \frac{\partial \mathcal{L}}{\partial \dot{q}_i} d\dot{q}_i) = \dot{q}_i dp_i - \frac{\partial \mathcal{L}}{\partial q_i} dq_i$, or in other words, $\frac{\partial H}{\partial q_i} = -\frac{\partial \mathcal{L}}{\partial q_i}$ and $\frac{\partial H}{\partial p_i} = \dot{q}_i$. Thus Newton's Equations become $\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}$, $\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}$. These are called Hamilton's Equations.

Poisson Brackets

If F and G are two smooth real-valued functions on $T\mathcal{C}$, we define a third such function $\{F, G\}$ called their Poisson Bracket, by $\{F, G\} := \sum_i \frac{\partial F}{\partial p_i}\frac{\partial G}{\partial q_i} - \frac{\partial G}{\partial p_i}\frac{\partial F}{\partial q_i}$.

Remark. It can be shown that the function we get in this way does not depend on the choice of canonical coordinates used to define it.

Clearly, $(F, G) \mapsto \{F, G\}$ is bilnear and skew-symmetric, and in particular $\{F, F\} = 0$.

Exercise. If $F : T\mathcal{C} \to \mathbf{R}$ is a smooth function, show that its Poisson Bracket with the Hamiltonian H is its time derivative along solution curves of Hamilton's Equations. That is, if $\sigma(t)$ is any solution of Hamilton's Equations then $\frac{d}{dt}F(\sigma(t)) = \{H, F\}(\sigma(t))$.

Constants of the Motion

A function $F : T\mathcal{C} \to \mathbf{R}$ is called a **constant of the motion** if it is constant along every solution curve of Hamilton's Equations.

Exercise. Show that F is a constant of the motion if and only if $\{H, F\} = 0$. Deduce that H is always a constant of the motion. (This is the Conservation of Energy Theorem.)

Exercise. A canonical coordinate q_i is called **ignorable** if the Hamiltonian H is independent of q_i. Prove that if this is the case the the conjugate momentum p_i is a constant of the motion.

Exercise. Show that if F and G are constants of the motion then so is $\{F, G\}$.

Complete Integrability

Given real numbers $a < b$ we will denote by $H_{(a,b)}$ the set of points (x, v) in $T\mathcal{C}$ where $a < H(x, v) < b$. Since H is a constant of the motion, each $H_{(a,b)}$ is invariant under the flow given by Hamilton's Equations.

Definition. Canonical coordinates (q_i, p_i) defined in $H_{(a,b)}$ are called **action-angle variable** if all the q_i are ignorable, i.e., $H = H(p_1, \ldots, p_n)$. If action-angle variables exist in $H_{(a,b)}$ we say that the corresponding Hamiltonian system is **completely integrable** in $H_{(a,b)}$

Assume that this is the case, and choose E satisfying $a < E < b$, and $a = (a_1, \ldots, a_n)$ such that $H(a_1, \ldots, a_n) = E$, and define $\Sigma(E, a)$ to be the submanifold of $T\mathcal{C}$ where $H = E$ and $p_i = a_i$ for $i = 1, \ldots, n$.

Complete Integrability (Cont.)

Since H and the p_i are all constants of the motion, each such $\Sigma(E, a)$ will be an invariant submanifold of the Hamiltonian flow. Let's see what the flow looks like on such a manifold when looked at using action-angle variables.

If we define $\Omega_i(a) := \frac{\partial H}{\partial p_i}(a)$, then since Hamilton's Equations are $\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}$, $\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}$, on $\Sigma(E, a)$ Hamilton's Equations become $\frac{dq_i}{dt} = \Omega_i(a)$, $\frac{dp_i}{dt} = 0$, so the solutions are just $q_i(t) = q_i(t_0) + \Omega_i(a)t$. In other words, expressed in the action-angle variables, (q_i, p_i), the solution curves are just "straight line motion with constant velocity".

The Arnold-Liouville Theorem

Smooth functions on $T\mathcal{C}$ are said to be in **involution** if their Poisson brackets all vanish, Clearly, if (q_i, p_i) are action-angle variables, then the p_i are in involution and are of course functionally independent (i.e., their differentials dp_i are everywhere linearly independent). So if a Hamiltonian system is completely integrable in $H_{(a,b)}$ then there exist in $H_{(a,b)}$ $n := \dim(\mathcal{C})$ functionally independent constants of the motion $F_1, F_2, \ldots, F_n$ that are in involution. The Arnold-Liouville Theorem is the converse statement: given such $F_1, F_2, \ldots, F_n$, one can use them to construct (by quadratures) action-angle variables in $H_{(a,b)}$, and so linearize the Hamiltonian flow defined by H.

Ergodicity and the KAM Theorem.

At the other extreme from completely integrable systems is the class of Hamiltonian systems called ergodic. There is a natural measure on $T\mathcal{C}$, the so-called Liouville measure, which with respect to any set of canonical coordinates is just Lebesgue measure. Moreover, the Liouville measure is preserved by the flow defined by any Hamiltonian system in $T\mathcal{C}$. If H is a Hamiltonian on $T\mathcal{C}$, the Liouville measure induces a measure on each of its level surfaces Σ, and this induced Liouville measure is also preserved by the Hamiltonian flow restricted to Σ. The Hamiltonian flow is called ergodic if every invariant measureable set of a level surface Σ either has measure zero or its complement has measure zero.

Ergodicity and KAM (Cont.).

It is easy to see that if the Hamiltonian flow generated by H is ergodic then H is the "only" constant of the motion—meaning that any constant of the motion is a function of H. For a long time it was believed that generic Hamiltonian systems would be ergodic, and in fact there was a famous theorem of Oxtoby and Ulam in the 1930s that could be read as supporting that view. However, in the late 1950s, Kolmagorov, Arnold, and Moser proved a truly remarkable result. It follows in particular from the KAM theory that if H_s is a smooth one parameter family of Hamiltonian functions on $T\mathcal{C}$ such that H_0 is completely integrable, then H_s is not ergodic for s sufficiently close to 0. (In fact, KAM prove something much stronger and more precise.)

Example Integrable Systems

Completely integrable systems are the only ones that one can expect to analyze in detail, but one can hope their study will provide clues to the behavior of more general Hamiltonian systems. Perhaps for this reason the nineteenth and early twentieth centuries saw a flourishing theory of complete integrable systems. Below are a few important examples.

Example 1. One-dimensional Systems. The Arnold-Liouville criterion for complete integrability is the existence of $n = \dim(\mathcal{C})$ functionally independent constants of the motion. Since H is always a constant of the motion, it follows that all one-dimensional Hamiltonian systems are completely integrable.

Exercise. Show directly how to solve the general Newton's Equation by quadrature in one dimension. (Hint: Solve $\frac{1}{2}(\frac{dx}{dt})^2 + U(x) = E$ for $\frac{dx}{dt}$.)

Example Integrable Systems (Cont.)

Example 2. Uncoupled Products. Let $\mathcal{C} := \mathcal{C}_1 \oplus \cdots \oplus \mathcal{C}_k$ and identify $T\mathcal{C}$ with $T\mathcal{C}_1 \oplus \cdots \oplus T\mathcal{C}_k$. If H_i is a Hamiltonian on $T\mathcal{C}_i$, $i = 1, \ldots, k$, define a Hamiltonian $H = H_1 + \cdots + H_k$ on $T\mathcal{C}$. The resulting Hamiltonian system $(T\mathcal{C}, H)$ is called the uncoupled product of the "component" systems $(T\mathcal{C}_i, H_i)$.

Exercise. Show that uncoupled products of completely integrable systems are completely integrable, and construct action-angle variables for the product from action-angle variables for each component,

Remark. One way to analyze a complicated Hamiltonian system $(T\mathcal{C}, H)$ is to write H as a sum $H_u + H_I$ where H_u is an uncoupled Hamiltonian and H_I, the "interaction" term is in some sense small. This approach is called perturbation theory

Harmonic Oscillators

Perhaps the most basic Hamiltonian system is the so-called "simple harmonic oscillator". The configuraton space $\mathcal{C}$ is $\mathbf{R}$, so $T\mathcal{C}$ is $\mathbf{R}\times\mathbf{R}$, and the kinetic and potential energies are respectively $K(x,v) := \frac{m}{2}v^2$, and $U(x,v) := \frac{1}{2}kx^2$. Corresponding to the coordinate x on $\mathcal{C}$ we have the canonical coordinates $q(x,v) := x$ and $\dot{q}(x,v) := v$ for $T\mathcal{C}$, so $K(q,\dot{q}) := \frac{m}{2}\dot{q}^2$, $U(q,\dot{q}) := \frac{k}{2}q^2$, the Lagrangian is $\mathcal{L}(q,\dot{q}) = \frac{m}{2}\dot{q}^2 - \frac{k}{2}q^2$, and the total energy is $H = \frac{m}{2}\dot{q}^2 + \frac{k}{2}q^2$ It follows that the momentum conjugate to q is $p := \frac{\partial\mathcal{L}}{\partial\dot{q}} := m\dot{q}$, so the Hamiltonian H (which is just the total energy expressed as a function of q and p) is therefore $H(p,q) = \frac{p^2}{2m} + \frac{k}{2}q^2$. This gives the (linear!) Hamilton's Equations, $\frac{dq}{dt} = \frac{\partial H}{\partial p} = \frac{p}{m}$, $\frac{dp}{dt} = -\frac{\partial H}{\partial q} = -kq$, or $\frac{d^2q}{dt^2} = -\omega^2 q$ where $\omega := \left(\frac{k}{m}\right)^{\frac{1}{2}}$.

Harmonic Oscillators (Cont.)

Exercise Show that the solution of the simple harmonic oscillator with the initial conditions $q(0) = q_0$ and $p(0) = p_0$ is:

$$q = q_0 \cos(\omega t) + \frac{p_0}{m\omega} \sin(\omega t)$$

$$p = -m\omega q_0 \sin(\omega t) + p_0 \cos(\omega t).$$

Remark. The oscillator period is the time T it takes to go through one cycle, i.e., $\omega T = 2\pi$ so $T = 2\pi(\frac{m}{k})^{\frac{1}{2}}$.

Exercise Show that if we rescale q and t (i.e., choose new units for space and time) by $q \mapsto \sqrt{k}q$ and $t \mapsto \omega t$, then in the new canonical coordinates $H = \frac{1}{2}p^2 + \frac{1}{2}q^2$, i.e., effectively $m = k = \omega = 1$. Thus the Hamiltonian flow becomes rotation with unit angular velocity. Show that if (r, θ) are the polar coordinates associated to these rescaled q and p, then $\frac{r^2}{2}$ and θ are action-angle variables.

Harmonic Oscillators (Cont.)

It is just as easy to solve is a system of n uncoupled harmonic oscillators, Here $\mathcal{C} = \mathbf{R}^n$, $T\mathcal{C} = \mathbf{R}^n \times \mathbf{R}^n$, and $H(q_i, p_i) := \sum_{i=1}^n (\frac{p_i^2}{2m_i} + \frac{k_i}{2} q_i^2)$, giving the Hamiltonian system $\frac{dq_i}{dt} = \frac{p_i}{m_i}$ and $\frac{dp_i}{dt} = -k_i q_i$. Since the oscillators are uncoupled, the solutions are just the solutions above of the original simple harmonic operator with the n different choices of the mass m_i and spring constant k_i, giving n different angular frequencies $\omega_i := (\frac{k_i}{m_i})^{\frac{1}{2}}$.

Apparently much more complicated is the system of n **coupled** harmonic oscillators, where we replace the term $\frac{1}{2} \sum_{i=1}^n k_i q_i^2$ in the above Hamiltonian by the more general $\frac{1}{2} \sum_{i,j=1}^n k_{ij} q_i q_j$, where k_{ij} is a positive definite symmetric matrix.

Harmonic Oscillators (Cont.)

However this apparent extra complexity in the coupled case is just an illusion. While it is true that the off-diagonal "coupling constants", k_{ij} give a qualitatively different physical behavior to the system, mathematically their is no difference between the coupled and the uncoupled systems. For by the spectral theorem, by an appropriate change of variables we can simultaneously diagonalize the two quadratic forms $\frac{1}{2}\sum_{i=1}^{n}\frac{p_i^2}{m_i}$ and $\frac{1}{2}\sum_{i,j=1}^{n}k_{ij}q_iq_j$, and this effectively reduces the coupled to the uncoupled case. The new basis (in which the kinetic energy and the potential energy quadratic forms are both diagonal) is called the **normal modes** for the coupled oscillator system and the frequencies ω_i of the equivalent uncoupled oscillators are called the **eigen-frequencies** or the characteristic frequencies of the system.

Physics Near Equilibrium

It is a very remarkable fact that highly complex and fundamentally non-linear physical systems frequently appear to behave as if they were composed of uncoupled harmonic oscillators, We want to explain why that is **not** so surprising (and in fact why it should be expected) and also how we can predict the normal modes and characteristic frequencies of the oscillators.

Let us suppose that the dynamics of our system is controlled by a Lagrangian of the familiar form $\mathcal{L}(x, v) = K(x, v) - U(x)$ with $K(x, v) := \frac{1}{2} \|v\|^2$. If we choose orthonormal coordinates with respect to the kinetic energy term then it is exactly the same as the corresponding term for coupled harmonic oscillators, so what we have to see is why it is often possible to approximate the potential $U(x)$ by one having the form $\frac{1}{2} \sum_{i,j=1}^{n} k_{ij} q_i q_j$ of coupled harmonic oscillators.

Physics Near Equilibrium (Cont.)

The key to understanding this phenomenon is the fact that we normally observe the dynamical behavior of physical systems when they are evolving close to a position x_0 in $\mathcal{C}$ of **stable equilibrium**, i.e., a non-degenerate local minimum of U, so that in particular all of the first partial derivatives of U vanish at x_0 and the Hessian matrix $k_{ij} := \frac{\partial^2 U}{\partial x_i \partial x_j}(x_0)$ is positive definite. We can choose our coordinates so that x_0 is the origin, and since U is only defined up to a constant, we can also suppose $U(0) = 0$, so if we expand U in a Taylor series about x_0, we find $U(x) = \frac{1}{2}\sum_{i,j=1}^{n} k_{ij} x_i x_j + O(\|x\|^3)$.

Physics Near Equilibrium (Cont.)

Exercise. Use conservation of total energy to show that given $\epsilon > 0$ there is a $\delta > 0$ such that if the initial position and velocity have norms less than δ, then the solution will stay within ϵ of the origin for all time. Then use the Gronwall inequality to show that as ϵ goes to zero, solutions will behave more and more like solutions of the harmonic oscillator system that we get if we drop the $O(\|x\|^3)$ term.

We should still explain **why** we normally see systems when they are close to an equilibrium. The reason is that no realistic macroscopic physical system can be totally isolated; there are always interactions with its surroundings (friction) that continually reduce the total energy of the system, and unless energy is added to the system by some "forcing", the system gradually drifts toward a stable equilibrium.

Physics Near Equilibrium (Example)

Consider a pendulum of length L and mass m with its configuration described by the angle θ it makes with the vertical direction. Its kinetic energy is $\frac{1}{2}mL^2\dot{\theta}^2$ and its potential energy is $U(\theta) = mgL(1 - \cos(\theta))$. Clearly $\theta = 0$ is a stable equilibrium and the Taylor series for U there is $U(\theta) = \frac{1}{2}mgL\theta^2 + O(\theta^4)$, so the approximating harmonic oscillator has the Lagrangian $\mathcal{L}(\theta, \dot{\theta}) = \frac{1}{2}M\dot{\theta}^2 + \frac{1}{2}k\theta^2$ where $M = mL^2$ and $k = mgL$. Since $\frac{M}{k} = \frac{L}{g}$, the associated period is $2\pi\sqrt{\frac{M}{k}} = 2\pi\sqrt{\frac{L}{g}}$.

Symmetries of Lagrangians

A diffeomorphism $\Phi : \mathcal{C} \to \mathcal{C}$ induces a naturally associated diffeomorphism we will call $T\Phi : T\mathcal{C} \to T\mathcal{C}$, namely $T\Phi(x, v) = (\Phi(x), D\Phi_x(v))$.

Exercise. Show that if $\sigma : I \to \mathcal{C}$ is a smooth path in $\mathcal{C}$ and $\dot{\sigma} : I \to T\mathcal{C}$ is its natural lifting, then $T\Phi \circ \dot{\sigma}$ is the natural lifting of $\Phi \circ \sigma$. (This is the sense in which $T\Phi$ is "natural".)

Definition. If $\mathcal{L}$ is a Lagrangian function on $T\mathcal{C}$, then a diffeomorphism Φ of $\mathcal{C}$ is called a **symmetry** of $\mathcal{L}$ if $\mathcal{L} \circ T\Phi = L$.

Exercise. Show that if $\Phi : \mathcal{C} \to \mathcal{C}$ is a symmetry of a Lagrangian function $\mathcal{L}$ on $T\mathcal{C}$ then Φ preserves the action functional $\mathcal{A}_{\mathcal{L}}$ defined by $\mathcal{L}$, i.e., if $\sigma : [a, b] \to \mathcal{C}$ is any smooth path in $\mathcal{C}$, then $\mathcal{A}_{\mathcal{L}}(\Phi \circ \sigma) = \mathcal{A}_{\mathcal{L}}(\sigma)$. (Recall that $\mathcal{A}_{\mathcal{L}}(\sigma) := \int_a^b \mathcal{L}(\dot{\sigma}(t)\, dt$.)

Symmetries of Lagrangians (Cont.)

Suppose that V is an autonomous vector field in $\mathcal{C}$. With respect to coordinates x_i in $\mathcal{C}$, the corresponding ODE takes the form $\frac{dx_i}{dt} = V_i(x_1, \ldots, x_n)$, and the n functions V_i are called the components of V in this coordinate system. It is customary to write $V = \sum_i V_i \frac{\partial}{\partial x_i}$, the reason being that if g is any smooth real-valued function in $\mathcal{C}$, then along any solution curve $\sigma(t)$ of V, the chain-rule gives $\frac{dg}{dt} = \sum_i V_i \frac{\partial g}{\partial x_i}$. If we have a Lagrangian function $\mathcal{L}$ on $T\mathcal{C}$, we associate to a smooth vector field V on $\mathcal{C}$ a smooth function $\hat{V}$ on $T\mathcal{C}$ called its conjugate momentum (with respect to $\mathcal{L}$) by $\hat{V} = \sum_i V_i \frac{\partial \mathcal{L}}{\partial \dot{q}_i}$. (Note that when $V = \frac{\partial}{\partial x_i}$, this gives just the momentum p_i conjugate to q_i.)

Exercise. The definition of $\hat{V}$ seems to depend on a choice of coordinates, but show that it has the following intrinsic definition: $\hat{V}(x, v) := \langle v, V \rangle$.

Noether's Principle

Now suppose that the vector field V is complete and that the one-parameter group of diffeomorphisms ϕ_t generated by V are symmetries of $\mathcal{L}$. In that case we call V an **infinitesimal symmetry** of $\mathcal{L}$. (Note that by definition, if $p \in \mathcal{C}$ then $V(p) := \frac{d}{ds}\big|_{s=0}\phi_s(p)$.)

Emmy Noether's Symmetry Principle. *If V is an infinitesimal symmetry of the Lagrangian $\mathcal{L}$, then the conjugate momentum $\hat{V}$ is a constant of the motion of Lagrange's Equations.*

PROOF. Let $\sigma : [a,b] \to \mathcal{C}$ be a solution of the Lagrange equations. What we have to show is that $\hat{V}(\sigma(t))\big]_a^b := \hat{V}(\sigma(b)) - \hat{V}(\sigma(a)) = 0$. Let ϕ_s denote the one-parameter group of symmetries of $\mathcal{L}$ generated by V and define a variation $\sigma_s : [a,b] \to \mathcal{C}$ by $\sigma_s := \phi_s \circ \sigma$.

Noether's Principle (Cont.)

Since the ϕ_s are symmetries of $\mathcal{L}$, $\mathcal{A}_{\mathcal{L}}(\sigma_s)$ is constant in s and hence $\frac{d}{ds}\big|_{s=0}\mathcal{A}_{\mathcal{L}}(\sigma_s) = 0$. But we evaluated $\frac{d}{ds}\big|_{s=0}\mathcal{A}_{\mathcal{L}}(\sigma_s)$ earlier in the course of deriving Lagrange's equations, and we found:

$$\left.\frac{d\mathcal{A}_{\mathcal{L}}(\sigma_s)}{ds}\right|_{s=0} = \sum_i \int_a^b \left(\frac{\partial \mathcal{L}}{\partial q_i} - \frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}_i}\right)\right) \cdot \delta\sigma_i(t)\, dt + \sum_i \frac{\partial \mathcal{L}}{\partial \dot{q}_i} \cdot \delta\sigma_i(t)\Bigg]_a^b,$$

where by definition $\delta\sigma(t) := \frac{d}{ds}\big|_{s=0}\phi_s(\sigma(t))$, which is in turn, by definition of V, just $V(\sigma(t))$. Since σ is a solution of Lagrange's equations, the first sum above vanishes, leaving just $\sum_i \frac{\partial \mathcal{L}}{\partial \dot{q}_i} \cdot V_i(\sigma(t))\Big]_a^b = \hat{V}(\sigma(t))$, proving that $\hat{V}(\sigma(t))\Big]_a^b = 0$.

Particle Mechanics

We next consider briefly the mechanics of N particles P_i in $\mathbf{R}^k$. Configuration space $\mathcal{C}$ is now $(\mathbf{R}^k)^N$. If $x = (x^1, \ldots, x^N)$ is a point of $\mathcal{C}$, then $x^i = (x^i_1, \ldots, x^i_k)$ represents the position of P_i. If m_i is the mass of P_i, then $K(x, v) = \frac{1}{2} \sum_{i=1}^{N} m_i \left\| v^i \right\|^2$. We assume that for $1 \leq i < j \leq N$ there is a potential function $U_{ij} : \mathbf{R} \to \mathbf{R}$ giving the interaction between P_i and P_j. Namely, the force F_{ij} on P_i due to P_j is $U'_{ij}(\left\| x^i - x^j \right\|) r_{ij}$, where r_{ij} is the unit vector pointing in the direction from x^j to x^i. (Notice that $F_{ji} = -F_{ij}$ in accordance with Newton's Third Law of Motion.) The total force on P_i is then $F_i := \sum_j F_{ij}$, and is given by $F_i = \frac{\partial U}{\partial x^i}(x)$, where $U(x) = U(x^1, \ldots, x^N) := \sum_{i<j} U_{ij}(\|x_i - x_j\|)$. With these definitions, the generalized Newton's Equations for the Lagrangian $\mathcal{L} := K - U$ are the usual Newton's Equations $m_i \frac{d^2 x^i}{dt^2} = F_i$.

Conservation Laws

What conservation laws can we deduce from these assumptions? Because of Noether's Principle, we can rephrase that as: "What diffeomorphisms of $\mathcal{C}$ are symmetries of the Lagrangian?" We shall see that this approach leads to the so-called classical conservation laws, i.e., the fact that the components of total linear and total angular momentum are constants of the motion.

If γ is any diffeomorphism of $\mathbf{R}^k$, let Γ denote the corresponding "diagonal" diffeomorphism of $\mathcal{C} = (\mathbf{R}^k)^N$, i.e., $\Gamma(x^1, \ldots, x^N) = (\gamma(x^1), \ldots, \gamma(x^N))$. Note that if γ is a translation or a rotation then $\gamma(x^i) - \gamma(x^j) = \gamma(x^i - x^j)$ and so $\big\|\gamma(x^i) - \gamma(x^j)\big\| = \big\|x^i - x^j\big\|$.

Exercise. Show that if γ_t is any one-parameter group of Euclidean motions of $\mathbf{R}^k$ then Γ_t is a symmetry of the above N-particle Lagrangian on $T\mathcal{C}$.

Conservation Laws (Cont.)

Exercise. Show that the invariance of the N-particle Lagrangian under translations leads to "conservation of total linear momentum", i.e., the fact that each component of the total linear momentum vector $P = m_1\dot{q}^1 + \cdots + m_N\dot{q}^N$ is a constant of the motion.

Exercise. In the case, $k = 3$, the vector field V on $\mathbf{R}^3$ that generates the one parameter group of rotations about an axis x is $V(v) = x \times v$. Deduce that invariance of the N-particle Lagrangian under rotations leads to "conservation of total angular momentum", i.e., the fact that each component of the total angular momentum vector $\Omega = m_1(q^1 \times \dot{q}^1) + \cdots + m_N(q^N \times \dot{q}^N)$ is a constant of the motion.

Scaling Invariance

Multiplying a Lagrangian by some non-zero constant factor leaves the Euler-Lagrange equations and hence the extremals unchanged. This simple fact can be combined with some scaling symmetries to derive many non-obvious conclusions. Kepler's Third Law is an example.

Let $\mathcal{S}$ be an N-particle system with Lagrangian $\mathcal{L} = K(\dot{x}) - U(x)$ and let us assume that the potential function U is positively homogeneous of degree k, i.e., for all $\rho > 0$, $U(\rho x) = \rho^k U(x)$. Let's investigate how various quantities change under the transformations (corresponding to a change of length and time units) $x \to \rho x$, $t \to \sigma t$.

Scaling Invariance (Cont.)

By assumption, $U \to \rho^k U$, and clearly the velocities satisfy $\dot{x} \to (\rho/\sigma)\dot{x}$, so the kinetic energy K satisfies $K \to (\frac{\rho}{\sigma})^2 K$. Thus if $(\frac{\rho}{\sigma})^2 = \rho^k$, i.e., if $\sigma = \rho^{1-\frac{1}{2}k}$, then both K and U (and hence $\mathcal{L}$) are multiplied by the same factor, and so, by the above observation we derive the following:

Scaling Principle. *If the potential energy U is positively homogeneous of degree k and $x(t)$ is an extremal of the Lagrangian $\mathcal{L} = K - U$, then for any positive ρ, $\xi(t) = \rho\, x(\rho^{1-\frac{1}{2}k} t)$ is also an extremal.*

Scaling Invariance (Cont.)

Corollary. *If the potential U is positively homogeneous of degree k and $x(t)$ is a periodic trajectory of period T, then $\xi(t) = \rho\, x(\rho^{1-\frac{1}{2}k}t)$ is a periodic trajectory of period $\rho^{1-\frac{1}{2}k}T$.*

For an inverse square force law, $U(x) \sim 1/\left\|x\right\|$, so U is homogeneous of degree $k = -1$. Hence if we scale an orbit of period T by a factor ρ we get another closed orbit whose period T' is $T\rho^{3/2}$, that is, the square of the period scales as the cube of the size of the orbit. This is Kepler's Third Law of Planetary Motion.

Exercise. Suppose a particle attracted to the origin by a force $F(x) = -kx$ (Hookes Law). This is a central force with potential $U(x) = \frac{1}{2}k\left\|x\right\|^2$, i.e., an harmonic oscillator. Show that all of its closed orbits have the same period.

The Two Body Problem

A system of two particles in $\mathbf{R}^3$ is referred to as a "two-body problem". Newton's Equations are a system of six non-linear, second order ODE (for the three components of $x^1(t)$ and the three components of $x^2(t)$). It is a highly remarkable fact that we can reduce this system of ODE to a much simpler form, eliminating many of the variables, and then explicitly solving the simplified system by quadratures.

This is not an accident—the explicit integration rests on the six classical constants of the motion, the three components of total linear momentum, and the three components of total angular momentum. And these integrals in turn have their origin in the six parameter Euclidean group of symmetries of the problem.

The Two Body Problem (Cont.)

We make use of conservation of total linear momentum by choosing center of mass coordinates, i.e., we force the center of mass to be at the origin by the constraint equation $m_1 x^1 + m_2 x^2 = 0$.

This cuts out a three-dimensional subspace $\mathcal{C}'$ of the six dimensional configuration space $\mathcal{C} = (\mathbf{R}^3)^2$, and we define so-called relative position coordinates $X = (X_1, X_2, X_3)$ on $\mathcal{C}'$ by $X = x^1 - x^2$

It is easy to check that on $\mathcal{C}'$, $x^1 = \frac{m_2}{m_1+m_2} X$ and $x^2 = -\frac{m_1}{m_1+m_2} X$, so if we define $\dot{X} := \dot{x}^1 - \dot{x}^2$, the kinetic energy is $K = \frac{1}{2} m \left\| \dot{X} \right\|^2$, where m is the so-called **reduced mass**: $m := \frac{m_1 m_2}{m_1+m_2}$, and the Lagrangian is $\mathcal{L}(X, \dot{X}) = \frac{1}{2} m \left\| \dot{X} \right\|^2 + U(\|X\|)$ on $\mathcal{C}'$.

The Two Body Problem (Cont.)

If we can find the trajectories in $\mathcal{C}'$ for this Lagrangian, then the above formulas for x^1 and x^2 in terms of X give the trajectories of the original Lagrangian, so this is a true reduction of the problem.

We now use conservation of angular momentum to reduce the problem further, namely to motion in a plane Π.

Note that for the reduced problem, the linear momentum is $p = m\frac{dX}{dt}$ so p is parallel to $\frac{dX}{dt}$ and hence $\frac{dX}{dt} \times p = 0$. Now the angular momentum is $J = X \times p$.

Since the force $-\nabla U$ is in the radial direction X, $\frac{dp}{dt}$ is parallel to X, so $X \times \frac{dp}{dt}$ also vanishes and it follows that $\frac{dJ}{dt} = \frac{dX}{dt} \times p + X \times \frac{dp}{dt} = 0$, proving that J is a constant of the motion (as we already knew).

The Two Body Problem (Cont.)

Thus the plane Π normal to J is fixed in space, and of course the radius vector X lies in this plane.

With respect to polar coordinates (r, θ) in Π, the Lagrangian takes the form $L(r, \theta, \dot{r}, \dot{\theta}) = \frac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}^2) - U(r)$, so the Euler-Lagrange equations are now:

$$m\frac{d^2r}{dt^2} = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{r}}\right) = \frac{\partial L}{\partial r} = mr\dot{\theta}^2 - \frac{\partial U}{\partial r},$$

$$\frac{d}{dt}(mr^2\dot{\theta}) = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}}\right) = \frac{\partial L}{\partial \theta} = 0.$$

The Two Body Problem (Cont.)

The second of these equations integrates immediately, and seems to give yet another constant of the motion: $mr^2\dot{\theta} = C$. However, it is easy to verify that $mr^2\dot{\theta}$ is just the magnitude $\|J\|$ of the total angular momentum vector, so this is really not an independent new conservation law. In fact, note that choosing the plane Π only fixed the **direction** of J (two real parameters) and the equation $mr^2\dot{\theta} = \|J\|$ gives the third remaining relation that follows from the conservation of angular momentum.

The Two Body Problem (Cont.)

Since m is a constant, we can also write the second of the Euler-Lagrange equations as $\frac{d}{dt}(r^2\dot{\theta}) = 0$, and as such it has an interesting geometric interpretation.

In fact, $\frac{1}{2}r^2\frac{d\theta}{dt}$ is clearly the rate at which the ray from the origin to the point (r, θ) is sweeping out area, so the second equation is just Kepler's Second Law of Planetary Motion (equal areas in equal times).

An important thing to note is that Kepler's Second Law holds for **any** central force, not just for a $\frac{1}{r^2}$ force, and also that it is really just another way of stating the constancy of the magnitude of total angular momentum.

The Two Body Problem (Cont.)

Finally, the integrated form of the second equation allow us to write the total energy $E = K + U$ as

$$\begin{aligned} E &= \frac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}^2) + U(r) \\ &= \frac{1}{2}m\dot{r}^2 + \frac{\|J\|^2}{2mr^2} + U(r). \end{aligned}$$

If we solve this for $\dot{r} = \frac{dr}{dt}$, then invert and integrate, we get an explicit formula for t as a function of r:

$$t = \int_{r_0}^{r} \frac{dr}{\sqrt{\frac{2}{m}[E - U(r)] - \left(\frac{\|J\|}{mr}\right)^2}}.$$

The Two Body Problem (Cont.)

On the other hand, from $d\theta = \frac{\|J\|}{r^2} dt$ and the preceding formula for dt we find:

$$\theta = \int_{r_0}^{r} \frac{\|J\| \; dr}{r^2 \sqrt{2m[E - U(r)] - \left(\frac{\|J\|}{r}\right)^2}}.$$

These latter two equations give a fairly explicit answer of the problem of finding the motion of a single particle under the action of a general central force, and hence also of the problem of the motion of a closed system of two particles under a central force. However, for a $\frac{1}{r}$ potential, say $U(r) = -\frac{k}{r}$ we can give a more satisfying closed form solution.

The Two Body Problem (Cont.)

In fact in this case the right hand side of the above formula for θ reduces to a well-known elementary integral and we find:

$$\theta = \cos^{-1}\left(\frac{\frac{\|J\|}{r} + \frac{km}{\|J\|}}{\sqrt{2mE + \frac{k^2m^2}{\|J\|^2}}}\right) + \theta_0,$$

so that

$$r = \frac{p}{1 + e\cos(\theta - \theta_0)},$$

where $p = \frac{\|J\|^2}{km}$ and $e = \sqrt{1 + \frac{2E\|J\|^2}{k^2m}}$, the standard polar coordinate form of a conic section with eccentricity e and semi latus rectum p.

Lattice Models

Finally, we consider what is called a one-dimensional lattice of N oscillators with nearest neighbor interactions and zero boundary conditions.

We imagine a "string" consisting of particles that are positioned along the x-axis from 0 to its length ℓ. The N particles have equilibrium positions $p_i = ih$, $i = 0, \ldots, N-1$, where $h = \ell/(N-1)$ is the lattice spacing, so their positions at time t are $X_i(t) = p_i + x_i(t)$, (where the x_i represent the displacements of the oscillators from equilibrium). The force attracting any oscillator to one of its neighbors is taken as $k(\delta + \alpha\delta^2)$, δ denoting the "strain", i.e., the deviation of the distance separating these two oscillators from their equilibrium separation h. (Note that when $\alpha = 0$ this is just the Hooke's law force with spring constant k.)

Lattice Models (Cont.)

The force acting on the i-th oscillator due to its right neighbor is $F(x)_i^+ = k[(x_{i+1} - x_i) + \alpha((x_{i+1} - x_i)^2]$, while the force acting on the it due to its left neighbor is $F(x)_i^- = k[(x_{i-1} - x_i) - \alpha((x_{i-1} - x_i)^2]$. Thus the total force acting on the i-th oscillator will be the sum of these two forces, namely: $F(x)_i = k(x_{i+1} + x_{i-1} - 2x_i)[1 + \alpha(x_{i+1} - x_{i-1})]$, and assuming that all of the oscillators have the same mass, m, Newton's equations of motion read:

$$m\ddot{x}_i = k(x_{i+1} + x_{i-1} - 2x_i)[1 + \alpha(x_{i+1} - x_{i-1})],$$

with the boundary conditions $x_0(t) = x_{N-1}(t) = 0$.

Lattice Models (Cont.)

It will be convenient to rewrite Newton's equations in terms of parameters that refer more directly to the original string that we are trying to model. Namely, if ρ denotes the density of the string, then $m = \rho h$, while if κ denotes the Young's modulus for the string (i.e., the spring constant for a piece of unit length), then $k = \kappa/h$ will be the spring constant for a piece of length h. Defining $c = \sqrt{\kappa/\rho}$ we can now rewrite Newton's equations as:

(LE)

$$\ddot{x}_i = c^2 \left(\frac{x_{i+1} + x_{i-1} - 2x_i}{h^2} \right) [1 + \alpha(x_{i+1} - x_{i-1})],$$

and in this form we shall refer to them as the FPU Lattice Equations.

Lattice Models (Cont.)

We can now "pass to the continuum limit"; i.e., by letting N tend to infinity (so h tends to zero) we can attempt to derive a PDE for the function $u(x,t)$ that measures the displacement at time t of the particle of string with equilibrium position x. We shall leave the nonlinear case for later, and here restrict our attention to the linear case, $\alpha = 0$. If we take $x = p_i$, then by definition $u(x,t) = x_i(t)$, and since $p_i + h = p_{i+1}$ while $p_i - h = p_{i-1}$, with $\alpha = 0$, the latter form of Newton's equations gives:

$$u_{tt}(x,t) = c^2 \frac{u(x+h,t) + u(x-h,t) - 2u(x,t)}{h^2}.$$

Lattice Models (Cont.)

By Taylor's formula:

$$f(x \pm h) =$$
$$f(x) \pm h f'(x) + \frac{h^2}{2!} f''(x) \pm \frac{h^3}{3!} f'''(x) + \frac{h^4}{4!} f''''(x) + O(h^5),$$

and taking $f(x) = u(x,t)$ gives:

$$\frac{u(x+h,t)+u(x-h,t)-2u(x,t)}{h^2} =$$
$$u_{xx}(x,t) + \left(\frac{h^2}{12}\right) u_{xxxx}(x,t) + O(h^4),$$

so letting $h \to 0$, we find $u_{tt} = c^2 u_{xx}$, i.e., u satisfies the linear wave equation, with propagation speed c and the boundary conditions $u(0,t) = u(\ell,t) = 0$, and initial conditions $u_t(x,0) = 0$, $u(x,0) = u_0(x)$.

Appendix A

Smoothness with Respect to Initial Conditions

Suppose that V is a C^1 vector field on $\mathbf{R}^n$ and assume that the maximal solution σ_p of $\frac{dx}{dt} = V(x)$ is defined on $I = [a, b]$. For each $x \in \mathbf{R}^n$, the differential of V at x is a linear map $DV_x : \mathbf{R}^n \to \mathbf{R}^n$, and it is continuous in x since V is C^1. Thus $A(t) = DV_{\sigma_p(t)}$ defines a continuous map $A : I \to L(\mathbf{R}^n)$. The differential equation $\frac{dx}{dt} = A(t)x$ is an example of a non-autonomous linear ODE, studied in Part I. It is called the **variational equation** associated to the solution σ. By the general theory of such equations developed in Part I, we know that for each ξ in $\mathbf{R}^n$, the variational equation will have a unique solution $u(t, \xi)$ defined for $t \in I$, and satisfying the initial condition $u(t_0, \xi) = \xi$. For each t in I, the map $\xi \mapsto u(t, \xi)$ is a linear map of $\mathbf{R}^n$ to itself that we will denote by $\delta\sigma_p(t)$. What we are going to see next is that the map $(t, p) \mapsto \sigma_p(t)$ is C^1 and that $\delta\sigma_p(t)$ is the differential at p of the map $q \mapsto \sigma_q(t)$ of $\mathbf{R}^n$ to itself.

(Note that the derivative of $\sigma_p(t)$ with respect to t obviously exists and is continuous since $\sigma_p(t)$ satisfies $\sigma_p'(t) = V(\sigma_p(t))$.)

Exercise Check that if $q \mapsto \sigma_q(t)$ is indeed differentiable at p then its differential must in fact be $\delta\sigma_p(t)$. Hint: Calculate the differential of both sides of the differential equation with respect to p to see that $D\sigma_p(t)(\xi)$ satisfies the variational equation. On the right side of the equation use the chain rule and on the left side interchange the order of differentiation.

Recall that (by definition of the differential of a mapping) in order to prove that $q \mapsto \sigma_q(t)$ is differentiable at p, and that $u(t, \xi) = \delta\sigma_p(t)(\xi)$ is its differential at p in the direction ξ, what we need to show is that if $g(t) := \|(\sigma_{p+\xi}(t) - \sigma_p(t)) - u(t, \xi)\|$ then $\frac{1}{\|\xi\|} g(t)$ goes to zero with $\|\xi\|$. What we will show is that there are fixed positive constants C and M such that for any

positive ϵ there exists a δ so that $g(t) < C\epsilon \|\xi\| e^{Mt}$ provided $\|\xi\| < \delta$, which clearly implies that $\frac{1}{\|\xi\|} g(t)$ goes to zero with $\|\xi\|$, uniformly in t. To prove the latter estimate, it will suffice by Gronwall's inequality to show that $g(t) < C\epsilon \|\xi\| + M \int_0^t g(s)\, ds$.

Exercise Derive this estimate. Hint:
$\sigma_{p+\xi}(t) = p + \xi + \int_0^t V(\sigma_{p+\xi}(s))\, ds$ and $\sigma_p(t) = p + \int_0^t V(\sigma_p(s))\, ds$, while $u(t, \xi) = \xi + \int_0^t DV_{\sigma_p(s)} u(s, \xi)\, ds$. Taylor's Theorem with Remainder gives $V(q + x) - V(q) = DV_q(x) + \|x\| r(q, x)$ where $\|r(q, x)\|$ goes to zero with x, uniformly for q in some compact set. Take $q = \sigma_p(s)$ and $x = \sigma_{p+\xi}(s) - \sigma_p(s)$ and verify that $g(t) =$
$\|\xi\| \int_0^t \rho(\sigma_p(s), \sigma_{p+\xi}(s) - \sigma_p(s))\, ds + \int_0^t DV_{\sigma_p(s)} g(s)\, ds$.
Now choose $M = \sup_{s \in I} \left\| DV_{\sigma_p(s)} \right\|$ and recall that from the theorem on continuity with respect to initial

conditions we know that $\|\sigma_{p+\xi}(s) - \sigma_p(s)\| < \|\xi\|\, e^{Ks}$. The rest is easy, and we have now proved the case $r = 1$ of the following theorem.

Smoothness w.r.t. Initial Conditions.
Let V be a C^r vector field on $\mathbf{R}^n$, $r \geq 1$, and let $\sigma_p(t)$ denote the maximal solution curve of $\frac{dx}{dt} = V(x)$ with initial condition p. Then $(p, t) \mapsto \sigma_p(t)$ is a C^r map.

Exercise Prove the general case by induction on r. Hint: As we saw, the first order partial derivatives are solutions of an ODE whose right hand side is of class C^{r-1}.