Physical Simulations

Computer Science and Engineering
University of California San Diego

Albert Chern

Last update: April 6, 2023

Contents

I Principles

1. Dimensional Analysis

In computational physics, mechanics and graphics, we generate computer simulated data that mimic what one would observe in the natural world. This process allows scientists and engineers to make predictions or to virtually conduct experiments, and visual effect artists to create believable animations. The process takes several stages:

1. **Modeling** *Turning a physical phenomenon into a set mathematical equations.* The goal is to have the phenomenon described quantitatively through a clever choice of variables, whose values are specified by some equations they satisfy. This mathematical representation is derived based on known physical laws or based on educated guesses based on the observations.
2. **Analysis** *Making sense of the mathematical equations.* The goal is to get a general idea how the solution behaves for a given equation without really solving it. It may answer questions such as whether the solution of an equation exists (or else the model breaks down) and whether the solution is robust and unique (to have any scientific prediction power).
3. **Computation** *Analytically or numerically approximate the solution.* The goal is to generate the solutions to the equations accurately and efficiently.

In this chapter, we describe a powerful tool for modeling and analysis called **dimensional analysis**. This tool will help us find the *dimensionless quantities*, reduce the number of parameters, and discover the associated scaling symmetries within the equation.

1.1 Dimensions and Units

A (physical) **dimension** is the collection of measurements of one type of physical quantity. A **unit** is a way to assign a number for each measurement in the dimension.

Geometrically, a dimension is a line with a distinguished point called zero (origin); *i.e.* it is a one-dimensional vector space equipped with addition and scaling. Each

physical measurement is a point on the line. A unit is a choice of basis for the dimension (just a point that is not the origin on the line). With a unit, every other point on the line can be written as a unique scalar times the unit.

For example, length is a dimension and meters and feet are two possible units for length.

Two different dimensions can be multiplied together or be taken quotient to produce another dimension. For example, length divided by time is the dimension of speed. What happens if we divide length by length? What we get is a special dimension called the **dimensionless** dimension. The dimensionless dimension is just the real number line $\mathbb{R}$ whose elements are called **numbers**, and it has a canonical unit of $1 \in \mathbb{R}$, *i.e.* it doesn't have a choice for a different unit.

We will label each dimension by a symbol using capital letter in the sans serif font. For example the dimension of mass is M. Another notation for a dimension is $[x]$: if x is a physical quantity, then $[x]$ is the dimension x lies in. For example, $[5\,\mathrm{kg}] = [2\,\mathrm{kg}] = \mathsf{M}$. The dimensionless dimension is denoted by $[1]$. With multiplication being the operator and the dimensionless dimension being the identity, the set of all dimensions form an *Abelian group*.[1]

In classical physics, the group of dimensions is a *finitely generated Abelian group*. The conventional seven **primary (base) dimensions** are:

Table 1.1

Primary dimension	Symbol	SI Unit
Mass	M	kg (kilogram)
Length	L	m (meter)
Time	T	s (second)
Temperature	Θ	K (Kelvin)
Electric current	I	A (Ampere)
Amount of light	C	cd (candela)
Amount of matter	N	mol (mole)

Other dimensions can be expressed as multiplicative combinations of the primary dimensions. For example, force has the dimension $[f] = \mathsf{MLT}^{-2}$. In general, every dimension in classical physics takes the form

$$\mathsf{M}^{n_1}\mathsf{L}^{n_2}\mathsf{T}^{n_3}\Theta^{n_4}\mathsf{I}^{n_5}\mathsf{C}^{n_6}\mathsf{N}^{n_7}, \tag{1.1}$$

where $n_1, \ldots, n_7$ are integers and sometimes more generally real numbers. If we just look at the exponent $(n_1, \ldots, n_7)$, the multiplicative group of dimensions become just the (additive) 7D vector space $\{(n_1, \ldots, n_7) \in \mathbb{R}^7\}$. For example, $[f] = \mathsf{MLT}^{-2}$ corresponds to the vector $(1, 1, -2, 0, 0, 0, 0) \in \mathbb{R}^7$. The dimensionless dimension $[1]$

[1] A group $(G, \cdot)$ is a set G with an operator $\cdot\colon G \times G \to G$ so that
- associative: $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ for all $a, b, c \in G$;
- identity: $\exists 1 \in G$ such that $1 \cdot a = a \cdot 1$ for all $a \in G$;
- inverse: for each $a \in G$ there exists $b \in G$ so that $a \cdot b = b \cdot a = 1$.

If in addition we have commutativity $a \cdot b = b \cdot a$ for all $a, b \in G$, then we say the group is an Abelian group.

corresponds to $(0,0,0,0,0,0,0) \in \mathbb{R}^7$. Changing a different set of primary dimensions amount to changing basis for $\mathbb{R}^7$.

1.2 Dimensional Homogeneity

In an equation, or in an inequality relation, every additive term must have the same dimension.

■ **Example 1.1 — Bernoulli equation.** The Bernoulli equation for incompressible and irrotational fluid flow is

$$\frac{1}{2}\rho v^2 + p + \rho g z = C \tag{1.2}$$

where ρ is the density (mass per volume) $[\rho] = \mathsf{ML}^{-3}$, v the speed (distance per time) $[v] = \mathsf{LT}^{-1}$, p the pressure (force per area) $[p] = \mathsf{MLT}^{-2}\mathsf{L}^{-2}$, g the gravitational acceleration $[g] = \mathsf{LT}^{-2}$, z the depth $[z] = \mathsf{L}$, and C a constant.

Let us check that the Bernoulli equation satisfies dimensional homogeneity. The first term has dimension $[\frac{1}{2}\rho v^2] = [\rho][v]^2 = \mathsf{ML}^{-3}(\mathsf{LT}^{-1})^2 = \mathsf{ML}^{-1}\mathsf{T}^{-2}$. The second term $[p] = \mathsf{ML}^{-1}\mathsf{T}^{-2}$ also has the same dimension. The third term $[\rho][g][z] = \mathsf{ML}^{-3}\mathsf{LT}^{-2}\mathsf{L} = \mathsf{ML}^{-1}\mathsf{T}^{-2}$ also shares the same dimension. The constant C must also has the same dimension. ■

■ **Example 1.2 — Sobolev inequality.** If a function u defined on an n-dimensional bounded domain Ω has an L^p-integrable derivative (*i.e.* $\|\nabla u\|_p = (\int_\Omega |\nabla u|^p \, dV)^{1/p} < \infty$), then u must be L^q integrable. Can you guess what q is in terms of p, n?

The statement is commonly written as an inequality: there exists a dimensionless constant C such that

$$\|u\|_q \leq C\|\nabla u\|_p + \left(\begin{smallmatrix}\text{sometimes there}\\ \text{are more terms}\end{smallmatrix}\right) \tag{1.3}$$

for all functions u on Ω. We can use dimensional homogeneity to determine q. Note that taking derivative divides the function by length dimension, and integrating over an n-dimension multiplies L^n.

The left-hand side of the inequality (1.3) is

$$\|u\|_q = \left(\int_\Omega |u|^q \, dV\right)^{1/q} \quad \Longrightarrow \quad [\|u\|_q] = ([u]^q \mathsf{L}^n)^{1/q} = [u]\mathsf{L}^{n/q}. \tag{1.4}$$

The right-hand side is

$$\|\nabla u\|_p = \left(\int_\Omega |\nabla u|^p \, dV\right)^{1/p} \quad \Longrightarrow \quad [\|\nabla u\|_p] = \left(([u]\mathsf{L}^{-1})^p \mathsf{L}^n\right)^{1/p} = [u]\mathsf{L}^{\frac{n-p}{p}}. \tag{1.5}$$

Now, by equating the dimensions, we find

$$\frac{n}{q} = \frac{n-p}{p} \quad \Longrightarrow \quad q = \frac{np}{n-p}. \tag{1.6}$$

Note that this is not a proof of the Sobolev inequality (1.3). But, should any inequality like (1.3) hold, q must be given by (1.5). ∎

1.3 Dimensionless Equations

Since each term in an equation has the same dimension, we can divide all terms of the equation by a constant in the same dimension. The resulting equation becomes dimensionless; that is, every term is dimensionless.

After turning an equation into a dimensionless one, the key dimensionless parameters appear. Usually, the number of dimensionless parameters will be much fewer than the number original parameters. Hence the process helps us understand the relationship between the key variables, and it also reduces the number of parameters in the problem.

∎ **Example 1.3 — Projectile motion.** Suppose at time 0 an object has an initial height z_0 and initial vertical velocity v_0. Under a uniform gravity g, the height z of the object at time t will be

$$z = z_0 + v_0 t - \frac{1}{2} g t^2, \tag{1.7}$$

which is the solution to the differential equation

$$\frac{d^2 z(t)}{dt^2} = -g, \quad z(0) = z_0, \quad \frac{dz}{dt}(0) = v_0. \tag{1.8}$$

The problem has a dependent variable z, an independent variable t, and three additional parameters z_0, v_0, g. The *parameters* z_0, v_0, g have dimensions $[z_0] = \mathsf{L}$, $[v_0] = \mathsf{LT}^{-1}$, $[g] = \mathsf{LT}^{-2}$, influencing the *variables* z, t whose dimensions are $[z] = \mathsf{L}$, $[t] = \mathsf{T}$. At the level of dimension, we can see how the dependent variables are functions of independent parameters:

$$[z] = [z_0], \quad [t] = \frac{[z_0]}{[v_0]} \left(= \frac{[v_0]}{[g]}\right). \tag{1.9}$$

(For a large system this is harder to find, and the solution is not unique. We will discuss how to solve this problem systematically later.) Using the information of (1.9), we introduce a change of variable:

$$z^* := \frac{z}{z_0}, \quad t^* := \frac{t v_0}{z_0}. \tag{1.10}$$

The new variables are dimensionless. Substituting them into (1.8) yields the

dimensionless equation

$$\frac{d^2 z^*(t^*)}{dt^{*2}} = \underbrace{-\frac{g z_0}{v_0^2}}_{=-\frac{1}{\mathrm{Fr}^2}}, \quad z^*(0) = 1, \quad \frac{dz^*}{dt^*}(0) = 1. \tag{1.11}$$

Now it is revealed that there is only one dimensionless parameter gz_0/v_0^2. The dimensionless number

$$\mathrm{Fr} = \frac{v_0}{\sqrt{g z_0}} \tag{1.12}$$

is called the **Froude number**, appearing in many hydrodynamical engineering problems involving gravity and inertia (Froude number is the ratio of inertia and gravitational effects). The solution (1.7) in dimensionless form is given by

$$z^* = 1 + t^* - \frac{1}{2\mathrm{Fr}^2} t^{*2}. \tag{1.13}$$

The maximum height of the projectile is $z^*_{\max} = 1 + \frac{\mathrm{Fr}^2}{2}$. ∎

∎ **Example 1.4 — Navier–Stokes equation.** The Navier–Stokes for the flow velocity field $\mathbf{u}(t, \mathbf{x})$ and pressure $p(t, \mathbf{x})$ of an incompressible viscous fluid is given by

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} = -\frac{\nabla p}{\rho} + \frac{\mu}{\rho} \nabla \cdot \nabla \mathbf{u}, \quad \nabla \cdot \mathbf{u} = 0. \tag{1.14}$$

Here the constant ρ is the density and the constant μ is the viscosity. The dimensions of each variable and parameter are

- $[t] = \mathsf{T}$, $[\frac{\partial}{\partial t}] = \mathsf{T}^{-1}$
- $[\mathbf{x}] = \mathsf{L}$, $[\nabla] = \mathsf{L}^{-1}$, $[\nabla \cdot \nabla] = \mathsf{L}^{-2}$.
- $[\mathbf{u}] = \mathsf{L}\mathsf{T}^{-1}$ (distance per time).
- $[p] = \mathsf{M}\mathsf{L}^{-1}\mathsf{T}^{-2}$ (force per area).
- $[\rho] = \mathsf{M}\mathsf{L}^{-3}$ (mass per volume).
- $[\mu] = \mathsf{M}\mathsf{L}^{-1}\mathsf{T}^{-1}$ (stress per speed gradient, *i.e.* (force per area) per (speed per length))

One can check that every additive term has the same dimension of $\mathsf{L}\mathsf{T}^{-2}$. Now, let L be any constant with $[L] = \mathsf{L}$ called a characteristic length, and let U be any any constant with $[U] = \mathsf{L}\mathsf{T}^{-1}$ called a characteristic speed. For example, in a experimental setup of flow passing around an obstacle, we can set L as the diameter of the obstacle and U be the flow speed at the inlet. Using the new constants, perform the changes of variables to get dimensionless variables and operators:

- $t^* := \frac{U}{L} t$, $\frac{\partial}{\partial t^*} = \frac{L}{U} \frac{\partial}{\partial t}$.
- $\mathbf{x}^* := \frac{\mathbf{x}}{L}$, $\nabla^* = L\nabla$, $\nabla^* \cdot \nabla^* = L^2 \nabla \cdot \nabla$.
- $\mathbf{u}^* := \frac{\mathbf{u}}{U}$.
- $p^* := \frac{p}{\rho U^2}$.

Then the Navier–Stokes equation (1.14) becomes

$$\frac{\partial \mathbf{u}^*}{\partial t^*} + \mathbf{u}^* \cdot \nabla^* \mathbf{u}^* = -\nabla^* p^* + \underbrace{\frac{\mu}{\rho UL}}_{=\frac{1}{\mathrm{Re}}} \nabla^* \cdot \nabla^* \mathbf{u}^*, \quad \nabla^* \cdot \mathbf{u}^* = 0 \tag{1.15}$$

The dimensionless parameter $\mathrm{Re} = \frac{\rho UL}{\mu}$ is called the **Reynolds number**. As we now see that even though the original equation depends on many parameters including length scale L and a characteristic speed U, density ρ and viscosity μ, the final behavior of the flow only depends on one parameter Re. Reynolds discovered that for small Reynolds number (less than 100) the flow tends to be laminar, and flor large Reynolds number (greater than 1,000 or 10,000) the flow tends to be turbulent. ∎

Reduction of parameters

A major advantage of dimensionless equation is the reduction of parameters. Suppose we need to generate the projectile solutions for all possible combinations of parameters. Taking the original problem directly, we would have to enumerate combinations of 3 parameters. In the dimensionless form, we only need to deal with one parameter. Similarly, for generating simulated results of fluid flow, we only need to report it per Reynolds number, instead of enumerating all combinations of density, length, speed and viscosity.

Similarity between systems

Another advantage of dimensionless form is that we can use it to map a set of physical parameters to another one to emulate inaccessible environments. For example, suppose we want to shoot a movie scene with a falling object on Mars, which has a different gravity from Earth's. We can calculate what the Froude number of the scene, replace the gravity by Earth's gravity and adjust either v_0 or z_0 to keep the Froude number the same. Then we can shoot the scene on Earth with the adjusted v_0, z_0. Finally, one reinterpretes the footage by rescaling z and t.

For aerodynamics experiments such as testing an airplane, one builds a smaller model of the airplane and place it in a wind tunnel. The density, wind speed are set to compensate a smaller length scale of the model in such a way that the Reynolds number is the same as that of an actual flying airplane.

Extrapolation of data

In some scenario we have already have data for the parameters and variables in a certain range. We can use their dimension relation to obtain a scaling law to obtain a trend of how the data would continue beyond the known data.

In most experiments, to save time and money, tests are performed on the **scaled model** rather than the full-scale **prototype**. This is possible thanks to the dimensionless representation of the system.

After nondimensionalization, the dimensionless equation is telling us a relation between the dimensionless dependent variables (output) Π_1 and the dimensionless dependent variables and parameters (input) $\Pi_2, \ldots, \Pi_n$ as some generic function

$\Pi_1 = f(\Pi_2, \ldots, \Pi_n)$. *If the sets of dimensionless numbers* $(\Pi_2, \ldots, \Pi_n)$ *for the scaled model and for the prototype are the same, then the output number* Π_1 *for the scaled model and for the prototype must also be the same.* We first *encode* the original full-scale prototype input as $(\Pi_2, \ldots, \Pi_n)$, solve the equation or perform the experiment using the scaled-model representative, and *decode* the output Π_1 back to the full-scale prototype.

The primary purposes of dimensional analysis are

- To generate dimensionless parameters that help in designing the experiments or numerical simulators and in reporting the results.
- To obtain scaling laws to map the results from the scaled model to a prediction of results of the prototype.
- To predict trends in the relationship between parameters and variables.

1.4 Buckingham Π Theorem

In this section, we study the main technique to find the dimensionless parameters.

Theorem 1.1 — Buckingham Π Theorem. If a physical equation $f(q_1, \ldots, q_n) = 0$ for n variables $q_1, \ldots, q_n$ involving in total k independent physical dimensions, then the equation can be restated as $F(\Pi_1, \ldots, \Pi_p) = 0$ for $p = n - k$ dimensionless parameters $\Pi_1, \ldots, \Pi_p$.

The proof for the theorem will be clear by following the construction of the dimensionless parameters.

Let $\mathsf{B}_1, \ldots, \mathsf{B}_\ell$ be the base physical dimensions (*e.g.* Table 1.1). Write the dimension $[q_i]$ of each q_i in terms of the base dimensions:

$$[q_i] = \mathsf{B}_1^{m_{1i}} \mathsf{B}_2^{m_{2i}} \cdots \mathsf{B}_\ell^{m_{\ell i}} \tag{1.16}$$

giving rise to the ℓ-by-n **dimensional matrix**

$$\mathbf{M} \coloneqq \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1n} \\ \vdots & \vdots & & \vdots \\ m_{\ell 1} & m_{\ell 2} & \cdots & m_{\ell n} \end{bmatrix}. \tag{1.17}$$

Any combination of the n variables (with exponents $a_1, \ldots, a_n \in \mathbb{R}$) has a dimension

$$[q_1]^{a_1} \cdots [q_n]^{a_n} = \mathsf{B}_1^{b_1} \mathsf{B}_2^{b_2} \cdots \mathsf{B}_\ell^{b_\ell}, \quad \text{where} \quad \begin{bmatrix} b_1 \\ \vdots \\ b_\ell \end{bmatrix} = \mathbf{M} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}. \tag{1.18}$$

The precise meaning for the statement that the equation "involves k independent physical dimensions" is that the rank of M is k. Each dimensionless parameter Π_j will be constructed as

$$\Pi_j = q_1^{a_{1j}} \cdots q_n^{a_{nj}}, \quad \text{or} \quad \log(\Pi_j) = \begin{bmatrix} \log(q_1) & \cdots & \log(q_n) \end{bmatrix} \begin{bmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{bmatrix}, \tag{1.19}$$

so that the the resulting dimension exponents $\mathbf{M}\begin{bmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{bmatrix} = \mathbf{0}$. Indeed, by the rank-nullity theorem, there are precisely $p = (n-k)$ linearly independent null vectors of $\mathbf{M}$, each of which corresponds to a dimensionless parameter:

$$\begin{bmatrix} \log(\Pi_1) & \cdots & \log(\Pi_p) \end{bmatrix} = \begin{bmatrix} \log(q_1) & \cdots & \log(q_n) \end{bmatrix} \underbrace{\begin{bmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{np} \end{bmatrix}}_{\mathbf{A}}. \tag{1.20}$$

The n-by-p matrix $\mathbf{A}$ has linearly independent columns (full column rank); *i.e.* it is surjective. In particular, for every $(\log(\Pi_1), \ldots, \log(\Pi_p))$ we can linearly reconstruct some $(\log(q_1), \ldots, \log(q_n))$. Removing the logarithm, we can re-express each $q_1, \ldots, q_n$ by some $\Pi_1^{c_1} \cdots \Pi_p^{c_p}$. Substitute them into $f(q_1, \ldots, q_n) = 0$ to obtain the equation $F(\Pi_1, \ldots, \Pi_p) = 0$.

■ **Example 1.5 — Aerodynamic drag of a car.** A moving car will experience some aerodynamic drag. In this system the variations are the length of the car L, the speed of the car V, the density of the air ρ, the viscosity of the air μ, and the drag force F_D. We postulate that these quantities obey some unknown functional relation. For these 5 quantities (L, V, ρ, μ, F_D) the dimensions are

$$[L] = \mathsf{M}^0\mathsf{L}^1\mathsf{T}^0, \quad [V] = \mathsf{M}^0\mathsf{L}^1\mathsf{T}^{-1}, \quad [\rho] = \mathsf{M}^1\mathsf{L}^{-3}\mathsf{T}^0, \tag{1.21}$$

$$[\mu] = \mathsf{M}^1\mathsf{L}^{-1}\mathsf{T}^{-1}, \quad [F_D] = \mathsf{M}^1\mathsf{L}^1\mathsf{T}^{-2}. \tag{1.22}$$

The dimensional matrix is given by

$$\mathbf{M} = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & -3 & -1 & 1 \\ 0 & -1 & 0 & -1 & -2 \end{bmatrix}. \tag{1.23}$$

The matrix has rank 3 and thus there are 2-dimensional null space. A pair of null vectors can be (in columns)

$$\mathbf{A} = \begin{bmatrix} 1 & -2 \\ 1 & -2 \\ 1 & -1 \\ -1 & 0 \\ 0 & 1 \end{bmatrix} \tag{1.24}$$

which corresponds to two dimensionless parameters:

$$\Pi_1 = \frac{\rho L V}{\mu}, \quad \Pi_2 = \frac{F_D}{\rho L^2 V^2}. \tag{1.25}$$

The Buckingham Π Theorem assets that these two parameters satisfy some (unknown) functional relation. Recall that $\Pi_1 = \mathrm{Re}$ is the **Reynolds number**. To test the drag of a car, build a model car with length L' and put it in a wind tunnel.

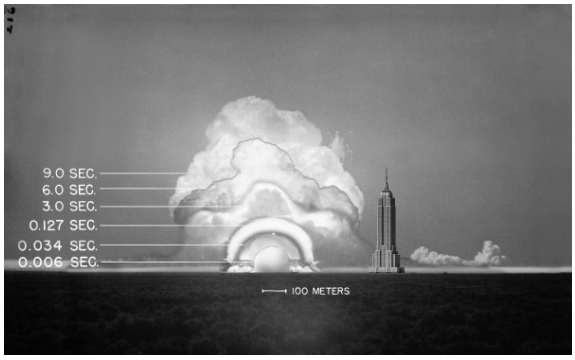

Figure 1.1 High speed photographs of the Trinity nuclear test (July 16th 1945). From the length and time scale of the footage, G. I. Taylor was able to calculate the energy yield using dimensional analysis.

Suppose the air in the wind tunnel is the same as the air on road, *i.e.* $\rho = \rho'$ and $\mu = \mu'$. Now, set the wind tunnel speed as $V' = V\frac{L}{L'}$. In that case we have $\Pi_1' = \Pi_1$. Since (Π_1, Π_2) and (Π_1', Π_2') obey the same functional relation, we must have $\Pi_2' = \Pi_2$. We can measure the drag F_D' for the model car, and find the drag of the full-scale car by $F_D = F_D' \left(\frac{V}{V'}\right)^2 \left(\frac{L}{L'}\right)^2$. ∎

∎ **Example 1.6 — Atomic bomb.** British fluid physicist G. I. Taylor was sent to the Manhattan project during World War II. Taylor was not directly involved in the development of nuclear bombs, and for security reasons worked independently from the U.S. Many data were highly classified after the war, such as the energy released in the Trinity nuclear test in 1945. However, the high speed photographs and films of the Trinity nuclear test were released to public 1947 (Figure 1.1). In 1950, Taylor published articles on estimating the energy from those magazine figures using Buckingham's Π Theorem.

The photographs show an expanding shock front of the nuclear explosion. Assume the relevant variables are

- r: the radius of the shock front. $[r] = \mathsf{L}$.
- ρ: the density of the surrounding air. $[\rho] = \mathsf{ML}^{-3}$.
- E: the energy released by the bomb. $[E] = \mathsf{ML}^2\mathsf{T}^{-2}$.
- t: the time at which the front reaches r. $[t] = \mathsf{T}$.

The dimensional matrix for (r, ρ, E, t) is given by

$$\begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & -3 & 2 & 0 \\ 0 & 0 & -2 & 1 \end{bmatrix} \tag{1.26}$$

which has a one dimensional null space spanned by

$$\begin{bmatrix} -5 \\ -1 \\ 1 \\ 2 \end{bmatrix}, \quad \Longrightarrow \quad \Pi = \frac{Et^2}{r^5\rho}. \tag{1.27}$$

Since there are no other dimensionless quantities, the functional relation is simply that

$$\Pi = \frac{Et^2}{r^5\rho} = C\,(\text{constant}) \tag{1.28}$$

where the dimensionless constant C does not depend on the energy of the bomb, the stage of the fireball, or the surrounding air. In other words, all explosions share the same universal constant C. Taylor had experimental data that indicated $C \approx 1.033$.

Using just one photograph, which shows the fireball radius $r = 80\,\text{m}$ at $t = 0.006\,\text{s}$, and the fact that $\rho = 1.1\,{}^{\text{kg}}/_{\text{m}^3}$ at that altitude, Taylor got

$$E \approx \frac{Cr^5\rho}{t^2} \approx \frac{1.033 \cdot 80^5 \cdot 1.1}{0.006^2}\text{J} \approx 10^{14}\text{J} \quad (\text{J} = {}^{\text{kg}\,\text{m}^2}/_{\text{s}^2}). \tag{1.29}$$

A kiloton of TNT is approximately 4.2×10^{12}J. Therefore,

$$E \approx 24\,\text{kilotons of TNT}. \tag{1.30}$$

Taylor averaged over more frames, and got 22 kilotons. The actual yield of the nuclear test is 20 kilotons, which is close. ∎

2. Vectors and Covectors

This chapter is a review of linear algebra with a special emphasis in distinguishing covectors from vectors. We state a few theorems from linear algebra along the exposition. Most linear algebra textbooks would state the theorems for matrices, and hence inherently basis dependent. We reprise the theorems using the language of vectors and covectors. We skip the proofs of most of these theorems. Readers can easily reconstruct the proofs by following the matrix counterpart from any matrix-based linear algebra textbook.

2.1 Dual Pairing

The *dual pairing* is a recurring operation throughout linear algebra. Linear combinations of vectors $a_1\vec{v}_1 + a_2\vec{v}_2 + a_3\vec{v}_3$ are pairing between a list of coefficients and a list of vectors. Matrix multiplications pair row vectors and column vectors:

$$\begin{bmatrix} — & \mathbf{r}_1^\intercal & — \\ — & \mathbf{r}_2^\intercal & — \end{bmatrix} \begin{bmatrix} | & | & | \\ \mathbf{c}_1 & \mathbf{c}_2 & \mathbf{c}_3 \\ | & | & | \end{bmatrix} = \begin{bmatrix} \mathbf{r}_1^\intercal\mathbf{c}_1 & \mathbf{r}_1^\intercal\mathbf{c}_2 & \mathbf{r}_1^\intercal\mathbf{c}_3 \\ \mathbf{r}_2^\intercal\mathbf{c}_1 & \mathbf{r}_2^\intercal\mathbf{c}_2 & \mathbf{r}_2^\intercal\mathbf{c}_3 \end{bmatrix}. \tag{2.1}$$

In a linear system of equations, for example

$$\begin{bmatrix} 1\,\mathrm{kg} & 1\,\mathrm{kg} \\ 2\,\mathrm{kg} & 3\,\mathrm{kg} \end{bmatrix} \begin{bmatrix} x\,\$/\mathrm{kg} \\ y\,\$/\mathrm{kg} \end{bmatrix} = \begin{bmatrix} \$\,3 \\ \$\,8 \end{bmatrix}, \tag{2.2}$$

we see pairings between arrays quantities (rows with unit of kilograms) and an array of prices (x and y) that produce scalars in costs (right-hand side). From these examples, we observe that the pairings are between *different types* of objects. In particular the pairings are *not* inner product. It does not make sense to talk about "cosine of angles" between two vectors of different units!

To distinguish the two types of objects in the linear pairing, we call one of the object *vector* and the other one *covector*. A covector is to be paired with a vector.

Definition 2.1 A covector is a scalar linear function of vectors. The evaluation, also known as the dual pairing, of a covector α and a vector $\vec{v}$ is denoted by

$$\alpha(\vec{v}) = \langle \alpha | \vec{v} \rangle = \langle \vec{v} | \alpha \rangle. \tag{2.3}$$

Definition 2.2 Let V be a real vector space. The space of all covectors on V is called the *dual space* $V^* = \{\alpha \colon V \xrightarrow{\text{linear}} \mathbb{R}\}$.

Proposition 2.1 *For a finite dimensional space V, we have $V^{**} \cong V$. In infinite dimensional cases, we may have $V^{**} \supseteq V$.*

Throughout the note, we will postulate that $V^{**} = V$.

2.1.1 Geometric picture of vectors and covectors

The standard geometric picture of a vector $\vec{v} \in V$ is an arrow based at the origin $\vec{0} \in V$. The geometric postulates for a vector space V is that it is an affine space with a distinguished point called the origin $\vec{0}$. An affine space is a space equipped with the notion of parallelism, or a projective space with a distinguished hyperplane at infinity, or equivalently a Euclidean space without metric. The vector addition $\vec{u} + \vec{v}$ is constructed by drawing a parallelogram formed by $\vec{u}$, $\vec{v}$ and connecting the origin to its opposite vertex. The scaling of a vector $a\vec{v}$ can be constructed using similar triangles and diagonal bisection in parallelograms.

The geometric picture of a covector $\alpha \in V^*$ is a foliation of equidistant parallel hyperplanes in V, each of which represents a level set $\{\alpha = k\} \subset V$ of the scalar linear function α. The dual paring $\langle \alpha | \vec{v} \rangle$ measures the number of planes of α pierced through by the arrow $\vec{v}$.

Since we know how to geometrically construct equidistant parallel hyperplanes (again via the parallel postulate) we only need to draw two special level sets: $\{\alpha = 0\}$ and $\{\alpha = 1\}$ for illustrating α. Note that $\{\alpha = 0\} = \ker \alpha$ passes through the origin. Hence each covector is a slab with one side leaning against the origin.

The covector addition $\alpha + \beta$ is constructed by first intersecting the two slabs, which forms a parallelogram prism, followed by building a new slab whose far-side from the origin passes through the diagonal of the parallelogram that is not incident to the origin. The scaling of a covector $a\alpha$ is defined by scaling the thickness of the slab by $1/a$. In particular, the narrower the slab is, the more dense the foliation of level sets of α is, and the larger the slope the linear function α has.

2.1.2 Annihilator

It is important to note that the dual pairing is different from an inner product. The dual pairing pairs a vector and a covector, whereas the inner product measures how aligned two vectors are. We do not need the inner product structure (a metric) to perform dual pairing. In fact, many inner product related notions we have seen in a first course in linear algebra are independent of metric. Here, we give an example: the notion of "orthogonal complement" of a subspace will be metric-independent if we treat it as a subspace in the dual space.

Definition 2.3 Let $U \subset V$ be a linear subspace of the vector space V. The **annihilator** $U^\circ \subset V^*$ of U is a linear subspace in the dual space defined by

$$U^\circ = \{\alpha \in V^* \,|\, \langle \alpha | \vec{u} \rangle = 0 \text{ for all } \vec{u} \in U\} \tag{2.4}$$

Proposition 2.2 $U^{\circ\circ} = U$.

2.1.3 Adjoint

A linear map $A\colon U \xrightarrow{\text{linear}} V$ from a vector space U to another vector space V induces a dual (adjoint) map from the dual target space V^* back to the dual domain U^*.

Definition 2.4 The adjoint of a linear map $A\colon U \xrightarrow{\text{linear}} V$ is the linear map $A^*\colon V^* \xrightarrow{\text{linear}} U^*$ defined by

$$\langle A^*\lambda | \vec{u} \rangle = \langle \lambda | A\vec{u} \rangle \quad \text{for all } \vec{u} \in U \text{ and } \lambda \in V^*. \tag{2.5}$$

Proposition 2.3 $A^{**} = A$.

In case U, V are equipped with bases that allow a matrix representation $\mathbf{A} = (A_{ij})$ for A, the adjoint A^* is simply the transpose $\mathbf{A}^\intercal$.

Definition 2.5 The **four fundamental subspaces** of a linear map $A\colon U \xrightarrow{\text{linear}} V$ are

- kernel: $\ker(A) \coloneqq \{\vec{u} \in U \,|\, A\vec{u} = 0\} \subset U$,
- image: $\mathrm{im}(A) \coloneqq \{A\vec{u} \,|\, \vec{u} \in U\} \subset V$,
- cokernel: $\ker(A^*) \coloneqq \{\lambda \in V^* \,|\, A^*\lambda = 0\} \subset V^*$,
- coimage $\mathrm{im}(A^*) \coloneqq \{A^*\lambda \,|\, \lambda \in V^*\} \subset U^*$.

Theorem 2.4 — Theorem of four fundamental subspaces.

$$\ker(A)^\circ = \mathrm{im}(A^*), \quad \mathrm{im}(A)^\circ = \ker(A^*) \tag{2.6}$$

2.2 Endomorphisms and Bilinear Forms

In a first course in linear algebra, we have seen many square matrices. A square matrix may represent a linear transformation of vectors. It may also represent a map from vector to a covector. These two types of square matrices are vastly different.

A linear map $A\colon V \xrightarrow{\text{linear}} V$ that maps a vector space to itself is called an **endomorphism**. The identity map is an endomorphism. An eigenvalue problem $A\vec{v} = \lambda\vec{v}$ is formulated for endomorphisms A. Powers A^k, exponentials e^A and so on are defined only for endomorphisms.

A linear map $B\colon V \xrightarrow{\text{linear}} V^*$ that maps a vector space to its dual space is called a **bilinear form**. For each $\vec{v} \in V$, the resulting value $B(\vec{v}) \in V^*$ is a covector. Recall that a covector is a scalar-valued linear function that can take another vector and return a scalar. That is, a bilinear map is to be evaluated like $B(\vec{v})(\vec{w})$ bi-linearly dependent on the two input vectors $\vec{v}, \vec{w} \in V$. In other words, a bilinear map can also be viewed as $B\colon V \times V \xrightarrow{\text{bilinear}} \mathbb{R}$.

Note that the adjoint of a bilinear form $B\colon V \xrightarrow{\text{linear}} V^*$ would have the same type $B^*\colon V \xrightarrow{\text{linear}} V^*$ (using $V^{**} = V$). Only for bilinear forms we can compare B with B^* as they are objects of the same type.

A bilinear form B is said to be **symmetric** or **self-adjoint** if $B = B^*$. Equivalently, a bilinear form is symmetric if $B(\vec{u})(\vec{v}) = B(\vec{v})(\vec{u})$ for all $\vec{u}, \vec{v} \in V$. A bilinear form is said to be **skew-symmetric** if $B = -B^*$; *i.e.* $B(\vec{u})(\vec{v}) = -B(\vec{v})(\vec{u})$.

2.2.1 Quadratic form

A symmetric bilinear form is equivalent to a quadratic form. A quadratic form $q\colon V \xrightarrow{\text{quadratic}} \mathbb{R}$ is a function such that $q(a\vec{v}) = a^2 q(\vec{v})$ for all scalar $a \in \mathbb{R}$. A bilinear form $B\colon V \xrightarrow{\text{linear}} V^*$ gives rise to a quadratic form by $q(\vec{v}) \coloneqq B(\vec{v})(\vec{v})$. Conversely, a quadratic form $q\colon V \xrightarrow{\text{quadratic}} \mathbb{R}$ gives rise to a symmetric bilinear form $B\colon V \xrightarrow{\text{linear}} V^*$ using the formula for expanding a square

$$q(\vec{u} + \vec{v}) = q(\vec{u}) + q(\vec{v}) + 2B(\vec{u})(\vec{v}). \tag{2.7}$$

Therefore, a symmetric bilinear form is fundamentally the same as a quadratic form.

2.3 Vector–Covector Conversion 1: Dual Basis

Vectors and covectors are different. The former is an arrow based at the origin, and the latter is a slab bounded by a pair of parallel hyperplanes, one of which passes through the origin. Without any additional structure, there is no method converting a vector to a covector.

One scenario where we can convert vectors to covectors is when we have a full set of basis vectors. Suppose V is an n-dimensional vector space. Suppose $\vec{v}_1, \dots, \vec{v}_n$ are n linearly independent vectors. Then they form a basis. Geometrically, they span a parallelepiped. This parallelepiped has faces giving rise to n pairs of parallel hyperplanes, one of which passes through the origin. As such, there emerges n covectors. This construction is canonical and unique. The resulting covectors is called the **dual basis** for the dual space V^*.

Conversely, given n linearly independent covectors, as a basis for V^*, we can geometrically visualize them as n slabs which intersect into parallelepiped with a corner being the origin. By reading off the edge vectors fo the parallelepiped based at the origin, we get the find the dual basis for V.

The following is the algebraic definition of the above geometric construction.

Theorem 2.5 — Dual basis. Let $\vec{v}_1, \dots, \vec{v}_n \in V$ be a basis for a vector space V. Then there exists a unique set of covectors $\alpha_1, \dots, \alpha_n \in V^*$ forming a basis for V^* such that

$$\langle \alpha_i | \vec{v}_j \rangle = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases} \tag{2.8}$$

If $V = \mathbb{R}^n$ and $\vec{v}_1, \dots, \vec{v}_n$ are column vectors $\mathbf{v}_1, \dots, \mathbf{v}_n$, then the relationship

between $\mathbf{v}_1, \dots, \mathbf{v}_n$ and the dual basis as row vectors $\alpha_1, \dots, \alpha_n$ is a matrix inversion

$$\begin{bmatrix} — & \alpha_1 & — \\ & \vdots & \\ — & \alpha_n & — \end{bmatrix} = \begin{bmatrix} | & & | \\ \mathbf{v}_1 & \cdots & \mathbf{v}_n \\ | & & | \end{bmatrix}^{-1}. \tag{2.9}$$

Note the we need a full set of basis vectors to produce a full set of basis covectors. We cannot turn a single vector to a single covector without any additional structure.

2.4 Vector–Covector Conversion 2: Metric

A **metric** or an **inner product** is an additional structure for a vector space.

Definition 2.6 A **metric** is an injective symmetric bilinear form $\flat\colon V \xrightarrow{\text{linear}} V^*$ (injectivity means that $\ker(\flat) = \{0\}$ or that $\flat$ is invertible). We usually denote $\flat(\vec{u})(\vec{v}) = \langle \vec{u}, \vec{v} \rangle$ and $\flat(\vec{u}) = \vec{u}^\flat$. The inverse $\flat^{-1}$ of the bilinear form $\flat$ is denoted by $\sharp = \flat^{-1}\colon V^* \xrightarrow{\text{linear}} V$.

The symbols $\flat, \sharp$ are called **flat** and **sharp** respectively, and they are both called **musical isomorphisms**.

Recall that a symmetric bilinear form is equivalent to a quadratic form. In the case of a metric, we denote the quadratic form by $|\cdot|^2$:

$$|\vec{v}|^2 = \flat(\vec{v})(\vec{v}) = \langle \vec{v}^\flat | \vec{v} \rangle = \langle \vec{v}, \vec{v} \rangle. \tag{2.10}$$

A metric $|\cdot|^2$ on V induces a metric on V^* given by

$$|\alpha|^2_{V^*} \coloneqq \langle \alpha | \alpha^\sharp \rangle. \tag{2.11}$$

Geometrically, we can visualize a metric structure $|\cdot|^2$ on V by the "unit sphere"

$$Q \coloneqq \left\{ \vec{x} \in V \,\middle|\, |\vec{x}|^2 = 1 \right\} \subset V, \tag{2.12}$$

which is a non-degenerate quadratic surface in the affine space centered at the origin. In addition to the parallelism structure in the vector space, we have a "compass" allowing us to define distances and angles. The relationship between the arrow $\vec{v} \in V$ and the slab $\vec{v}^\flat \in V^*$ is that the tip $\vec{v}$ of the arrow and the $\{\vec{v}^\flat = 1\}$ hyperplane of the slab are in **polarity** with respect to the quadratic surface Q.

Definition 2.7 In projective geometry, the **polar hyperplane** h to a point x with respect to a quadratic hypersurface Q is the hyperplane that passes through all the points on Q at which the tangent plane to Q contains x. If x is inside Q so that there is no tangent plane of Q containing x, then h is defined by the collection of points (outside of Q) whose polar hyperplane contains x. The point x is called the **pole** of the plane h with respect to Q if h is the **polar** of x.

If $\alpha = \vec{v}^\flat$, then the parallel hyperplanes of α are orthogonal to $\vec{v}$ with respect to the metric $\flat$.

Musical isomorphisms generally give different vector–covector conversions from the dual basis construction. In fact:

Theorem 2.6 Let $\vec{v}_1, \dots, \vec{v}_n \in V$ be a basis for a vector space V and let $\alpha_1, \dots, \alpha_n$ be its dual basis. Let $\flat$ be a metric on V. Then $\alpha_i = \vec{v}_i^\flat$ if and only if $\vec{v}_1, \dots, \vec{v}_n$ is an **orthonormal basis**.

Definition 2.8 A metric is called **positive definite** if $|\vec{v}|^2 > 0$ for all $\vec{v} \neq 0$.

For most discussion we do not need to assume that a metric is positive definite.

2.5 Miscellaneous Theorems in Linear Algebra

The typical exposition of the following theorems in linear algebra are stated in terms of matrices under an orthonormal basis, assuming a metric defined for the vector space. In those matrix representation, we lose the distinction between endomorphism and bilinear forms, and between metric-dependent and metric-independent phenomena. Here, we use a more type-sensitive exposition for these theorems.

2.5.1 Characteristic polynomial

Theorem 2.7 — Cayley–Hamilton Theorem. Let $A \in \mathrm{End}(V)$ be an endomorphism on an n-dimensional vector space V. Then there exists a unique monic polynomial (polynomial with unit highest degree coefficient) $P(t) = t^n + c_1 t^{n-1} + \cdots + c_{n-1} t + c_n$ such that $P(A) = 0$. P is called the **characteristic polynomial**.

Definition 2.9 The **trace** of an endomorphism A is defined by $\mathrm{tr}(A) \coloneqq -c_1$ from the coefficient of the characteristic polynomial. The **determinant** of A is defined as $\det(A) \coloneqq (-1)^n c_n$.

Theorem 2.8 The roots of the characteristic polynomial of an endomorphism A are the eigenvalues of A.

Theorem 2.9 Let $V_1, V_2, \dots, V_m$ be vector spaces, and let $A_1 \colon V_1 \xrightarrow{\text{linear}} V_2$, $V_2 \colon V_2 \xrightarrow{\text{linear}} V_3$, …, $V_m \colon V_m \xrightarrow{\text{linear}} V_1$ be a cyclic sequence of linear maps. In particular, for each $k = 1, \dots, m$, the composition $L_k \coloneqq A_{k-1} \circ A_{k-2} \circ \cdots \circ A_{k+1} A_k$ is an endomorphism on V_k. Then the characteristic polynomial of L_k is the same for all k. That is, the characteristic polynomial, as well as the derived notions such as trace, determinant, and eigenvalues, is a property of the entire cyclic loop independent of where the map starts.

Corollary 2.10 For linear maps A, B, C, we have $\mathrm{tr}(ABC) = \mathrm{tr}(BCA) = \mathrm{tr}(CAB)$.

2.5.2 Eigenvalue problems for symmetric matrices

Eigenvalue problems for symmetric matrices make no sense. Eigenvalue problems are posed for endomorphisms and symmetry are adjective only for bilinear forms. So, what is the deal with the famous theorem that "if a matrix is symmetric, then it has

real eigenvalues and the eigenvectors are orthogonal." The answer is that there is a hidden metric that is often omitted. Here is the restored version of the statement.

Let B, C be two bilinear forms of type $V \xrightarrow{\text{linear}} V^*$. Then it is sensible to pose the eigenvalue problem as

$$B\vec{v} = \lambda C\vec{v}. \tag{2.13}$$

Theorem 2.11 If $\flat$ is a positive definite metric on V, and $B\colon V \xrightarrow{\text{linear}} V^*$ is a symmetric bilinear form. Then there exists an orthonormal basis $\vec{v}_1, \ldots, \vec{v}_n$ with respect to $\flat$ and real numbers $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ such that

$$B\vec{v}_i = \lambda_i \flat \vec{v}_i. \tag{2.14}$$

$\vec{v}_i$ and $\vec{v}_j$ are perpendicular under both $\flat$ and B: $\flat(\vec{v}_i)(\vec{v}_j) = B(\vec{v}_i)(\vec{v}_j) = 0$ for $i \neq j$.

Theorem 2.12 The result of Theorem 2.11 holds for the more general **normal bilinear forms**. A bilinear form $B\colon V \xrightarrow{\text{linear}} V^*$ is called normal if

$$B^* \circ \sharp \circ B = B \circ \sharp \circ B^*. \tag{2.15}$$

2.5.3 Rotation matrices

A matrix $\mathbf{R}$ is called a special orthogonal matrix or a rotation matrix if $\mathbf{R}^\intercal\mathbf{R} = \mathbf{I}$. Rotation is supposed to be an endomorphism, as it maps a vector to a rotated vector in the same space. But then the adjoint $\mathbf{R}^\intercal$ would be an endomorphism on the dual space V^*. It does not make sense to compose $\mathbf{R}^\intercal$ with $\mathbf{R}$. What is missing here is some metric dependency.

Definition 2.10 Let V be a vector space with metric $\flat\colon V \xrightarrow{\text{linear}} V^*$. An endomorphism $R \in \mathrm{End}(V)$ is said to be **orthogonal** or **unitary** if the composition of the following cyclic sequence of linear maps is the identity endomorphism

$$\begin{array}{ccc} V & \xrightarrow{R} & V \\ {\scriptstyle\sharp}\uparrow & & \downarrow{\scriptstyle\flat} \\ V^* & \xleftarrow[R^*]{} & V^* \end{array} \tag{2.16}$$

If $\det(R) = 1$, then R is called a special unitary transform. Another way to describe unitarity is that the pullback metric $R^* \circ \flat \circ R$ by R is the same as the metric $\flat = R^* \circ \flat \circ R$. In other words, it is a **linear isometry**.

Definition 2.11 In general, for two spaces U, V both equipped with metric $\flat_U, \flat_V$, a linear map $A\colon U \xrightarrow{\text{linear}} V$ is an **isometry** if

$$A^* \circ \flat_V \circ A = \flat_U. \tag{2.17}$$

Equivalently,

$$\langle A\vec{x}, A\vec{y}\rangle_V = \langle \vec{x}, \vec{y}\rangle_U \quad \text{for all } \vec{x}, \vec{y} \in U. \tag{2.18}$$

2.5.4 Diagonalization

$\mathbb{R}^n = \mathbb{R} \oplus \mathbb{R} \oplus \cdots \oplus \mathbb{R}$ is a special vector space called Cartesian space. Each element is an n-tuple of numbers. A map $\mathbf{D}\colon \mathbb{R}^n \to \mathbb{R}^n$ is called a **diagonal matrix** if the map is just a channel-wise scaling

$$\mathbf{D}\colon (r_1, \ldots, r_n) \mapsto (\lambda_1 r_1, \ldots, \lambda_n r_n). \tag{2.19}$$

A basis $\vec{v}_1, \ldots, \vec{v}_n$ for a vector space V induces a map $X\colon \mathbb{R}^n \xrightarrow{\text{linear}} V$ by $X((r_1, \ldots, r_n)) \coloneqq \sum_{i=1}^n r_i \vec{v}_i$. Note that the inverse map $X^{-1}\colon V \xrightarrow{\text{linear}} \mathbb{R}^n$ is given by $X^{-1}(\vec{u}) = (\langle \alpha_1 | \vec{u}\rangle, \ldots \langle \alpha_n | \vec{u}\rangle)$ where $\alpha_1, \ldots, \alpha_n \in V^*$ is the dual basis.

A diagonalization of an endomorphism $A \in \mathrm{End}(V)$ is to express the endomorphism by

$$A = X \circ \mathbf{D} \circ X^{-1}, \tag{2.20}$$

which is achieved by solving the eigenvalue problem.

Definition 2.12 The trivial Cartesian metric on $\mathbb{R}^n$ and $\mathbb{R}^{n*} = \mathbb{R}^n$ is the identity map $\mathbf{I}$.

If V is equipped with a metric, then a basis $X\colon \mathbb{R}^n \xrightarrow{\text{linear}} V$ is orthonormal if and only if $X\colon (\mathbb{R}^n, \mathbf{I}) \to (V, \flat)$ is an isometry.

Theorem 2.13 If $X\colon (\mathbb{R}^n, \mathbf{I}) \xrightarrow{\text{isometry}} (V, \flat)$ is an orthonormal basis, then

$$X^{-1} = X^* \circ \flat \tag{2.21}$$

2.5.5 Singular value decomposition

Let U, V be two spaces equipped with metric $\flat_U, \flat_V$. The **singular value decomposition (SVD)** of a linear map $A\colon U \to V$ is to find an orthonormal basis $E_U\colon \mathbb{R}^m \xrightarrow{\text{linear}} U$ and an orthonormal basis $E_V\colon \mathbb{R}^n \xrightarrow{\text{linear}} V$ such that

$$A = E_V \circ \begin{bmatrix} \sigma_1 & & \\ & \sigma_2 & \\ & & \ddots \end{bmatrix} \circ \underbrace{E_U^{-1}}_{E_U^* \circ \flat_U} \tag{2.22}$$

where $\sigma_1 \geq \sigma_2 \geq \cdots$ and the diagonal matrix may be truncated to a rectangular matrix if $m \neq n$.

2.5.6 Polar decomposition

Let U, V be two spaces equipped with metric $\flat_U, \flat_V$. The **polar decomposition** of a linear map $A\colon U \to V$ is

$$A = RY, \quad \text{where } R\colon U \xrightarrow{\text{isometry}} V \text{ and } Y \in \mathrm{End}(U) \tag{2.23}$$

such that $\flat_U \circ Y$ is a symmetric positive definite bilinear form.

2.6 Vector Fields and Covector Fields

A typical picture of a vector field is many arrows scattered over a domain, where each of the arrows represents a direction and a magnitude of some flow at the specific location where the arrow is based. Note that the arrows composing the vector field are based at different based points, which is in contrast to vectors in a vector space which are all based at the same origin. Therefore, to describe vector fields sensibly, we must consider a collection of vector spaces V_p indexed by the base point p in the domain. A vector field $\vec{v}$ is an assignment of a vector $\vec{v}_p \in V_p$ for every point p in the domain.

With the above picture in mind, let us introduce the standard mathematical notations for vector fields.

Let $M \subset \mathbb{R}^n$ be a region in an n-dimensional space representing a domain. For each $p \in M$, define a vector space T_pM given by a copy of $\mathbb{R}^n$. We call T_pM the **tangent space** to the domain M at p. The elements of T_pM are called **tangent vectors** at p.

The elements of T_pM are associated with *velocities* at which a particle at p can travel. For each smooth parameterized curve $\gamma\colon (-\epsilon, \epsilon) \to M$ that passes through p at $t = 0$, that is $\gamma(0) = p$, the instantaneous velocity $\frac{d\gamma}{dt}$ at $t = 0$ is an element of T_pM; that is, $\dot{\gamma}(0) \in T_pM$. Conversely, for each tangent vector $\vec{v}_p \in T_pM$, there exists (non-uniquely) a parameterized curve $\gamma\colon (-\epsilon, \epsilon) \to M$ such that $\gamma(0) = p$ and $\dot{\gamma}(0) = \vec{v}_p$.

The collection $TM = \bigsqcup_{p \in M} T_pM$ of all tangent vectors at various base points p is called the **tangent bundle** over M. There is a projection operator $\pi\colon TM \to M$, $\pi\colon \vec{v}_p \mapsto p$. That is, π takes in a tangent vector $\vec{v}_p \in TM$ and returns the information $p \in M$ of which base point $\vec{v}_p$ has. The tangent space T_pM at p is the same as the preimage $\pi^{-1}(\{p\})$ of the point p via the projection. The preimage $\pi^{-1}(\{p\})$ of a point via the projection of a bundle is often called a **fiber** of the bundle.

A vector field is an assignment of an element in each fiber of the tangent bundle. Such an object is called a **section** of the bundle.

Definition 2.13 A **section** $\vec{v}$ of a bundle $\pi\colon TM \to M$ is a map $\vec{v}\colon M \to TM$ such that $\pi \circ \vec{v} = \mathrm{id}_M$. That is, $\pi(\vec{v}(p)) = p$ for all $p \in M$.

The space of all sections of the tangent bundle (*i.e.* the space of all vector fields defined over M) is denoted by $\Gamma(TM)$. So, a vector field would be instanced as $\vec{v} \in \Gamma(TM)$.

A covector field $\alpha \in \Gamma(T^*M)$ is a section of the **cotangent bundle** T^*M over M. The cotangent bundle is defined such that each of its fiber T_p^*M is the dual space of the tangent space T_pM at the same base point.

2.7 Differential of a Function

One of the most important applications of distinguishing vectors and covectors is to understand the meaning of taking derivatives of a function.

Let $M \subset \mathbb{R}^n$ be a region in an n-dimensional space representing a domain, on which we will consider generic scalar-valued non-linear functions $g\colon M \to \mathbb{R}$.

Definition 2.14 The differential of a function $g\colon M \to \mathbb{R}$ at a point $p \in M$ in a domain M is a covector $dg|_p$ based at p. This covector takes in a vector $\vec{v}_p$ based at p (representing a small displacement about p) and returns the rate of change $\langle dg_p|\vec{v}_p\rangle$ of g in the direction of $\vec{v}_p$. Concretely, for each $\vec{v}_p \in T_pM$ consider a parameterized curve $\gamma\colon (-\epsilon,\epsilon) \to M$ with $\gamma(0) = p$ and $\dot{\gamma}(0) = \vec{v}_p$, and define

$$\left\langle dg|_p \middle| \vec{v}_p \right\rangle \coloneqq \left.\frac{d}{dt}\right|_{t=0} g(\gamma(t)). \tag{2.24}$$

The differential dg of a scalar function $g\colon M \to \mathbb{R}$ is a covector field $dg \in \Gamma(T^*M)$ describing the "slope" of g at every point by the best fitting linear function on every tangent space.

2.7.1 Conversion 1: to partial derivatives

A **coordinate system** on M is a set of scalar functions $x_1\colon M \to \mathbb{R}, \ldots, x_n\colon M \to \mathbb{R}$ such that the covectors $dx_1, \ldots, dx_n$ form a basis for every cotangent space T_p^*M.

Let $\vec{e}_1, \ldots, \vec{e}_n \in \Gamma(TM)$ be the dual basis of $dx_1, \ldots, dx_n$.

Definition 2.15 The partial derivative of g along the k-th direction in a coordinate system $(x_1, \ldots, x_n)$ is defined by

$$\frac{\partial g}{\partial x_k} \coloneqq \langle dg|\vec{e}_k\rangle. \tag{2.25}$$

Equivalently, $\frac{\partial g}{\partial x_1}, \ldots, \frac{\partial g}{\partial x_n}$ are the coefficients when expressing the covector dg in terms of the covector basis from the coordinate system

$$dg = \frac{\partial g}{\partial x_1} dx_1 + \cdots \frac{\partial g}{\partial x_n} dx_n. \tag{2.26}$$

R The partial derivative $\frac{\partial g}{\partial x_1}$ depends not only on the coordinate function x_1 but it depends on the entire coordinate system $x_1, \ldots, x_n$. This is because the construction relies on taking the dual basis. Intuitively, the partial derivative depends on which other variables are fixed during the variation.

2.7.2 Conversion 2: to gradient vector

The **gradient vector** of a function g is the vector whose direction is the steepest ascending direction and whose magnitude is the slope along that direction.

Definition 2.16 Let M be a domain equipped with metric $\flat$ on every tangent space. Then the **gradient** $\operatorname{grad} g$ of a function $g\colon M \to \mathbb{R}$ is defined by

$$\operatorname{grad} g \coloneqq (dg)^\sharp. \tag{2.27}$$

One may use the polarity geometric picture of the musical isomorphism about the unit sphere to see why the gradient vector points in the steepest ascending direction.

Note that the gradient of a function is independent of the coordinate. It does however depend on a choice of metric.

R In many expositions in optimization or machine learning, the gradient is defined by the array of partial derivatives

$$\operatorname{grad} g \stackrel{?}{=} \left(\frac{\partial g}{\partial x_1}, \dots, \frac{\partial g}{\partial x_n} \right). \tag{2.28}$$

This is not true unless the coordinate system is orthonormal and therefore it is misleading. Note that partial derivatives arise from the conversion of dual basis, and the gradient arises from the conversion using a metric.

2.7.3 Pushforward of vectors

The idea of the differential of a function $dg_p \colon T_pM \xrightarrow{\text{linear}} \mathbb{R}$ of a scalar function $g \colon M \to \mathbb{R}$ can be extended for general maps.

Let $\phi \colon M \to N$ be a general nonlinear map from a space M to another space N. Then the differential $d\phi$ of the map ϕ is a linear map between the corresponding tangent spaces

$$d\phi_p \colon T_pM \xrightarrow{\text{linear}} T_{\phi(p)}N. \tag{2.29}$$

Such a linear map transforms a rate of change $\vec{v}_p \in T_pM$ at the input $p \in M$ to a rate of change of $(d\phi_p)(\vec{v}) \in T_{\phi(p)}N$ at the output $\phi(p) \in N$. That is, $d\phi_p(\vec{v}_p)$ describes how sensitive the value of ϕ is depending on variation in the $\vec{v}$ direction.

Definition 2.17 Let $\phi \colon M \to N$ be a general nonlinear map. The linear map $\phi_* = d\phi \colon TM \to TN$, with $(\phi_*)_p \colon T_pM \xrightarrow{\text{linear}} T_{\phi(p)}N$, is called the **pushforward** map of ϕ. It is defined such that for every scalar function $g \colon N \to \mathbb{R}$, the composited scalar function $f \coloneqq (g \circ \phi) \colon M \to \mathbb{R}$ has the differential given by the compositions of linear maps

$$df = dg \circ \phi_*. \tag{2.30}$$

Theorem 2.14 — Jacobian matrix. Suppose we put a coordinate system $x_1, \dots, x_m \colon M$ on M and a coordinate system $y_1, \dots, y_n \colon N \to \mathbb{R}$ on N. Let $\vec{a}_1, \dots, \vec{a}_m$ be the coordinate vectors as the dual basis of the covector basis $dx_1, \dots, dx_m$. Similarly, let $\vec{e}_1, \dots, \vec{e}_n$ be the dual basis of $dy_1, \dots, dy_n$. Now, call $\phi_i \coloneqq y_i \circ \phi \colon M \to \mathbb{R}$ as the resulting coordinate value of the mapping ϕ. For each vector $\vec{v}$ on M, we can write it under the basis as $\vec{v} = \sum_{i=1}^m v_i \vec{a}_i$. The resulting vector after being applied by pushforward is $\phi_* \vec{v} = \sum_{j=1}^n w_j \vec{e}_j$. Then,

$$\begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} = \underbrace{\begin{bmatrix} \frac{\partial \phi_1}{\partial x_1} & \cdots & \frac{\partial \phi_1}{\partial x_m} \\ \vdots & & \vdots \\ \frac{\partial \phi_n}{\partial x_1} & \cdots & \frac{\partial \phi_n}{\partial x_m} \end{bmatrix}}_{\mathbf{F}} \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix} \tag{2.31}$$

where the partial derivatives are defined in the way of Section 2.7.1. The matrix

$\mathbf{F} \in \mathbb{R}^{n\times m}$ is called the **Jacobian matrix** of ϕ under the coordinates $(x_1, \dots, x_m)$ and $(y_1, \dots, y_n)$.

Theorem 2.15 The sensitivity analysis for a composition $\psi \circ \phi \colon M \to L$ of maps $\phi \colon M \to N$ and $\psi \colon N \to L$ is simply the composition of their pushforwards:

$$(\psi \circ \phi)_* = \psi_* \phi_*, \quad \text{or} \quad d(\psi \circ \phi)_p = d\psi_{\phi(p)} d\phi_p. \tag{2.32}$$

This rule is called the **chain rule**.

Evaluating the differential by composing the chain of differentials of subcomponents is the basis for the **forward mode auto-differentiation** programs.

2.7.4 Pullback of covectors

Definition 2.18 The adjoint

$$\phi^* \colon T^*_{\phi(p)} N \xrightarrow{\text{linear}} T^*_p M \tag{2.33}$$

of the pushforward $\phi_* \colon T_p M \xrightarrow{\text{linear}} T_{\phi(p)} N$ is called the **pullback** operator (for covectors) via the mapping $\phi \colon M \to N$. For each covector (field) α on N, $\phi^*\alpha$ becomes a covector (field) on M and is called the pullback covector through ϕ. By definition of adjoint,

$$\langle \phi^* \alpha | \vec{v} \rangle = \langle \alpha | \phi_* \vec{v} \rangle \quad \text{for all } \alpha \in T^*_{\phi(p)} N \text{ and } \vec{v} \in T_p M. \tag{2.34}$$

When written in coordinates, the pullback operator is the transpose of the Jacobian matrix.

The word *pullback* can also be used for composition of function. Suppose $g \colon N \to \mathbb{R}$ is a function on N and suppose $\phi \colon M \to N$. Then $f \colon M \to \mathbb{R}$ defined by $f = g \circ \phi =: \phi^* g$ is called the **pullback** of g via the mapping ϕ.

Theorem 2.16 The pullback ϕ^* for functions g and covectors dg are defined so that

$$d(\phi^* g) = \phi^*(dg) \tag{2.35}$$

Theorem 2.17 Let $\phi \colon M \to N$ and $\psi \colon N \to L$. Then

$$(\psi \circ \phi)^* = \phi^* \circ \psi^*. \tag{2.36}$$

■ **Example 2.1 — Back-propagation.** Suppose we have the sensitivity $\phi_* = d\phi$, $\psi_* = d\psi$ of maps $\phi \colon M \to N$ and $\psi \colon N \to L$. Suppose $g \colon L \to \mathbb{R}$ is some "cost function." Through ϕ and ψ we can pull the function back and formulate a cost function f on the parameter space M defined by $f = g \circ \psi \circ \phi$.

How do we compute the differential df_p at a particular set of parameters $p \in M$?

In a forward-mode differentiation method, one would first say that df_p is, under a coordinate system $(x_1, \ldots, x_m)$, given by $df = \frac{\partial f}{\partial x_1} dx_1 + \cdots + \frac{\partial f}{\partial x_m} dx_m$ where $\frac{\partial f}{\partial x_i} = df(\vec{a}_i)$ where $\vec{a}_i$ is the coordinate vector. To find each of these partial derivatives, we expand $df = (dg)(d\psi)(d\phi)$, and evaluate $df(\vec{a}_i) = (dg)(d\psi)(d\phi)(\vec{a}_i)$ for each i. These pushforward operations have to be performed for m numbers of time.

An alternative approach is the so-called **reversed-mode differentiation** (a.k.a. back-propagation)

$$df_p = (\phi_p^*)(\psi_{\phi(p)}^*)(dg_{\psi(\phi(p))}). \tag{2.37}$$

In this case, we only need to pushback one covector (since g is scalar valued), rather than pushing forward m vectors like in the forward mode differentiation. ∎

3. Calculus of Variations

Calculus of variations is differential calculus applied on general spaces such as a function space. Just like low dimension calculus, variational calculus allows us to study the response in a functional due to small changes in its input. Here, a **functional** is just another name for function with an emphasis that it may be a function over a space of functions. That is, a functional's argument can be a function. The primary application of calculus of variations is for optimization problems. The optimality condition for a minimization problem is that the loss functional has zero derivative at the optimum.

■ **Example 3.1 — The brachistochrone problem.** This problem posed by Johann Bernoulli (1696) is regarded as the first example that gives birth to the calculus of variations.

> *Given two points A and B in a vertical plane, what is the curve traced out by a point acted on only by gravity, which starts at A and reaches B in the shortest time*

Let the coordinates of the two fixed points be $A = (a, y_a)$, $B = (b, y_b)$. For the rolling particle to travel from A to B we assume $y_b > y_a$. The goal is to find a continuously differentiable curve $y = y(x), a \leq x \leq b$ subject to the constraint $y(a) = y_a, y(b) = y_b$. Among all of these tracks, we find the one that would take the least time for the particle to travel along. Using energy conservation, we can figure out the speed of the particle at any height. The gravitation potential that has been released at height y is given by $mg(y_a - y)$, where $g > 0$ is the gravity. Converting these gravitational potential to kinetic energy $\frac{mv^2}{2}$ we find that the speed $v = \sqrt{2g(y_a - y)}$. The arclength ds traveled by particle that moves by dx

in the x coordinate is given by $ds = \sqrt{dx^2 + dy^2} = \sqrt{1 + y'(x)^2}\, dx$. So, the time dt lapsed over as the particle moves by dx is given by $dt = \frac{ds}{v} = \sqrt{\frac{1+y'(x)^2}{2g(y_a - y)}}\, dx$. Hence the total time minimization problem is formulated as

$$\underset{y \in \mathcal{V}}{\text{minimize}} \underbrace{\int_a^b \sqrt{\frac{1 + y'(x)^2}{2g(y_a - y)}}\, dx}_{\mathcal{E}(y)}, \quad \mathcal{V} = \{y \colon [a, b] \to \mathbb{R} \,|\, y(a) = y_a, y(b) = y_b\} \tag{3.1}$$

Here the energy functional $\mathcal{E} \colon \mathcal{V} \to \mathbb{R}$ is a function defined over a function space $\mathcal{V}$. Its minimizer will be a function that describes the curve of least traveling time. ■

Many optimization problems such as the brachistochrone problem take the following form

$$\underset{y \in \mathcal{V}}{\text{minimize}} \underbrace{\int_a^b L(y(x), y'(x))\, dx}_{\mathcal{E}(y)}, \quad \mathcal{V} = \{y \colon [a, b] \to \mathbb{R} \,|\, y(a) = y_a, y(b) = y_b\} \tag{3.2}$$

where $L \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is some expression. For example, for (3.1) we have $L(u_1, u_2) = \sqrt{\frac{1+u_2^2}{2g(y_a - u_1)}}$.

3.1 Variations and Functional Gradients

Let $\mathcal{V}$ be a vector space, which is typically a function space. Let $\mathcal{E} \colon \mathcal{V} \to \mathbb{R}$ be a functional defined over $\mathcal{V}$. We call $\mathcal{E}$ an energy functional. The differential $d\mathcal{E}$ of $\mathcal{E}$ at any point $u \in \mathcal{V}$ is a covector $d\mathcal{E}|_u \in \mathcal{V}^*$, which is defined similar to the construction in a finite dimensional setting. Consider any vector $\mathring{u} \in \mathcal{V}$ as the direction of variation. Then

$$d\mathcal{E}_u[\![\mathring{u}]\!] = \langle d\mathcal{E}_u | \mathring{u} \rangle \coloneqq \left. \frac{d}{d\epsilon} \right|_{\epsilon=0} \mathcal{E}(u + \epsilon \mathring{u}). \tag{3.3}$$

We use the notation of $\mathring{u}$ to indicate that it is a variation in the object u. Likewise, we will often write

$$\mathring{\mathcal{E}} = \langle d\mathcal{E}_u | \mathring{u} \rangle \tag{3.4}$$

to denote the corresponding variation in $\mathcal{E}$.

■ **Example 3.2** Let $\mathcal{V} = \{u \colon [a, b] \to \mathbb{R}\}$. Define $\mathcal{E} \colon \mathcal{V} \to \mathbb{R}$ as $\mathcal{E}(u) = \int_a^b u^2(x)\, dx$. Then the linear function $d\mathcal{E}|_u \colon \mathcal{V} \xrightarrow{\text{linear}} \mathbb{R}$ is $d\mathcal{E}_u[\![\mathring{u}]\!] = \int_a^b 2u(x)\mathring{u}(x)\, dx$. ■

■ **Example 3.3** Let $\mathcal{V} = \{u \colon [a, b] \to \mathbb{R}\}$. Define $\mathcal{E} \colon \mathcal{V} \to \mathbb{R}$ as $\mathcal{E}(u) = \int_a^b u'(x)^2\, dx$. Then the linear function $d\mathcal{E}|_u \colon \mathcal{V} \xrightarrow{\text{linear}} \mathbb{R}$ is $d\mathcal{E}_u[\![\mathring{u}]\!] = \int_a^b 2u'(x)\mathring{u}'(x)\, dx$. Through integration by parts we can also write $d\mathcal{E}_u[\![\mathring{u}]\!] = 2u'(b)\mathring{u}(b) - 2u'(a)\mathring{u}(a) -$

$\int_a^b 2u''(x)\mathring{u}(x)\,dx$. The expression after the integration by parts appear more directly as an L^2 linear pairing with $\mathring{u}$ together with some finite dimensional dot product pairing at the boundary. ∎

Note that $d\mathcal{E}|_u$ is the differential of $\mathcal{E}$ at u, which is a covector. It is not to be confused with the functional gradient, which is an element in $\mathcal{V}$ (which is a function). Let $\flat_{\mathcal{V}}\colon \mathcal{V} \to \mathcal{V}^*$, or "uncurried" as $\langle\!\langle \cdot, \cdot \rangle\!\rangle_{\mathcal{V}} = \flat_{\mathcal{V}}(\cdot)(\cdot)$, be some metric (inner product structure) defined on $\mathcal{V}$. Then we have a unique vector $\operatorname{grad}\mathcal{E}|_u := d\mathcal{E}|_u^{\sharp_{\mathcal{V}}}$ defined such that

$$\langle\!\langle \operatorname{grad}\mathcal{E}|_u, \mathring{u} \rangle\!\rangle = d\mathcal{E}|_u[\![\mathring{u}]\!] \quad \forall\, \mathring{u} \in \mathcal{V}. \tag{3.5}$$

■ **Example 3.4** Let $\mathcal{V} = \{u\colon [a,b] \to \mathbb{R}\}$. Define $\mathcal{E}\colon \mathcal{V} \to \mathbb{R}$ as $\mathcal{E}(u) = \int_a^b u^2(x)\,dx$, whose variation is known as $d\mathcal{E}_u[\![\mathring{u}]\!] = \int_a^b 2u(x)\mathring{u}(x)\,dx$. If we define an inner product structure by

$$\langle\!\langle u, v \rangle\!\rangle := \int_a^b u(x)v(x)m(x)\,dx \tag{3.6}$$

using some given positive function $m\colon [a,b] \to \mathbb{R}_{>0}$. Then the gradient of $\mathcal{E}$ is given by

$$\operatorname{grad}\mathcal{E}|_u(x) = \frac{2u(x)}{m(x)}. \tag{3.7}$$

∎

The domain for $\mathcal{E}$ does not need to be a vector space $\mathcal{V}$. One typical extension is that $\mathcal{A}$ is an affine space parallel to $\mathcal{V}$. That is, $\mathcal{V}$ is the tangent space of $\mathcal{A}$ everywhere on $\mathcal{A}$. Then $d\mathcal{E}|_u$ is a covector in $\mathcal{V}^*$, and $\operatorname{grad}\mathcal{E}|_u$ is a vector in $\mathcal{V}$.

■ **Example 3.5** Let $\mathcal{A} = \{u\colon [a,b] \to \mathbb{R}, u(a) = c_a, u(b) = c_b\}$ for some constants c_a, c_b. Note that $\mathcal{A}$ is an affine space. That is, the difference between any two elements in $\mathcal{A}$ form a vector space $\mathcal{V} = \{\mathring{u}\colon [a,b] \to \mathbb{R}, \mathring{u}(a) = \mathring{u}(b) = 0\}$ which is closed under additions and scalar multiplications. Let $\mathcal{E}\colon \mathcal{A} \to \mathbb{R}$ be $\mathcal{E}(u) = \int_a^b \frac{1}{2}u'(x)^2\,dx$ for $u \in \mathcal{V}$. Let $\langle\!\langle \mathring{u}, \mathring{v} \rangle\!\rangle = \int_a^b \mathring{u}(x)\mathring{v}(x)m(x)\,dx$ for $\mathring{u}, \mathring{v} \in \mathcal{V}$. Then

$$d\mathcal{E}_u[\![\mathring{u}]\!] = \int_a^b u'(x)\mathring{u}'(x)\,dx = -\int_a^b u''(x)\mathring{u}(x)\,dx, \quad u \in \mathcal{A}, \mathring{u} \in \mathcal{V}, \tag{3.8}$$

and

$$\operatorname{grad}\mathcal{E}_u(x) = -\frac{u''(x)}{m(x)}. \tag{3.9}$$

∎

■ **Example 3.6** Let $L\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}$, $L = L(u_1, u_2)$ be some expression. Let its partial derivatives be denoted by

$$dL = \frac{\partial L}{\partial u_1}\,du_1 + \frac{\partial L}{\partial u_2}\,du_2. \tag{3.10}$$

Now, consider $\mathcal{A} = \{y\colon [a,b]\to\mathbb{R} \,|\, y(a) = y_a, y(b) = y_b\}$ for some given y_a, y_b. Note that $\mathcal{A}$ is an affine space with tangent vector space $\mathcal{V} = \{\mathring{y}\colon [a,b]\to\mathbb{R} \,|\, \mathring{y}(a) = \mathring{y}(b) = 0\}$. Let $\mathcal{E}\colon \mathcal{A}\to\mathbb{R}$,

$$\mathcal{E}(y) \coloneqq \int_a^b L\left(y(x), y'(x)\right)\,dx. \tag{3.11}$$

Then its variation is

$$d\mathcal{E}|_y[\![\mathring{y}]\!] = \int_a^b \left(\frac{\partial L}{\partial u_1}(y(x), y'(x))\mathring{y}(x) + \frac{\partial L}{\partial u_2}(y(x), y'(x))\mathring{y}'(x)\right)\,dx \tag{3.12}$$

$$= \int_a^b \left(\frac{\partial L}{\partial u_1}(y(x), y'(x)) - \frac{d}{dx}\left(\frac{\partial L}{\partial u_2}(y(x), y'(x))\right)\right)\mathring{y}(x)\,dx. \tag{3.13}$$

A common shorthand notation is

$$d\mathcal{E}|_y[\![\mathring{y}]\!] = \int_a^b \left(\frac{\partial L}{\partial y} - \frac{d}{dx}\frac{\partial L}{\partial y'}\right)\mathring{y}\,dx. \tag{3.14}$$

■

In a more general case, the domain of $\mathcal{E}$ is an infinite dimensional manifold $\mathcal{M}$. Its variation is a covector field $d\mathcal{E}\in\Gamma(T^*\mathcal{M})$ and its gradient with respect some inner product structure is a vector field $\operatorname{grad}\mathcal{E}\in\Gamma(T\mathcal{M})$.

3.2 Euler–Lagrange Equation

Let $\mathcal{E}$ be a smooth function defined on a domain $\mathcal{M}$. Suppose $y_0\in\mathcal{M}$ be a minimizer to the optimization problem

$$\underset{y\in\mathcal{M}}{\text{minimize}}\,\mathcal{E}(y). \tag{3.15}$$

Then a necessary condition for y_0 is that any variation of the energy at y_0 must vanish

$$d\mathcal{E}|_{y_0}[\![\mathring{y}]\!] = 0, \quad \text{for all } \mathring{y}\in T_{y_0}\mathcal{M}, \tag{3.16}$$

or simply

$$d\mathcal{E}|_{y_0} = 0\in T^*_{y_0}\mathcal{M}. \tag{3.17}$$

Using any inner product, this vanishing variation condition becomes

$$\operatorname{grad}\mathcal{E}|_{y_0} = 0. \tag{3.18}$$

We call (3.18) the **Euler–Lagrange equation**. In most cases, solving the optimization problem (3.15) amounts to solving (3.18), which can be an algebraic equation or a differential equation.

3.3 Optimization with Equality Constraints

An optimization with equality constraints takes the following general form. Let $\mathcal{M}$ be a manifold, $\mathcal{E}\colon \mathcal{M} \to \mathbb{R}$ be the energy function, and let $\mathcal{G}\colon \mathcal{M} \to \mathcal{U}$ be a constraint function for some vector space $\mathcal{U}$. The optimization problem is stated as

$$\begin{cases} \underset{y\in\mathcal{M}}{\text{minimize}}\ \mathcal{E}(y) & \text{subject to} \\ \mathcal{G}(y) = \mathbf{0}_{\mathcal{U}}. \end{cases} \tag{3.19}$$

We assume both $\mathcal{E}$ and $\mathcal{G}$ are smooth. The necessary condition for the optimal solution y_0 is

$$d\mathcal{E}|_{y_0}[\![\mathring{y}]\!] = 0 \quad \text{for all } \mathring{y} \in (d\mathcal{G}|_{y_0})^{\perp} \tag{3.20}$$

where $(d\mathcal{G}|_{y_0})^{\perp}$ denotes the annihilator subspace

$$(d\mathcal{G}|_{y_0})^{\perp} = \left\{ \mathring{y} \,\middle|\, d\mathcal{G}|_{y_0}[\![\mathring{y}]\!] = 0 \right\}. \tag{3.21}$$

Note that the type of $d\mathcal{G}|_{y_0}$ is $d\mathcal{G}|_{y_0}\colon T_{y_0}\mathcal{M} \xrightarrow{\text{linear}} \mathcal{U}$. The condition implies that there exists a **Lagrange multiplier** $\lambda_0 \in \mathcal{U}^*$ such that $\mathcal{G}|_{y_0} = \mathbf{0}$ and

$$d\mathcal{E}|_{y_0} + \left\langle \lambda_0 \middle| d\mathcal{G}|_{y_0} \right\rangle_{\mathcal{U}^*\times\mathcal{U}} = 0 \in T_{y_0}\mathcal{M}. \tag{3.22}$$

■ **Example 3.7** Eigenvalue problems. ■

3.4 Karush–Kuhn–Tucker Conditions

Let us look at optimization problems with both equality constraints and inequality constraints. The energy function is still given as a function $\mathcal{E}\colon \mathcal{M} \to \mathbb{R}$ on a general manifold $\mathcal{M}$. The equality constraints are given through a function $\mathcal{G}\colon \mathcal{M} \to \mathcal{U}$ to a vector space $\mathcal{U}$. The inequality constraints are defined by a (proper) convex cone $\mathcal{C} \subset \mathcal{V}$ in a vector space $\mathcal{V}$ and a function $\mathcal{H}\colon \mathcal{M} \to \mathcal{V}$. The optimization problem takes the following form:

$$\begin{cases} \underset{y\in\mathcal{M}}{\text{minimize}}\ \mathcal{E}(y) & \text{subject to} \\ \mathcal{G}(y) = \mathbf{0}_{\mathcal{U}} \\ \mathcal{H}(y) \in \mathcal{C}. \end{cases} \tag{3.23}$$

The necessary stationary conditions for an optimal solution $y_0 \in \mathcal{M}$ are the following **Karush–Kuhn–Tucker (KKT) conditions**. There exist Lagrange multiplies $\lambda_0 \in \mathcal{U}^*$ and

$$\mu_0 \in \mathcal{C}^{\circ} \coloneqq \left\{ \mu \in \mathcal{V}^* \,\middle|\, \langle \mu | v \rangle_{\mathcal{V}^*\times\mathcal{V}} \le 0 \text{ for all } v \in \mathcal{C} \right\} \subset \mathcal{V}^* \tag{3.24a}$$

($\mathcal{C}^{\circ}$ is called the **polar cone** of $\mathcal{C}$) such that $\mathcal{G}|_{y_0} = \mathbf{0}$, $\mathcal{H}|_{y_0} \in \mathcal{C}$,

$$d\mathcal{E}|_{y_0} + \left\langle \lambda_0 \middle| d\mathcal{G}|_{y_0} \right\rangle_{\mathcal{U}^*\times\mathcal{U}} + \left\langle \mu_0 \middle| d\mathcal{H}|_{y_0} \right\rangle_{\mathcal{V}^*\times\mathcal{V}} = 0 \in T_{y_0}\mathcal{M}, \tag{3.24b}$$

and

$$\left\langle \mu_0 \middle| \mathcal{H}(y_0) \right\rangle_{\mathcal{V}^* \times \mathcal{V}} = 0. \tag{3.24c}$$

The last condition (3.24c) is called the **complementary slackness condition**. It implies that if $\mathcal{H}(y_0)$ is the interior of the cone $\mathcal{C}$, then the Lagrange multiplier term is deactivated $\mu_0 = 0$. Only when $\mathcal{H}(y_0)$ lies on the boundary of the cone, we see an associated "normal force" μ_0 emerges to keep $\mathcal{H}(y_0)$ in the interior of the cone. Condition (3.24c) is the pullback of the corresponding complementary slackness condition about the convex cone on $\mathcal{V}$ back to $\mathcal{M}$ via the map $\mathcal{H}$. In (3.24b), the 2nd and 3rd terms can also be interpreted as the pullbacks of the covectors $\lambda_0 \in \mathcal{U}^*, \mu_0 \in \mathcal{V}^*$ via $\mathcal{G}$ and $\mathcal{H}$ respectively.

4. Exterior Calculus

4.1 Differential Forms

Definition 4.1 Let M be a manifold. A **k-form** ω is a skew symmetric k-linear form field; that is, for each $p \in M$, we have a function linear in each of its k arguments

$$\omega_p[\![\cdot, \dots, \cdot]\!] \colon \underbrace{T_pM \times \cdots \times T_pM}_{k} \xrightarrow{\text{multilinear}} \mathbb{R} \tag{4.1}$$

satisfying

$$\omega_p[\![X_1, \dots, X_i, \dots, X_j, \dots, X_k]\!] = -\omega_p[\![X_1, \dots, X_j, \dots, X_i, \dots, X_k]\!]. \tag{4.2}$$

Such a skew-symmetric k-linear form at p is collectively

$$\omega_p \in \textstyle\bigwedge^k T_p^*M. \tag{4.3}$$

The space of k-forms is denoted by $\Omega^k(M) = \Gamma(\bigwedge^k T^*M)$.

Space $\Omega^0(M)$ of 0-forms are just space of functions $\Omega^0(M) = \{f \colon M \to \mathbb{R}\}$. Space $\Omega^1(M)$ of 1-forms are the space of covector fields.

For each p in an n-dimensional manifold M, the dimension of the vector space $\bigwedge^k T_p^*M$ is $\binom{n}{k}$. Every k-form would be 0 if $k > n$.

■ **Example 4.1** In the 3D Cartesian space, each vector written in coordinates $\mathbf{a} = (a_1, a_2, a_3)^\intercal$ can be converted into a 1-form $(\mathbf{a})_{\text{1-form}}$ or a 2-form $(\mathbf{a})_{\text{2-form}}$,

whose evaluation on a vector $\mathbf{u}$ or a pair of vectors $\mathbf{u}, \mathbf{v}$ are defined by

$$(\mathbf{a})_{\text{1-form}}[\![\mathbf{u}]\!] = a_1 u_1 + a_2 u_2 + a_3 u_3 \tag{4.4}$$

$$(\mathbf{a})_{\text{2-form}}[\![\mathbf{u}, \mathbf{v}]\!] = \det \begin{bmatrix} a_1 & u_1 & v_1 \\ a_2 & u_2 & v_2 \\ a_3 & u_3 & v_3 \end{bmatrix}. \tag{4.5}$$

A scalar function a, which is a 0-form $a = (a)_{\text{0-form}}$, can be converted into a 3-form $(a)_{\text{3-form}}$

$$(a)_{\text{3-form}}[\![\mathbf{u}, \mathbf{v}, \mathbf{w}]\!] = a \det \begin{bmatrix} u_1 & v_1 & w_1 \\ u_2 & v_2 & w_2 \\ u_3 & v_3 & w_3 \end{bmatrix}. \tag{4.6}$$

■

4.1.1 Integration and Pullback

Differential forms are to-be-integrated along a k-dimensional surface. Suppose $S\colon \mathbb{D}^k \to M$ is a parametric k-dimensional surface defined over a k-dimensional parameter space $\mathbb{D}^k$ with values in M. Recall that $dS = S_*$ is its pushforward map. Let $\theta_1, \dots, \theta_k$, $\theta_i\colon \mathbb{D}^k \to \mathbb{R}$, be an arbitrary coordinate system. That is, we have $d\theta_1, \dots, d\theta_k$ as basis covector field, whose dual basis $\vec{e}_1, \dots, \vec{e}_k \in \Gamma(T\mathbb{D}^k)$ is a coordinate vector field.

Definition 4.2 Let $\omega \in \Omega^k(M)$ be a k-form on M. The integral $\int_S \omega$ of ω over a surface S described above is defined by

$$\int_S \omega \coloneqq \left(\iint \cdots \int\right)_{\mathbb{D}^k} \omega_{S(\theta_1,\dots,\theta_k)}[\![S_*\vec{e}_1, \dots, S_*\vec{e}_k]\!]\, d\theta_1 \cdots d\theta_k. \tag{4.7}$$

This definition is independent of reparametrization of S.

■ **Example 4.2** In the 3D Cartesian space, for a scalar function a and a vector field $\mathbf{a}$ we have

- $\int_p a = a(p)$.
- $\int_C (\mathbf{a})_{\text{1-form}} = \int_C \mathbf{a} \cdot d\mathbf{l}$.
- $\int_S (\mathbf{a})_{\text{2-form}} = \iint_S \mathbf{a} \cdot \mathbf{n} dA$.
- $\int_V (a)_{\text{3-form}} = \iiint_V a\, dV$.

■

Definition 4.3 For each smooth map $\phi\colon M \to N$, we define the **pullback operator**

$$\phi^*\colon \Omega^k(N) \xrightarrow{\text{linear}} \Omega^k(M) \tag{4.8}$$

that for each $\omega \in \Omega^k(N)$,

$$(\phi^*\omega)_p[\![X_1, \dots, X_k]\!] \coloneqq \omega_{\phi(p)}[\![\phi_* X_1, \dots, \phi_* X_k]\!]. \tag{4.9}$$

Theorem 4.1

$$\int_{\phi(S)} \omega = \int_S \phi^* \omega. \tag{4.10}$$

■ **Example 4.3** Let ϕ be a smooth map between the 3D Cartesian space. Let $\mathbf{F} = \nabla\phi$, denoted as a matrix $F_{ij} = \frac{\partial \phi_i}{\partial x_j}$ called the **deformation gradient**. Let $J = \det(\mathbf{F})$. Then

- $\phi^*(a)_{0\text{-form}} = (a \circ \phi)_{0\text{-form}}$.
- $\phi^*(\mathbf{a})_{1\text{-form}} = (\mathbf{F}^\intercal \mathbf{a})_{1\text{-form}}$.
- $\phi^*(\mathbf{a})_{2\text{-form}} = (J\mathbf{F}^{-\intercal}\mathbf{a})_{2\text{-form}}$.
- $\phi^*(a)_{3\text{-form}} = (Ja)_{3\text{-form}}$.

■

4.1.2 Wedge Product and Interior Product

Definition 4.4 — Wedge product. Define a pointwise product between forms

$$\wedge \colon \Omega^k(M) \times \Omega^\ell(M) \xrightarrow{\text{bilinear}} \Omega^{k+\ell}(M), \quad \alpha_p \wedge \beta_p = (\alpha \wedge \beta)_p \tag{4.11}$$

that satisfies

- Associativity: $(\alpha \wedge \beta) \wedge \gamma = \alpha \wedge (\beta \wedge \gamma)$;
- Graded anti-commutative: $\alpha \wedge \beta = (-1)^{k\ell} \beta \wedge \alpha$ if α is a k-form and β is an ℓ-form.

Definition 4.5 — Interior product. For each vector field $X \in \Gamma(TM)$ (or a vector at a point), define a pointwise operation

$$i_X \colon \Omega^k(M) \xrightarrow{\text{linear}} \Omega^{k-1}(M), \quad i_{X_p} \omega_p = (i_X \omega)_p, \tag{4.12}$$

that satisfies

- $i_X \alpha = \langle \alpha | X \rangle$ for all 1-forms α;
- Nilpotent: $i_X i_X = 0$;
- Graded Leibniz rule: $i_X(\alpha \wedge \beta) = (i_X \alpha) \wedge \beta + (-1)^{\deg(\alpha)} \alpha \wedge (i_X \beta)$.

Theorem 4.2

$$(i_X \omega)[\![Y_1, \dots, Y_{k-1}]\!] = \omega[\![X, Y_1, \dots, Y_{k-1}]\!]. \tag{4.13}$$

Theorem 4.3 For $\alpha_1, \dots, \alpha \in \Omega^1(M)$,

$$(\alpha_1 \wedge \cdots \wedge \alpha_k)[\![X_1, \dots, X_k]\!] = \det \begin{bmatrix} \langle \alpha_1 | X_1 \rangle & \cdots & \langle \alpha_1 | X_k \rangle \\ \vdots & \ddots & \vdots \\ \langle \alpha_k | X_1 \rangle & \cdots & \langle \alpha_k | X_k \rangle \end{bmatrix}. \tag{4.14}$$

■ **Example 4.4** In the 3D Cartesian space $\mathbb{R}^3$ with covector basis $dx, dy, dz \in \Omega^1(\mathbb{R}^3)$, the conversion from Cartesian vectors/scalars to forms are explicitly given by

- $(a)_{0\text{-form}} = a.$
- $(\mathbf{a})_{1\text{-form}} = a_1\, dx + a_2\, dy + a_3\, dz.$
- $(\mathbf{a})_{2\text{-form}} = a_1\, dy \wedge dz + a_2\, dz \wedge dx + a_3\, dx \wedge dy.$
- $(a)_{3\text{-form}} = a\, dx \wedge dy \wedge dz.$

■

■ **Example 4.5** In the 3D Cartesian space,

- $(a)_{0\text{-form}} \wedge (b)_{0\text{-form}} = (ab)_{0\text{-form}}.$
- $(a)_{0\text{-form}} \wedge (\mathbf{b})_{1\text{-form}} = (a\mathbf{b})_{1\text{-form}}.$
- $(a)_{0\text{-form}} \wedge (\mathbf{b})_{2\text{-form}} = (a\mathbf{b})_{2\text{-form}}.$
- $(a)_{0\text{-form}} \wedge (b)_{3\text{-form}} = (ab)_{3\text{-form}}.$
- $(\mathbf{a})_{1\text{-form}} \wedge (\mathbf{b})_{1\text{-form}} = (\mathbf{a} \times \mathbf{b})_{2\text{-form}}.$
- $(\mathbf{a})_{1\text{-form}} \wedge (\mathbf{b})_{2\text{-form}} = (\mathbf{a} \cdot \mathbf{b})_{3\text{-form}}.$

■

■ **Example 4.6** In the 3D Cartesian space,

- $i_{\mathbf{v}}(\mathbf{a})_{1\text{-form}} = (\mathbf{a} \cdot \mathbf{v})_{0\text{-form}}.$
- $i_{\mathbf{v}}(\mathbf{a})_{2\text{-form}} = (\mathbf{a} \times \mathbf{v})_{1\text{-form}}.$
- $i_{\mathbf{v}}(a)_{3\text{-form}} = (a\mathbf{v})_{2\text{-form}}.$

■

Theorem 4.4

$$\phi^*(\alpha \wedge \beta) = (\phi^*\alpha) \wedge (\phi^*\beta). \tag{4.15}$$

Theorem 4.5

$$\phi^*(i_{\phi_*(X)}\omega) = i_X(\phi^*\omega). \tag{4.16}$$

Suppose $S\colon \mathbb{D}^k \to M$ is a $(k-1)$-dimensional surface. Let $X \in \Gamma(TM)$ be a vector field. Consider extruding the $(k-1)$-dimensional surface S into a k-dimensional surface $\mathrm{ext}^\epsilon_X S$ along X. Concretely, it is the solution to the initial value problem of flowing along X:

$$(\mathrm{ext}^\epsilon_X S)\colon [0,\epsilon] \times \mathbb{D}^k \to M, \tag{4.17}$$

$$\frac{\partial(\mathrm{ext}^\epsilon_X S)}{\partial t}(t, \theta_1, \ldots, \theta_{k-1}) = X|_{(\mathrm{ext}^\epsilon_X S)(t,\theta_1,\ldots,\theta_{k-1})} \tag{4.18}$$

$$(\mathrm{ext}^\epsilon_X S)(0, \theta_1, \ldots, \theta_{k-1}) = S(\theta_1, \ldots, \theta_{k-1}). \tag{4.19}$$

Theorem 4.6

$$\left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \int_{\mathrm{ext}_X^\epsilon S} \omega = \int_S i_X \omega. \tag{4.20}$$

4.1.3 Exterior Derivative

Definition 4.6 — Exterior derivative. Define $d\colon \Omega^k(M) \xrightarrow{\text{linear}} \Omega^{k+1}$ (not pointwise) satisfying

- df is the differential of f for any 0-form f;
- Nilpotent: $d \circ d = 0$;
- Graded Leibniz rule: $d(\alpha \wedge \beta) = (d\alpha) \wedge \beta + (-1)^{\deg(\alpha)} \alpha \wedge (d\beta)$.

■ **Example 4.7** In the 3D Cartesian space,

- $d(f)_{\text{0-form}} = (\nabla f)_{\text{1-form}}$.
- $d(\mathbf{v})_{\text{1-form}} = (\nabla \times \mathbf{v})_{\text{2-form}}$.
- $d(\mathbf{v})_{\text{2-form}} = (\nabla \cdot \mathbf{v})_{\text{3-form}}$.

■

Theorem 4.7

$$\phi^*(d\alpha) = d(\phi^*\alpha). \tag{4.21}$$

Theorem 4.8 — Stokes Theorem.

$$\int_S d\alpha = \int_{\partial S} \alpha. \tag{4.22}$$

4.2 Lie Derivative

For each vector field $X \in \Gamma(TM)$ on M, we may consider the flow map $\phi_t\colon M \to M$, $t \in (-\epsilon, \epsilon)$, generated by X; that is, it is the solution to $\frac{\partial \phi}{\partial t} = X \circ \phi$ and $\phi_0 = \mathrm{id}_M$.

Definition 4.7 — Lie derivative for differential forms. Let $X \in \Gamma(TM)$ be a vector field on M, and let $\phi_t\colon M \to M$ be the flow map generated by X. Define the **Lie derivative** $\mathscr{L}_X\colon \Omega^k(M) \xrightarrow{\text{linear}} \Omega^k(M)$ by

$$\mathscr{L}_X \alpha \coloneqq \left.\frac{\partial}{\partial t}\right|_{t=0} (\phi_t^* \alpha). \tag{4.23}$$

Definition 4.8 — Lie derivative for vector fields. Let $X \in \Gamma(TM)$ be a vector field on M, and let $\phi_t\colon M \to M$ be the flow map generated by X. Define the **Lie**

derivative $\mathscr{L}_X : \Gamma(TM) \xrightarrow{\text{linear}} \Gamma(TM)$ by

$$\mathscr{L}_X\, Y \coloneqq -\left.\frac{\partial}{\partial t}\right|_{t=0} (\phi_{t*}\, Y). \tag{4.24}$$

Another common notation for $\mathscr{L}_X\, Y$ is

$$[X, Y] \coloneqq \mathscr{L}_X\, Y \tag{4.25}$$

called the **Lie bracket**. This "commutator" notation will be justified in Theorem 4.11.

Theorem 4.9

$$\begin{aligned}\mathscr{L}_X(\alpha[\![Y_1, \ldots, Y_k]\!]) = (\mathscr{L}_X\, \alpha)[\![Y_1, \ldots, Y_k]\!] + \alpha[\![\mathscr{L}_X\, Y_1, \ldots, Y_k]\!] + \\ \cdots + \alpha[\![Y_1, \ldots, \mathscr{L}_X\, Y_k]\!]\end{aligned} \tag{4.26}$$

Definition 4.9 We say that a time-dependent k-form α_t is **Lie-advected** by X if

$$\tfrac{\partial}{\partial t}\alpha + \mathscr{L}_X\, \alpha = 0; \tag{4.27}$$

equivalently,

$$\phi_t^* \alpha_t = \alpha_0. \tag{4.28}$$

We say that a time-dependent vector field Y_t is **Lie-advected** by X if

$$\tfrac{\partial}{\partial t}\, Y + [X, Y] = 0; \tag{4.29}$$

equivalently,

$$\phi_{t*}\, Y_0 = Y_t. \tag{4.30}$$

Definition 4.10 Define the **Lie material derivative**

$$\tfrac{\partial}{\partial t} + \mathscr{L}_X, \tag{4.31}$$

which measures the rate of change of a field along the flow generated by X.

Theorem 4.10 — Cartan's formula. For $\alpha \in \Omega^k(M)$

$$\mathscr{L}_X\, \alpha = i_X\, d\alpha + d i_X \alpha. \tag{4.32}$$

Definition 4.11 — View vectors as derivations. Each vector field $X \in \Gamma(TM)$ is also viewed as an operator on functions $X : \Omega^0(M) \to \Omega^0(M)$ defined by

$$Xf \coloneqq df[\![X]\!]. \tag{4.33}$$

Theorem 4.11 For each $X, Y \in \Gamma(TM)$,

$$(\mathscr{L}_X\ Y)f = XYf - YXf. \tag{4.34}$$

In particular, if we denote **Lie bracket** between vectors

$$[\cdot,\cdot]\colon \Gamma(TM)\times\Gamma(TM)\to\Gamma(TM), \quad [X,Y] \coloneqq \mathscr{L}_X\ Y \tag{4.35}$$

then we obtain the following **Lie algebra** properties:

- $[X,Y] = -[Y,X]$;
- $[X,[Y,Z]] + [Y,[X,Z]] + [Z,[X,Y]] = 0$.

■ **Example 4.8** In the 3D Cartesian space, the Lie bracket is

- $[\mathbf{u},\mathbf{v}] = \mathbf{u}\cdot\nabla\mathbf{v} - \mathbf{v}\cdot\nabla\mathbf{u}$.

The Lie derivatives for k-forms are

- $\mathscr{L}_{\mathbf{u}}(a)_{0\text{-form}} = (\mathbf{u}\cdot\nabla a)_{0\text{-form}}$.
- $\mathscr{L}_{\mathbf{u}}(\mathbf{a})_{1\text{-form}} = (\mathbf{u}\cdot\nabla\mathbf{a} + (\nabla\mathbf{u})^\intercal\mathbf{a})_{1\text{-form}}$.
- $\mathscr{L}_{\mathbf{u}}(\mathbf{a})_{2\text{-form}} = (\mathbf{u}\cdot\nabla\mathbf{a} - \mathbf{a}\cdot\nabla\mathbf{u} + (\nabla\cdot\mathbf{u})\mathbf{a})_{2\text{-form}}$.
- $\mathscr{L}_{\mathbf{u}}(a)_{3\text{-form}} = (\mathbf{u}\cdot\nabla a + (\nabla\cdot\mathbf{u})a)_{3\text{-form}}$.

The traditional **material derivative** along a vector field $\mathbf{u}$ defined in most existing textbook is given by

$$\tfrac{D}{Dt} \coloneqq \tfrac{\partial}{\partial t} + \mathbf{u}\cdot\nabla. \tag{4.36}$$

The Lie material derivative, which measures the real rate of change of various forms, deviates from the material derivative by additional terms:

- $(\frac{\partial}{\partial t} + \mathscr{L}_{\mathbf{u}})(a)_{0\text{-form}} = (\frac{D}{Dt}a)_{0\text{-form}}$.
- $(\frac{\partial}{\partial t} + \mathscr{L}_{\mathbf{u}})(\mathbf{a})_{1\text{-form}} = (\frac{D}{Dt}\mathbf{a} + (\nabla\mathbf{u})^\intercal\mathbf{a})_{1\text{-form}}$.
- $(\frac{\partial}{\partial t} + \mathscr{L}_{\mathbf{u}})(\mathbf{a})_{2\text{-form}} = (\frac{D}{Dt}\mathbf{a} - \mathbf{a}\cdot\nabla\mathbf{u} + (\nabla\cdot\mathbf{u})\mathbf{a})_{2\text{-form}}$.
- $(\frac{\partial}{\partial t} + \mathscr{L}_{\mathbf{u}})(a)_{3\text{-form}} = (\frac{D}{Dt}a + (\nabla\cdot\mathbf{u})a)_{3\text{-form}}$.

■

4.3 Pairings, Inner Products and Hodge Stars

The dual space of k-forms is isomorphic to $(n-k)$-forms through the following dual pairing

$$\begin{aligned}\Omega^k(M)\times\Omega^{n-k}(M) &\xrightarrow{\text{bilinear}} \mathbb{R}\\ (\alpha,\beta) &\mapsto \int_M \alpha\wedge\beta.\end{aligned} \tag{4.37}$$

■ **Example 4.9** In the 3D Cartesian space, $\int_M (\mathbf{a})_{1\text{-form}}\wedge(\mathbf{b})_{2\text{-form}} = \iiint_M \mathbf{a}\cdot\mathbf{b}\,dV$.
■

Definition 4.12 Suppose $\flat$ is an inner product structure on T_pM for each $p \in M$. This inner product structure defines a natural **volume form**

$$\mu \in \Omega^n(M), \quad \mu[\![X_1, \ldots, X_n]\!] = 1 \text{ if } X_1, \ldots, X_n \text{ are positively orthonormal.} \tag{4.38}$$

The metric also defines a natural inner product on k-forms: for $\alpha, \beta \in \Omega^k(M)$

$$\langle \alpha, \beta \rangle \coloneqq \sum_{1 \le i_1 < \cdots < i_k \le n} \alpha[\![X_{i_1}, \ldots, X_{i_k}]\!]\, \beta[\![X_{i_1}, \ldots, X_{i_k}]\!] \tag{4.39}$$

using any orthonormal basis $X_1, \ldots, X_n$.

Definition 4.13 Suppose $\flat$ is an inner product structure on T_pM for each $p \in M$. Define the L^2 **inner product** for $\Omega^k(M)$ as

$$\langle\!\langle \alpha, \beta \rangle\!\rangle \coloneqq \int_M \langle \alpha, \beta \rangle \, \mu. \tag{4.40}$$

Define the pointwise linear operator called the **Hodge star**

$$\star_k \colon \Omega^k(M) \xrightarrow{\text{linear}} \Omega^{n-k}(M) \tag{4.41}$$

such that

$$\int_M \alpha \wedge \star\beta = \langle\!\langle \alpha, \beta \rangle\!\rangle, \quad \text{equivalently,} \quad (\alpha_p \wedge \star\beta_p) = \langle \alpha_p, \beta_p \rangle \mu_p. \tag{4.42}$$

■ **Example 4.10** In the 3D Cartesian space

- $\star(a)_{0\text{-form}} = (a)_{3\text{-form}}$.
- $\star(\mathbf{a})_{1\text{-form}} = (\mathbf{a})_{2\text{-form}}$.
- $\star(\mathbf{a})_{2\text{-form}} = (\mathbf{a})_{1\text{-form}}$.
- $\star(a)_{3\text{-form}} = (a)_{0\text{-form}}$.

■

Theorem 4.12

$$i_X \star_k \alpha = (-1)^k \star_{k+1} (X^\flat \wedge \alpha) \tag{4.43}$$

4.4 Covariant Derivatives and Vector-Valued Forms

So far, we have only been looking at scalar-valued k-forms. Another frequently occurring type of object is vector-valued k-forms. Recall a k-form is a section of the bundle $\bigwedge^k T^*M$. One may take a pointwise tensor product with the tangent bundle $(\bigwedge^k T^*M) \otimes TM$, whose sections are both a k-form and a vector field. These are vector-valued k-forms. For notation convenience we call

$$\Gamma\left((\textstyle\bigwedge^k T^*M) \otimes TM\right) = \Omega^k(M; TM). \tag{4.44}$$

A vector field $Y \in \Gamma(TM)$ can also be seen as a vector-valued 0-form $Y \in \Omega^0(M; TM)$. One may ask what is its directional derivative along a direction $X \in T_pM$ at a point $p \in M$.

The answer would be straightforward $dY[\![X]\!] = \lim_{\epsilon\to 0} \frac{1}{\epsilon}(Y(p+\epsilon X) - Y(p))$ if the underlying space M is a vector space or affine space. However, on a manifold, it is not clear how one takes the subtraction $Y(p+\epsilon X) - Y(p)$ between vectors from two different tangent spaces.

The answer would also be straightforward if X is an entire vector field and if the question is "what is the rate of change of Y along the flow generated by X?" That is the Lie derivative $\mathscr{L}_X Y = [X, Y] = \lim_{\epsilon\to 0}(-\phi_{\epsilon *} Y)$ where ϕ_ϵ is the flow map generated by X.

The natural definition of the directional derivative of a vector would depend on a metric.

Definition 4.14 Let $(M, \flat)$ be a Riemannian manifold. A differential operator that takes a vector field to a vector-valued 1-form

$$\nabla\colon \Omega^0(M; TM) \to \Omega^1(M; TM) \tag{4.45}$$

is a **Levi-Civita connection** or a **covariant derivative** if

- Compatible with differential of scalar: $\nabla(fY) = (df)\, Y + f \nabla Y$ for $f \in \Omega^0(M)$ and $Y \in \Omega^0(M; TM)$;
- Compatible with metric: $d\langle X, Y\rangle = \langle \nabla X, Y\rangle + \langle X, \nabla Y\rangle$;
- Compatible with Lie derivative (torsion free): $\nabla_X Y - \nabla_Y X = [X, Y]$.

Theorem 4.13 — Fundamental theorem of Riemannian geometry. The Levi-Civita connection exists and is unique.

So, on a manifold with metric, we can just take differential of a vector field like taking differential of a standard 0-form. The only difference in notation is that we use ∇ instead of d.

Definition 4.15 — Exterior covariant derivative. Using the covariant derivative ∇ we can apply exterior derivative to vector-coefficient k-forms (*i.e.* vector-valued k-forms). The operator

$$d^\nabla\colon \Omega^k(M; TM) \to \Omega^{k+1}(M; TM) \tag{4.46}$$

is characterized by that

- It is the same as ∇ when $k = 0$.
- Graded Leibniz rule: $d^\nabla(\alpha \wedge A) = (d\alpha) \wedge A + (-1)^{\deg(\alpha)} \alpha \wedge d^\nabla A$ for scalar-valued form α and vector-valued form A.

R Despite the similarity between d^∇ and d, there are some differences. For example,

$$d^\nabla \circ d^\nabla \tag{4.47}$$

is generally nonzero. In fact, $d^\nabla \circ d^\nabla = R\wedge$ where R is a matrix-valued 2-form called **Riemann curvature tensor**.

The following are some useful notions.

Definition 4.16 — Identity. The identity I is a vector-valued 1-form $I \in \Omega^1(M; TM)$ defined by

$$I(X) = X \quad \text{for all } X \in \Gamma(TM). \tag{4.48}$$

Theorem 4.14 $X^\flat = \langle X, I \rangle$.

Theorem 4.15 $d^\nabla I = 0$.

II Classical Mechanics

5. Newton's Laws of Motion

The standard task in physical simulations is to solve an **Initial Value Problem** (IVP) for an Ordinary Differential Equation (ODE):

$$\frac{d}{dt}\mathbf{y}(t) = \dot{\mathbf{y}}(t) = \mathbf{f}(\mathbf{y}(t)), \quad \mathbf{y}(0) = \mathbf{y}_0 \quad \text{given.} \tag{5.1}$$

Note that such an **evolution equation** can always be written such that it involves only first derivative in time. For example, an equation

$$\ddot{x}(t) + \sin(x(t)) = 0 \tag{5.2}$$

can be split into

$$\begin{bmatrix} \dot{x}(t) \\ \dot{v}(t) \end{bmatrix} = \begin{bmatrix} v(t) \\ -\sin(x(t)) \end{bmatrix}. \tag{5.3}$$

The evolution equation describes the dynamics of the variables. These evolution equations are often derived from classical physics. In this chapter, we describe the general principles for deriving the equations.

5.1 Laws of Motions

Newton's Laws of motions from *the Principia* (1687) are summarized as

- **Newton's 1st Law** or **law of inertia**. A body remains at rest or in motion at a constant speed in a straight line unless acted upon by a "force."
- **Newton's 2nd Law**. When the body is acted by a force, the rate of change of the "momentum" equals to the force.
- **Newton's 3rd Law**. Every action has an equal and opposite reaction.

These laws of motion is useful for deriving the dynamics of a simple system involving a few point masses connected by a few additional force laws (such as gravitation field or Hooke's law of spring). In that case, we define the momentum as mass times velocity, and the force laws come from some models. As the physical system becomes larger, it is no longer obvious how to define and account for all momenta and forces.

In the 18th and 19th century, there are a lot of developments in providing a more mathematically principled way of deriving the equation of motions. They take advantage of defining the **kinetic energy**, **potential energy**, **rate of dissipation** and so on. Each of these energies is a function defined over the moving state of object. The equations of motion are a balanced exchange of energies.

It turns out that Newton's 1st Law about inertia is completely characterized by the notion of kinetic energy. Kinetic energy also gives rise to the notion of momentum. Inertia, momenta and kinetic energies are inseparable from each other. Newton's 2nd and 3rd Law are derived when an additional potential energy of state is defined.

Note that the idea of energies was not introduced when Newton's three laws of motions were proposed. The derived motions from energies will be equivalent to Newton's law. The advantage of energy-based approach is that it is much more systematic to account for all physical forces in a system.

5.2 Law of Inertia

Let Q be the space of all positions. Every fixed point $q \in Q$ represents one configuration of the physical system at rest. Every tangent vector $\dot{q} \in T_q Q$ represents an instantaneous motion.

Definition 5.1 An **inertia tensor** is a Riemannian metric on Q. That is, it is an assignment of a positive definite quadratic form (*i.e.* symmetric positive definite bilinear form) $\flat_q \colon T_q Q \to T_q^* Q$ at every $q \in Q$. The **kinetic energy** of the instantaneous motion ($q \in Q, \dot{q} \in T_q Q$) is defined by

$$K_q(\dot{q}) = \frac{1}{2} \flat_q(\dot{q})(\dot{q}). \tag{5.4}$$

Definition 5.2 The **momentum** associated with $(q, \dot{q})$ is defined by

$$\flat_q \dot{q} \in T_q^* Q. \tag{5.5}$$

The inertia tensor as a metric gives rise to a notion of distances and angles on Q. The law of inertia says that a moving object should travel as straight and as constant speed as possible. Such a motion can be characterized geometrically as the shortest path using the notion of distances given by the inertia.

If one only wants to find the shortest path, one finds a path connecting two given points that minimizes the total arclength $\int_a^b \sqrt{\flat(\dot{q})(\dot{q})}\, dt$. If one wants to find the shortest path and the path travels at constant speed along the path, then one minimizes the total squared length $\int_a^b \flat(\dot{q})(\dot{q})$. The minimization of the latter will geometrically give the shortest path, and in addition to that the parameterization along the path will have constant speed.

Definition 5.3 Let q_a, q_b be two points on Q. Let

$$\mathcal{M}_{a,b,q_a,q_b} = \{q\colon [a,b] \to Q \mid q(a) = q_a, q(b) = q_b\} \tag{5.6}$$

be the space of all parametric curves connecting q_a, q_b. A **geodesic** path connecting q_a, q_b is a parametric curve $q\colon [a,b] \to Q$ that minimizes

$$\underset{q\in\mathcal{M}_{a,b,q_a,q_b}}{\text{minimize}} \int_a^b \frac{1}{2}\left\langle \flat_{q(t)}\dot{q}(t) \middle| \dot{q}(t)\right\rangle dt. \tag{5.7}$$

Any parametric curve $q\colon [0,T] \to Q$ is said to be a **geodesic** if at each time point $t \in (0,T)$ there is a neighborhood $[a,b] \ni t$ such that q is the geodesic connecting $q(a)$ and $q(b)$.

Postulate 5.1 — Law of inertia. The motion $q\colon [0,T] \to Q$ of a physical system follow a geodesic path unless acted upon by a force.

5.2.1 Examples of inertial motions

■ **Example 5.1 — Free particle system.** ■

■ **Example 5.2 — Reduced order systems.** ■

■ **Example 5.3 — Rotating rigid body.** ■

5.2.2 Geodesic equation

Let us derive the Euler–Lagrange equation for (5.7) in the most general form. Suppose there is a coordinate system on Q. In particular, Q is locally parameterized as $\mathbb{R}^m$, where each point is expressed in coordinates as $\mathbf{q} = (q^1, \dots, q^m)$. In this coordinate system, the inertia metric has a symmetric matrix representation $\flat = (g_{ij})_{ij}$

$$\mathbf{p}_i = (\flat\dot{\mathbf{q}})_i = g_{ij}\,q^j = \sum_j g_{ij}\,q^j. \tag{5.8}$$

Here, we used Einstein's summation convention (If a pair of repeated indices appear as one on the upper and one lower, then the index is summed; the summation symbol can be suppressed). In reverse, we can recover $\dot{\mathbf{q}}$ from the momentum $\mathbf{p}$ by

$$\dot{\mathbf{q}}^i = (\sharp\mathbf{p})^i = g^{ij}p_j. \tag{5.9}$$

As a common convention, the matrix (g^{ij}) with upper indices is the inverse matrix of (g_{ij}). Note that

$$g^{ij}g_{jk} = \delta^i_k. \tag{5.10}$$

The energy of the path is given by

$$\mathcal{E}(\mathbf{q}) = \int_a^b \underbrace{\frac{1}{2}g_{ij}(\mathbf{q}(t))\dot{q}^i(t)\dot{q}^j(t)}_{L(\mathbf{q}(t),\dot{\mathbf{q}}(t))}\,dt. \tag{5.11}$$

Next, we use another index notational convention that the index followed after a comma like $a_{ij,k}$ means partial derivative $a_{ij,k} = \frac{\partial a_{ij}}{\partial q^k}$. Now, the variation of $\mathcal{E}$ is

$$d\mathcal{E}[\![\mathring{\mathbf{q}}]\!] = \int_a^b \frac{1}{2}\left(g_{ij,k}\mathring{q}^k\dot{q}^i\dot{q}^j + g_{ij}\dot{\mathring{q}}^i\dot{q}^j + g_{ij}\dot{q}^i\dot{\mathring{q}}^j\right) dt \tag{5.12}$$

$$\text{(integration by parts)} \quad = \int_a^b \frac{1}{2}\Bigg(g_{ij,k}\dot{q}^i\dot{q}^j\mathring{q}^k - (\underbrace{\dot{g}_{ij}}_{g_{ij,\ell}\dot{q}^\ell}\dot{q}^j + g_{ij}\ddot{q}^j)\mathring{q}^i \tag{5.13}$$

$$- (\underbrace{\dot{g}_{ij}}_{g_{ij,\ell}\dot{q}^\ell}\dot{q}^i + g_{ij}\ddot{q}^i)\mathring{q}^j\Bigg) dt \tag{5.14}$$

$$\text{(relabel index)} = \int_a^b -\Bigg(\ddot{q}^k + \frac{1}{2}g^{k\ell}g_{i\ell,j}\dot{q}^i\dot{q}^j + \frac{1}{2}g^{k\ell}g_{j\ell,i}\dot{q}^j\dot{q}^i \tag{5.15}$$

$$- \frac{1}{2}g^{k\ell}g_{ij,\ell}\dot{q}^i\dot{q}^j\Bigg)g_{km}\mathring{q}^m\, dt \tag{5.16}$$

Define the **Christoffel symbol**

$$\Gamma^k_{ij} \coloneqq \frac{1}{2}g^{k\ell}\left(g_{\ell i,j} + g_{j\ell,i} - g_{ij,\ell}\right), \tag{5.17}$$

which simplifies the variation formula

$$d\mathcal{E}[\![\mathring{\mathbf{q}}]\!] = \int_a^b -\left(\ddot{q}^k + \Gamma^k_{ij}\dot{q}^i\dot{q}^j\right)g_{km}\mathring{q}^m\, dt. \tag{5.18}$$

Therefore, the equation of motion (the **geodesic equation**) is given by

$$\ddot{q}^k + \Gamma^k_{ij}\dot{q}^i\dot{q}^j = 0. \tag{5.19}$$

Definition 5.4 — Covariant time derivative. Let $(Q, \flat)$ be a Riemannian manifold. Let $\mathbf{q}\colon [a,b] \to Q$ be a path. The covariant derivative of a tangent vector field $\mathbf{v}\colon [a,b] \to TQ$, $\mathbf{v}(t) \in T_{\mathbf{q}(t)}Q$ be a vector field defined along the path. Then, under a coordinate system, the **covariant time derivative** of $\mathbf{v}$ is defined by

$$\left(\frac{\nabla}{dt}\mathbf{v}\right)^k \coloneqq \dot{v}^k + \Gamma^k_{ij}\dot{q}^i v^j. \tag{5.20}$$

Equivalently, in a coordinate-free style, let ∇ be the **Levi-Civita connection** defined in Definition 4.14. Then

$$\overset{\nabla}{\dot{\mathbf{v}}} = \frac{\nabla}{dt}\mathbf{v} \coloneqq \nabla_{\dot{\mathbf{q}}}\mathbf{v}. \tag{5.21}$$

Using the coordinate-free notation, the **geodesic equation** is given by

$$\overset{\nabla}{\ddot{\mathbf{q}}} = 0. \tag{5.22}$$

In fact, we can derive (5.22) purely free of coordinates. First, note the following convenient rule for mixed derivatives.

Theorem 5.1 — Mixed derivative. Suppose $q\colon (-\epsilon,\epsilon)\times[a,b]\to Q$ is a path depending on two parameters $q=q(s,t)$. Then

$$\frac{\nabla}{\partial t}\frac{\partial q}{\partial s}=\frac{\nabla}{\partial s}\frac{\partial q}{\partial t}. \tag{5.23}$$

That is,

$$\overset{\nabla}{\underset{\circ}{\dot{q}}}=\overset{\nabla}{\dot{\underset{\circ}{q}}}. \tag{5.24}$$

Proof. Under a coordinate system, the mixed derivative is given by

$$\left(\frac{\nabla}{\partial t}\frac{\partial q}{\partial s}\right)^k=\frac{\partial^2 q^k}{\partial t\partial s}+\Gamma^k_{ij}\frac{\partial q^i}{\partial t}\frac{\partial q^j}{\partial s} \tag{5.25}$$

which is symmetric in s and t (note that $\Gamma^k_{ij}=\Gamma^k_{ji}$). In the coordinate-free context, this theorem is equivalent to the torsion-free condition in the definition for the Levi-Civita connection (Definition 4.14). □

Now, let us derive the geodesic equation using coordinate-free notation. The energy (5.11) is given by

$$\mathcal{E}(q)=\int_a^b\frac{1}{2}|\dot{q}|^2\,dt \tag{5.26}$$

whose variation is

$$d\mathcal{E}[\![\mathring{\mathbf{q}}]\!]=\int_a^b\langle\overset{\nabla}{\underset{\circ}{\dot{q}}},\dot{q}\rangle\,dt\overset{(5.24)}{=}\int_a^b\langle\overset{\nabla}{\dot{\underset{\circ}{q}}},\dot{q}\rangle\,dt \tag{5.27}$$

$$=\int_a^b\left(\frac{d}{dt}\langle\mathring{q},\dot{q}\rangle-\langle\mathring{q},\overset{\nabla}{\ddot{q}}\rangle\right)dt\overset{\mathring{q}|_{a,b}=0}{=}\int_a^b-\langle\mathring{q},\overset{\nabla}{\ddot{q}}\rangle\,dt. \tag{5.28}$$

Therefore, we obtain (5.22) as the Euler–Lagrange equation.

5.2.3 Acceleration and the rate of change of momentum

The covariant time derivative $\overset{\nabla}{\ddot{q}}$ of the velocity $\dot{q}$ is called the (covariant) **acceleration**. Newton's first law Postulate 5.1 says that the acceleration is zero (5.22) for a force-free motion.

The vanishing acceleration can also be interpreted as that the rate of change of momentum is zero in the following sense.

Definition 5.5 The covariant derivative Definition 5.4 or Definition 4.14 on tangent vector fields induce a natural covariant derivative for covector fields $\nabla\colon\Gamma(TM)\times\Gamma(T^*M)\to\Gamma(T^*M)$:

$$(\nabla_X\alpha)[\![Y]\!]\coloneqq d_X(\alpha[\![Y]\!])-\alpha[\![\nabla_X Y]\!]. \tag{5.29}$$

Theorem 5.2 $\nabla_X(\flat Y) = \flat(\nabla_X Y)$.

Recall that $\dot{q}$ represents the velocity, $\flat\dot{q}$ is the momentum, and $\overset{\nabla}{\ddot{q}}$ is the covariant acceleration. By Theorem 5.2, we have

$$\flat(\overset{\nabla}{\ddot{q}}) = \frac{\nabla}{dt}(\flat\dot{q}). \tag{5.30}$$

Think of the conversion $\flat\colon \Gamma(TQ) \to \Gamma(T^*Q)$ from velocity to momentum as the mass. Then (5.30) describes that "mass times acceleration equals to the rate of change of momentum." Note that this statement is true even if one has spatially varying mass. Phenomena such as an apparent acceleration due to the loss of mass or decrease of moment of inertia is encoded in the covariant derivatives ∇.

5.3 Newton's 2nd Law

In Section 5.2, we defined the notion of inertia, velocity, momentum, acceleration, and the rate of change of momentum. We also described the governing equation for a moving position $q(t) \in Q$ in the absence of force. Here, we describe the general law of motion under a general force.

Definition 5.6 A **force model** is a base-point-preserving map

$$f\colon TQ \to T^*Q. \tag{5.31}$$

That is, for each $q \in Q$ and $\dot{q} \in T_qQ$, the value $f(q, \dot{q})$ is a covector in T_q^*Q.

Postulate 5.2 — Newton's 2nd Law. Let Q be the space of positions with inertia metric $\flat$. The motion of the object is described by a force model f over Q with the following equation of motion

$$\flat\overset{\nabla}{\ddot{q}} = f(q, \dot{q}), \quad \text{equivalently,} \quad \frac{\nabla}{dt}(\flat\dot{q}) = f(q, \dot{q}). \tag{5.32}$$

For simple physical systems, it is easy to find an intuitive force model depending on the position and velocity. In a general system, finding a force model using intuition is challenging and unreliable. Next, we introduce a mathematical way for deriving a special type of force models. The force models of this special type depend only on the position $f(q, \dot{q}) = f(q)$ and therefore f is a covector field. Furthermore, we assume f is exact; *i.e.* $f = -dU$ for some 0-form $U\colon Q \to \mathbb{R}$. We call such forces conservative forces, and they can describe all non-dissipative mechanical systems.

Definition 5.7 — Conservative force. A **conservative force** is an exact covector field $f \in \Omega^1(Q)$, $f = -dU$ for some $U \in \Omega^0(Q)$. The function U is called the **potential energy**.

5.4 Least Action Principle

The **least action principle**, **stationary action principle**, or **Hamilton's principle** states that the dynamical system of Postulate 5.2 can be derived using a variational principle. This variational principle resembles Definition 5.3 which characterizes the inertial motion.

Definition 5.8 — Lagrangian and Action. Let Q be the space of all positions with an inertial metric $\flat$. Let $U\colon Q \to \mathbb{R}$ be a potential energy. For each parametric curve $q\colon [a,b] \to Q$, we call

$$S(q) \coloneqq \int_a^b \left(\frac{1}{2} \left\langle \flat_{q(t)} \dot{q}(t) \middle| \dot{q}(t) \right\rangle - U(q(t)) \right) dt \tag{5.33}$$

the **action** of the path q. The integrand $L(q,\dot{q}) = \frac{1}{2}|\dot{q}|^2 - U(q)$ is called the **Lagrangian**.

Definition 5.9 Following Definition 5.3, let $\mathcal{M}_{a,b,q_a,q_b} = \{q\colon [a,b] \to Q \mid q(a) = q_a, q(b) = q_b\}$ be the space of all parametric curves connecting two given points $q_a, q_b \in Q$. A **least-action path** connecting q_a, q_b is a parametric curve $q\colon [a,b] \to Q$ that minimizes the action $S(q)$ defined in (5.33). Any parametric curve $q\colon [0,T] \to Q$ is said to be a **least-action path** if at each time point $t \in (0,T)$ there is a neighborhood $[a,b] \ni t$ such that q is the least-action path connecting $q(a)$ and $q(b)$.

Theorem 5.3 The equation for a least-action path is given by (5.32) where the force f is a conservative force with $f = -dU$. That is, the Euler–Lagrange equation as the equation of motion

$$\frac{\nabla}{dt} \frac{\partial L(q,\dot{q})}{\partial \dot{q}} + \frac{\partial L(q,\dot{q})}{\partial q} = 0 \tag{5.34}$$

is given by

$$\flat \overset{\nabla}{\dot{q}} = -dU. \tag{5.35}$$

5.4.1 Examples of least action principle

5.5 Hamiltonian Formulation

The Euler–Lagrange equation (5.34) is the minimizing condition for the path's action in the general form $S(q) = \int_a^b L(q(t),\dot{q}(t))\,dt$ with some Lagrangian function $L\colon TQ \to \mathbb{R}$. This equation (5.34), or in the more familiar "$F = ma$" equation (5.35), is an ODE that typically 2nd order in time. In the following, we convert the Euler–Lagrange equation into a system of equation that is 1st order in time.

First, we define **momentum** as the differential of $L\colon T_qQ \to \mathbb{R}$ for each fixed $q \in Q$:

$$p(q,v) \coloneqq \frac{\partial L(q,v)}{\partial v} \in T_q^*Q. \tag{5.36}$$

We make the following assumption that for each fixed q, the function $p(q,v)$ of v is invertible. Concretely, there exists $F_q\colon T_qQ \to T_q^*Q$ and its inverse function $G_q\colon T_q^*Q \to T_qQ$ such that

$$p(q,v) = F_q(v), \quad v(q,p) = G_q(p). \tag{5.37}$$

In our case, F_q happens to be a pure differential of a function $F_q = \frac{\partial L}{\partial v}$. This condition ensures that its inverse function G_q defined over the dual space T_q^*Q is also a pure differential of a function, which is now defined on T_q^*Q. That is,

$$G_q(p) = \frac{\partial H(q,p)}{\partial p} \quad \text{for some } H(q,\cdot)\colon T_q^*Q \to \mathbb{R}. \tag{5.38}$$

We call this function $H(q,\cdot)\colon T_q^*Q \to \mathbb{R}$ the **Legendre transform** of the function $L(q,\cdot)\colon T_qQ\colon \mathbb{R}$. An explicit construction of the Legendre transform is given by

$$H(q,p) \coloneqq \sup_{v\in T_qQ} \langle p|q\rangle - L(q,v). \tag{5.39}$$

One can check that $\frac{\partial H}{\partial p}$ and $\frac{\partial L}{\partial v}$ are indeed inverse functions of each other.

In the context of least action system, the Legendre transform H of the Lagrangian L is called the **Hamiltonian**.

■ **Example 5.4** If $L(q,v) = \frac{1}{2}\langle \flat_q v|v\rangle - U(q)$, then

$$H(q,p) = \frac{1}{2}\langle \sharp_q p|p\rangle + U(q), \tag{5.40}$$

which is also known as the total energy of the system. Here $\sharp_q = \flat_q^{-1}$. ■

Now, we rewrite the Euler–Lagrange equation $\frac{\nabla}{dt}\frac{\partial L}{\partial \dot q} = -\frac{\partial L}{\partial q}$ in terms of q,p:

$$\begin{cases} \dot q(t) = \frac{\partial H(q,p)}{\partial p} \\ \dot p(t) = -\frac{\partial H(q,p)}{\partial q}. \end{cases} \tag{5.41}$$

This is equation is called **Hamilton's equation**. Hamilton's equation can be derived by direct substitution of the equation $\frac{\nabla}{dt}\frac{\partial L}{\partial \dot q} = -\frac{\partial L}{\partial q}$. One may also derive Hamilton's equation by first writing the action in terms of q,p variables followed by taking its Euler–Lagrange equation:

$$S(q,p) = \int_a^b \Big(\langle p(t)|\dot q(t)\rangle - H(q(t),p(t))\Big)\, dt \tag{5.42}$$

5.5.1 Canonical Coordinate

In most practices, one puts a coordinate system on Q, making an identification $Q = \mathbb{R}^m$. In that case q is written as $\mathbf{q} = (q_1,\dots,q_m)$. The natural basis for T_q^*Q is the dual basis $dq_1,\dots,dq_m$ and $\mathbf{p} = \frac{\partial L}{\partial \dot{\mathbf{q}}} = (\frac{\partial L}{\partial \dot q_1},\dots,\frac{\partial L}{\partial \dot q_m})$ is the associated momentum. In this so called **canonical coordinate**, (5.41) becomes

$$\begin{cases} \dot{\mathbf{q}}(t) = \frac{\partial H(\mathbf{q},\mathbf{p})}{\partial \mathbf{p}} \\ \dot{\mathbf{p}}(t) = -\frac{\partial H(\mathbf{q},\mathbf{p})}{\partial \mathbf{q}}. \end{cases} \tag{5.43}$$

Note that under this canonical coordinate, the covariant derivative $\frac{\nabla}{dt}$ becomes trivial $\frac{d}{dt}$.

This is somewhat unusual. In a general Riemannian manifold $(Q, \flat)$ the covariant derivative ∇ does not reduce to a simple partial derivative in any coordinate system, unless the metric $\flat$ is flat (has no curvature). In contrast, when we describe the dynamical system for (q, p) in the cotangent bundle T^*Q, one can always find coordinate (canonical coordinate) so that ∇ becomes simple derivatives. One also get a hint of metric independence from (5.42), which is expressed only using dual pairing.

5.5.2 Poisson system

The position-momentum space T^*Q is called the **phase space**. An function defined over the phase space T^*Q is called an observable. For example, $H\colon T^*Q$ is an observable. The value of a coordinate component, say $q_1\colon (q_1, \dots, q_m, p_1, \dots, p_m) \mapsto q_1$, is an observable. For simplicity, let us assume Q has a coordinate $\mathbf{q}$ and we use the associated canonical coordinate $(\mathbf{q}, \mathbf{p})$ for T^*Q. Each observable is some function $f(\mathbf{q}, \mathbf{p})$.

Theorem 5.4 Following the time evolution of (5.43), the value $f(t) = f(\mathbf{q}(t), \mathbf{p}(t))$ of an observable f has the following rate of change:

$$\frac{df}{dt} = \left\langle \frac{\partial f}{\partial \mathbf{q}} \middle| \frac{\partial H}{\partial \mathbf{p}} \right\rangle - \left\langle \frac{\partial f}{\partial \mathbf{p}} \middle| \frac{\partial H}{\partial \mathbf{q}} \right\rangle. \tag{5.44}$$

This equation is called the **Poisson system**.

Definition 5.10 A standard notion is the (Lie)-Poisson bracket $\{\cdot, \cdot\}\colon C^\infty(T^*Q) \times C^\infty(T^*Q) \to C^\infty(T^*Q)$, which takes two observable and spits out another observable. Under a canonical coordinate, it is given by

$$\{f, g\} \coloneqq \left\langle \frac{\partial f}{\partial \mathbf{q}} \middle| \frac{\partial g}{\partial \mathbf{p}} \right\rangle - \left\langle \frac{\partial f}{\partial \mathbf{p}} \middle| \frac{\partial g}{\partial \mathbf{q}} \right\rangle. \tag{5.45}$$

In fact, the Poisson bracket on the vector space $C^\infty(T^*Q)$ of observables makes $C^\infty(T^*Q)$ a Lie algebra. That is, $\{\cdot, \cdot\}$ is bilinear, skew symmetric, and satisfies the Jacobi identity.[1]

The Poisson system can be written in terms of the Poisson bracket as

$$\dot{f} = \{f, H\}. \tag{5.46}$$

Corollary 5.5 Under (5.43), the Hamiltonian $H(\mathbf{q}(t), \mathbf{p}(t))$ is a constant (independent of t).

[1] The axiomatic definition of a Poisson bracket is that it is a Lie bracket on the (infinite dimensional) space $C^\infty(P)$ of observables on a phase space P, and that $\{f, \cdot\}\colon C^\infty(P) \to C^\infty(P)$ is a derivation (Definition 4.11).

5.5.3 Liouville's Theorem

Theorem 5.6 — Liouville's Theorem of Area Conservation. Let $C = C_t$ be a closed 1-dimensional loop in the phase space that moves following the time evolution of (5.43). Then the circulation

$$\oint_{C_t} \langle \mathbf{p} | d\mathbf{q} \rangle \tag{5.47}$$

is a constant of motion. Equivalently,

$$\mathscr{L}_{(\partial H/\partial \mathbf{p}, -\partial H/\partial \mathbf{q})} \langle d\mathbf{p} \wedge d\mathbf{q} \rangle = 0. \tag{5.48}$$

Let us call $\sigma = \langle d\mathbf{p} \wedge d\mathbf{q} \rangle \in \Omega^2(T^*Q)$, which is $d\langle \mathbf{p} | d\mathbf{q} \rangle$. In particular $d\sigma = 0$. This $\langle \mathbf{p} | d\mathbf{q} \rangle$ is called **Liouville's tautological 1-form**; and σ is called the **symplectic 2-form**. Let us call $X_H = (\frac{\partial H}{\partial \mathbf{p}}, -\frac{\partial H}{\partial \mathbf{q}})$. One can check that

$$i_{X_H} \sigma = -dH. \tag{5.49}$$

One can easily verify that $\mathscr{L}_{X_H} \sigma = 0$ using Cartan's formula.

By the way, the relationship between the symplectic form and the Poisson bracket is

$$\{F, G\} = \sigma(X_F, X_G). \tag{5.50}$$

Let $\phi_t \colon T^*Q \to T^*Q$ denote the flow generated by X_H. Then the vanishing Lie derivative implies that $\phi_t^* \sigma = \sigma$. In other words, ϕ_t preserves the symplectic form; such ϕ_t is said to be a **symplectomorphism**. Therefore, the flux of σ remains constant on any surface bordered by C_t.

Also note that the $2m$-dimensional volume form in a canonical coordinate (as the standard $\mathbb{R}^{2m}$ determinant) can be expressed as $\mu_{T^*Q} = \underbrace{\sigma \wedge \cdots \wedge \sigma}_{m}$. Since pullback distributes over wedge products, we have

$$\phi_t^* \mu = \mu. \tag{5.51}$$

Theorem 5.7 — Liouville's Theorem of Volume Preservation. The flow given by (5.43) on the phase space is volume preserving. Equivalently, the $2m$-dimensional vector field $X_H = (\frac{\partial H}{\partial \mathbf{p}}, -\frac{\partial H}{\partial \mathbf{q}})$ is divergence-free.

If $\rho_0 \in \Omega^{2m}(T^*Q)$ is some probability (mass) distribution for the initial condition for (5.43), then how would this probability ρ_t evolve over time t? Let us represent $\rho_t = f_t \mu$ using a distribution function f_t. Since $\phi_t^* \mu = \mu$, and by mass conservation we also have $\phi_t^* \rho_t = \rho_0$. This implies $\phi_t^* f_t = f_0$. Note that pullback of functions is just function composition $\phi^* f_t = f_t \circ \phi_t$. In the context of dynamical system on the phase space, ϕ_t^* is called the **Koopman operator**. The equation of motion for f_t that ensures $\phi_t^* f_t = f_0$ is **Liouville's equation**

$$\frac{\partial}{\partial t} f_t + \mathscr{L}_{X_t} f_t = 0 \quad \Leftrightarrow \quad \frac{\partial}{\partial t} f_t = -\{f, H\}. \tag{5.52}$$

The equation is similar to the Poisson system. The difference is that in Poisson system the observable is fixed while the observer is moving; here in Liouville's equation the density f is moving while the observer is fixed. That gives an extra minus sign.

Let $g\colon T^*Q \to \mathbb{R}$ be an observable, and let ρ_t be a probability distribution evolved by the Hamiltonian dynamical system. Then the expectation $\mathbb{E}(g) = \int_{T^*Q} g(\mathbf{q},\mathbf{p})\rho_t$ has the rate of change

$$\frac{d}{dt}\mathbb{E}(g) = \mathbb{E}(\mathscr{L}_{X_H} g) = \mathbb{E}(\{g, H\}). \tag{5.53}$$

5.5.4 Noether's Theorem

Suppose we find a vector field Y on the phase space such that $\mathscr{L}_Y \sigma = 0$. That is, Y will generate symplectomorphisms. This is a continuous symmetry of the phase space: the raw ingredient σ that can turn a Hamiltonian function H into an equation of motion X_H through (5.49) is the same after shifting the space by flowing along Y. If we further find that $dH[\![Y]\!] = 0$, then the dynamical system has a symmetry of Y. Each of such symmetry will lead to a conserved quantity, which is a major revelation discovered by Noether.

Suppose $\mathscr{L}_Y \sigma = 0$. By Cartan's formula $0 = i_Y \underbrace{d\sigma}_{=0} + d i_Y \sigma = 0$, we find that the 1-form $i_Y\sigma$ is closed. Hence, it is locally the differential of a function $i_Y\sigma = dJ$. The function $J = J_Y$ is called the **Noether current** associated to Y. In fact, J_Y depends linearly in Y if we have a linear space of Y's.

Theorem 5.8 Let $J\colon T^*Q \to \mathbb{R}$ be the Noether current of an infinitesimal symplectomorphism Y on the phase space T^*Q. Suppose, in addition, that $dH[\![Y]\!] = 0$. Then under (5.43), we have

$$\frac{d}{dt} J(\mathbf{q}(t), \mathbf{p}(t)) = 0. \tag{5.54}$$

Examples of Noether's theorem include that the translational invariance of a physical system implies the conservation of total linear momentum, and that the rotational invariance implies the conservation of total angular momentum.

5.5.5 Hamilton–Jacobi equation

There is an interesting relation between the action and the Hamiltonian. In the least action picture, let q_0 be fixed, but let q_T at time T be a variable. We will call $q_T = y \in Q$. Each choice of (T, y) gives rise to a least action path connecting $(0, q_0)$ and (T, y). This defines an action function depending on the choice of destination $(T, y) \in \mathbb{R} \times Q$:

$$S(T, q) \coloneqq \int_0^T L(q(t), \dot{q}(t))\, dt, \quad q(t) \text{ satisfies (5.34) and } q(0) = q_0, q(T) = y. \tag{5.55}$$

Theorem 5.9 $S(T, y)$ satisfies the **Hamilton–Jacobi equation**

$$\frac{\partial S}{\partial T} = -H\left(y, \frac{\partial S}{\partial y}\right). \tag{5.56}$$

In fact, many PDE (especially in fluid dynamics, control system, *etc.*) for functions $u(t, \mathbf{x})$ on space and time takes the form of

$$\frac{\partial u}{\partial t} + H\left(u, \frac{\partial u}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}\right) = 0. \tag{5.57}$$

A way to solve (or at least simplify) the equation is to convert this Hamilton–Jacobi equation back to an ODE about least action paths. These paths are called characteristic lines. This technique is called the **method of characteristics**.

6. Incremental Potential

Section 5.4 provides a first-principle method for deriving the equation of motion for a conservative and non-dissipative system. In this chapter, we consider a different first-principle approach to deriving the dynamics that turns out to be more general. This approach first discretizes time, and then turns each variable update of each time step into an optimization problem. Note that this optimization problem happens only in each small time interval of time step, instead of a global path optimization like Section 5.4 or Section 5.2. As we focus on modeling the optimization problem in each small time interval, we can describe dissipative forces such as friction, and impose constraints for collision and contact.

6.1 Conservative Systems

For simplicity, we first consider a conservative system, determined only by an inertia and a potential. Let Q be the space of all positions equipped with an inertia metric $\flat$. Let $U\colon Q \to \mathbb{R}$ be a potential energy. Then the equation of motion is given by

$$\flat \overset{\nabla}{\ddot{q}} = f(q), \quad f = -dU. \tag{6.1}$$

Let us suppose we are under a coordinate of Q, so that $Q \cong \mathbb{R}^m$ where the inertia $\flat$ is just a constant symmetric positive definite matrix $\mathbf{M} \in \mathbb{R}^{m\times m}$. Then the equation of motion becomes

$$\mathbf{M}\ddot{\mathbf{q}}(t) = \mathbf{f}(\mathbf{q}(t)) = -\nabla U\big|_{\mathbf{q}(t)}, \tag{6.2}$$

where ∇ denotes the coordinate gradient $(\nabla U)_i = \frac{\partial U}{\partial q^i}$.

Now we discretize time into $t_0, t_1, \ldots$, where $t_k = kh$ for some fixed step size $h > 0$. Let us denote $\mathbf{q}(t_k) = \mathbf{q}^{(k)}$. Our goal is to determine $\mathbf{q}^{(k+1)}$ given $\mathbf{q}^{(0)}, \ldots, \mathbf{q}^{(k)}$. The

backward Euler method discretizes (6.2) into the following equation for solving the unknown $\mathbf{q}^{(k+1)}$

$$\mathbf{M}\frac{\mathbf{q}^{(k-1)} - 2\mathbf{q}^{(k)} + \mathbf{q}^{(k+1)}}{h^2} = \mathbf{f}(\mathbf{q}^{(k+1)}). \tag{6.3}$$

Let us perform the following renaming of variables

$$\mathbf{v}^{\text{old}} = \mathbf{v}^{(k-1/2)} := \frac{\mathbf{q}^{(k)} - \mathbf{q}^{(k-1)}}{h}, \quad \mathbf{v}^{\text{new}} = \mathbf{v}^{(k+1/2)} := \frac{\mathbf{q}^{(k+1)} - \mathbf{q}^{(k)}}{h}, \tag{6.4}$$

$$\mathbf{q}^{\text{pre}} := \mathbf{q}^{(k)} + h\mathbf{v}^{\text{old}}. \tag{6.5}$$

Here $\mathbf{q}^{\text{pre}}$ means a prediction of $\mathbf{q}^{(k+1)} = \mathbf{q}^{(k)} + h\mathbf{v}^{\text{new}}$ estimated using the old velocity $\mathbf{v}^{\text{old}}$. That is, the prediction is estimated with a pure inertial motion. Then (6.3) is equivalent to the following equation

$$\frac{1}{h^2}\mathbf{M}\left(\mathbf{q}^{(k+1)} - \mathbf{q}^{\text{pre}}\right) = \mathbf{f}(\mathbf{q}^{(k+1)}). \tag{6.6}$$

One can check that solving (6.6) for $\mathbf{q}^{(k+1)}$ is equivalent to solving a minimization

$$\mathbf{q}^{(k+1)} = \underset{\mathbf{q}}{\text{argmin}}\ \frac{1}{2h^2}\left(\mathbf{q} - \mathbf{q}^{\text{pre}}\right)^{\intercal}\mathbf{M}\left(\mathbf{q} - \mathbf{q}^{\text{pre}}\right) + U(\mathbf{q}). \tag{6.7}$$

That is, the new position $\mathbf{q}^{(k+1)}$ is the **proximal map** of the prediction position with respect to the potential energy with the metric defined by the inertia.

Another equivalent formulation is the following. The problem of finding $\mathbf{q}^{(k+1)}$ can be rephrased as finding the new velocity $\mathbf{v}^{\text{new}}$, and the new position is just $\mathbf{q}^{(k+1)} = \mathbf{q}^{(k)} + h\mathbf{v}^{\text{new}}$. Then one can check that (6.7) is equivalent to

$$\mathbf{v}^{\text{new}} = \underset{\mathbf{v}}{\text{argmin}}\ \frac{1}{2}\left(\mathbf{v} - \mathbf{v}^{\text{old}}\right)^{\intercal}\mathbf{M}\left(\mathbf{v} - \mathbf{v}^{\text{old}}\right) + U(\mathbf{q}^{(k)} + h\mathbf{v}). \tag{6.8}$$

This formulation (6.8) can be interpreted as that, at each time step, the velocity is chosen to optimize two competing desires:

- The new velocity is close to the old velocity. The cost of changing velocity is $\frac{1}{2}|\mathbf{v}^{\text{new}} - \mathbf{v}^{\text{old}}|^2$ where $|\cdot|^2$ is given by the inertia metric. This is very close to the intuition of inertia.
- The new velocity would yield a new position, and the system wants to have a minimal energy for the new position. The cost of any choice of new velocity has a cost of $U(\mathbf{q}^{(k)} + h\mathbf{v}^{\text{new}})$.

Definition 6.1 The energy, as a function of $\mathbf{v}$ or $\mathbf{q}$,

$$\mathcal{E}^h = \frac{1}{2h^2}\left(\mathbf{q} - \mathbf{q}^{\text{pre}}\right)^{\intercal}\mathbf{M}\left(\mathbf{q} - \mathbf{q}^{\text{pre}}\right) + U(\mathbf{q}) \tag{6.9}$$

$$= \frac{1}{2}\left(\mathbf{v} - \mathbf{v}^{\text{old}}\right)^{\intercal}\mathbf{M}\left(\mathbf{v} - \mathbf{v}^{\text{old}}\right) + U(\mathbf{q}^{(k)} + h\mathbf{v}) \tag{6.10}$$

is called the **incremental potential** (respecting the backward Euler method) for a conservative system with inertia $\mathbf{M}$ and potential U.

Postulate 6.1 The dynamical system governed by an inertia and a potential is a consecutive optimization of the incremental potentials, with the limit $h \to 0$.

The incremental potential formulation is not limited to the backward Euler method. For example, the **Newmark scheme**[1] and the **PDF** family can be written in an incremental potential optimization. We will stick with the backward Euler method as it is more intuitive.

6.1.1 Coordinate free formulation

Let us write down the formal coordinate-free incremental potential on a general manifold Q with an inertia metric $\flat$ and a potential U. Note that without being in $\mathbb{R}^m$, we cannot perform addition and subtraction on positions. So, we employ the notion of exponential map in differential geometry.

Definition 6.2 Let $(M, \flat)$ be a Riemannian manifold. For each point $p \in M$ define $\exp_p \colon T_pM \to M$:

$$\exp_p(\vec{v}) \coloneqq x(1), \tag{6.11}$$

where $x \colon [0,1] \to M$ is the solution to the geodesic equation:

$$\overset{\nabla}{\ddot{x}} = 0, \quad x(0) = p, \quad \dot{x}(0) = \vec{v}. \tag{6.12}$$

Call its inverse the logarithm map.

Definition 6.3 Let $\log_p \colon M \to T_pM$ be the inverse function of $\exp_p \colon T_pM \to M$.

The incremental potential for a conservative system can be written as follows. Let Q be the space of all positions with inertia metric $\flat$. Let $U \colon Q \to \mathbb{R}$ be a potential energy. Then given a current position $q^{(k)}$, an old velocity $\vec{v}^{\text{old}} \in T_{q^{(k)}}Q$, and a step size, define

$$\mathcal{E}^h \coloneqq \frac{1}{2} \left| \vec{v} - \vec{v}^{\text{old}} \right|_\flat^2 + U\left(\exp_{q^{(k)}}(h\vec{v})\right) \tag{6.13}$$

where $q = \exp_{q^{(k)}}(h\vec{v})$. The new velocity arising from optimizing the incremental potential is an element of $T_{q^{(k)}}Q$. For the next iteration, one needs the velocity vector to be in $T_{q^{(k+1)}}Q$. This can be achieved by a parallel transport of the velocity along the geodesic path connecting $q^{(k)}$ and $q^{(k+1)}$.

6.2 Dissipative Systems

A variational principle for a dissipative system is the **maximum dissipation principle** explored by Onsager. Without diving into the statistical thermodynamic detail, the principle takes a simple form in the incremental potential formalism. All one has to

[1] We refer the readers to the paper "Variational Integrators and the Newmark Algorithm for Conservative and Dissipative Mechanical Systems" by Kane, Marsden, Ortiz and West, 2000.

do is to add an extra dissipation term in the incremental potential (6.10):

$$\mathcal{E}^h := \frac{1}{2}\left(\mathbf{v}-\mathbf{v}^{\text{old}}\right)^\intercal \mathbf{M}\left(\mathbf{v}-\mathbf{v}^{\text{old}}\right) + U(\mathbf{q}^{(k)}+h\mathbf{v}) + hR_{\mathbf{q}^{(k)}}(\mathbf{v}) \tag{6.14}$$

where

$$R\colon TQ \to \mathbb{R}_{\geq 0} \tag{6.15}$$

is called the **resistance function** or **Rayleigh's dissipation function**, which, at each position, is a (convex) function of velocity that satisfies

$$R_{\mathbf{q}}(\mathbf{0}) = 0. \tag{6.16}$$

In simple cases (referred to as the **viscosity law**), the Rayleigh dissipation function is modeled as a quadratic form

$$R_{\mathbf{q}}(\mathbf{v}) = \frac{1}{2}\mathbf{v}^\intercal \mathbf{R}_{\mathbf{q}}\mathbf{v} \tag{6.17}$$

for some symmetric positive semi-definite matrix $\mathbf{R}$ (that may depend on position).

Postulate 6.2 The dynamical system governed by an inertia $\mathbf{M}$, a potential U, and a dissipation rate D, is a consecutive optimization of the incremental potential defined by (6.14), with the limit $h \to 0$.

One can check that the corresponding force model is

$$\mathbf{f}(\mathbf{q},\mathbf{v}) = -\nabla U\big|_{\mathbf{q}} - \frac{\partial R_{\mathbf{q}}(\mathbf{v})}{\partial \mathbf{v}}. \tag{6.18}$$

Note that the second term is a partial derivative only on the $\mathbf{v}$ variable, and not on the $\mathbf{q}$ variable of D.

Theorem 6.1 The rate of change of the sum of the kinetic energy and potential energy is given by

$$\frac{d}{dt}\left(\frac{1}{2}\flat_{q(t)}(\dot{q}(t))(\dot{q}(t)) + U(q(t))\right) = -\left\langle \frac{\partial R_q(v)}{\partial v}\middle| \dot{q}\right\rangle \tag{6.19}$$

which is minus the directional derivative of the dissipation function at $\dot{q}$.

For quadratic (the case for **viscosity** or **lubricated friction**) Rayleigh dissipation function $R_{\mathbf{q}}(\mathbf{v})$, the corresponding force $-\frac{\partial R}{\partial \mathbf{v}}$ is a linear function of $\mathbf{v}$. Another type of interesting dissipation functions consists of functions whose function graph is a convex cone, *e.g.* $R(\mathbf{v}) = |\mathbf{v}|_C$ for some norm $|\cdot|_C$. In that case, the slope $\frac{\partial R}{\partial \mathbf{v}}$ would have an upper bound; *i.e.* the corresponding force can take values only with a bounded magnitude. Such dissipative forces model the **Coulomb's law of dry friction** and the **Prandtl–Reuss law for perfect plasticity**.

In the simple quadratic viscosity dissipation, R is differentiable and the dissipative force $-\frac{\partial R}{\partial \mathbf{v}}$ is a single-valued function. On the other hand, in the case of dry friction

and plasticity, the function graph of $R(\mathbf{v})$ is a cone that is not differentiable at the origin. Formally, $-\frac{\partial R}{\partial \mathbf{v}}$ is the subdifferential that can take any value in a convex set representing all slopes that support the cone at the tip. So, how does one determine that actual value of the friction force $-\frac{\partial R}{\partial \mathbf{v}}$? Well, in that case, one returns back to optimizing the incremental potential (6.14), which defines the dynamics without relying on the differentiability of R.

■ **Example 6.1 — Mass-spring-damper model.** A mass-spring-damper system is a mechanical model that consists of a mass (m), a spring with a spring constant (k), and a damper with a damping coefficient (b). The mass is connected to a fixed support via the spring and the damper. This system is commonly used to represent and analyze a wide range of physical systems, such as vehicle suspensions, bridges, and building structures.

Here is a classical exposition. When the mass is displaced from its equilibrium position, the spring and damper exert forces on the mass to restore it to its equilibrium position. The spring force is proportional to the displacement of the mass from its equilibrium position, while the damper force is proportional to the velocity of the mass. Let's define $x(t)$ as the displacement of the mass from its equilibrium position at any given time t. The spring force (F_s) acting on the mass is given by Hooke's law:

$$F_\mathrm{s} = -kx(t) \tag{6.20}$$

where k is the spring constant. The negative sign indicates that the force is opposite to the direction of displacement. The damping force (F_d) acting on the mass is given by:

$$F_\mathrm{d} = -bv(t) \tag{6.21}$$

where b is the damping coefficient and $v(t)$ is the velocity of the mass. The negative sign indicates that the force opposes the motion of the mass. Now, we'll apply Newton's second law of motion to the mass-spring-damper system:

$$F_\mathrm{total} = ma(t) \tag{6.22}$$

and the total force acting on the mass is the sum of the spring and damping forces $F_\mathrm{total} = F_\mathrm{s} + F_\mathrm{d}$. Substituting $a(t) = \frac{d^2x(t)}{dt^2}$, $v = \frac{dx(t)}{dt}$, and expressions for F_s and F_d, we get:

$$m\ddot{x}(t) + b\dot{x}(t) + kx(t) = 0. \tag{6.23}$$

This mass-spring-damper model can also be derived using variational principle. The kinetic energy, potential energy, and dissipation function are respectively given by

$$T(v) = \frac{1}{2}m\dot{v}^2, \quad U(x) = \frac{1}{2}kx^2, \quad R(v) = \frac{1}{2}bv^2. \tag{6.24}$$

We arrive at the same equation of motion after expanding

$$\frac{d}{dt}\frac{\partial T(v)}{\partial v} = -\frac{\partial U(x)}{\partial x} - \frac{\partial R(v)}{\partial v} \tag{6.25}$$

■

■ **Example 6.2 — Dry friction.** Dry friction model is a classical model that describes the interaction between two surfaces in contact, as they experience relative motion or attempt to move against each other. The model takes into account the static and kinetic friction forces between the two surfaces.

Physically, the dry friction model differentiates between two types of friction: static friction and kinetic friction. Static friction is the force that prevents the surfaces from sliding against each other when there is no relative motion. Kinetic friction, on the other hand, is the force that opposes the motion when the surfaces are already sliding against each other. The model assumes the following properties discovered from experiments:

- Amontons' First Law: The friction force is proportional to the normal force acting between the surfaces.
- Amontons' Second Law: The friction force is independent of the area of contact.
- Coulumb's Law of Friction: Kinetic friction is independent of the sliding velocity once the motion starts.

Coulomb's dry friction model can be expressed as follows:

$$F_{\mathrm{s}} \leq \mu_{\mathrm{s}} F_{\mathrm{n}}, \quad F_{\mathrm{k}} = -\mu_{\mathrm{k}} F_{\mathrm{n}} \frac{v}{|v|} \tag{6.26}$$

where

- F_{s}: static friction force; F_{k}: kinetic friction force;
- $\mu_{\mathrm{s}}, \mu_{\mathrm{k}}$: coefficients of static/kinetic friction (dimensionless quantity)
- F_{n}: normal force.
- v: relative velocity between the objects.

The inequality in the static friction equation represents that the static friction force can adjust itself up to a maximum value ($\mu_{\mathrm{s}} F_{\mathrm{n}}$) to prevent motion. When the applied force exceeds the maximum static friction force, the surfaces start sliding, and the kinetic friction force comes into play.

An approximate additional assumption that is often employed is that $\mu_{\mathrm{s}} = \mu_{\mathrm{k}}$ (which is usually the case between two dry metals). In that case, the friction model is expressed by

$$|F_{\mathrm{f}}| \leq \mu F_{\mathrm{n}}. \tag{6.27}$$

Let us also ignore the dependency of the normal force for now. For instance, we consider a block of mass resting or sliding on a floor with a constant normal force due to its weight. Then there is a distinguished maximal magnitude of friction force $F_{\max} \in \mathbb{R}$, given by μF_{n}, using which we express the Rayleigh dissipation

function

$$R(v_{\text{rel}}) = F_{\max}|v_{\text{rel}}| \tag{6.28}$$

where v_{rel} is the relative velocity between the sliding objects. The friction force conditions $|F_{\text{f}}| \leq \mu F_{\text{n}}$ and $F_{\text{f}} = -F_{\max}\frac{v_{\text{rel}}}{|v_{\text{rel}}|}$ are derived from the derivative of the dissipation function with respect to v_{rel}. ∎

6.3 Quasi-static Processes

A dynamical process of a physical system is called **quasi-static** if the effect of inertia or acceleration is ignored. This simplification is sometimes employed in a dissipative system with an external force, and the external force changes slowly. Then, at each time, the internal force $\mathbf{f}^{\text{int}} = -\nabla U|_{\mathbf{q}}$ of the system is balanced with the external force $\mathbf{f}^{\text{ext}}$:

$$\mathbf{f}^{\text{ext}} = -\mathbf{f}^{\text{int}} = \nabla U\big|_{\mathbf{q}} + \frac{\partial R_{\mathbf{q}}(\mathbf{v})}{\partial \mathbf{v}} \tag{6.29}$$

Quasi-static processes give the following simplification in mathematical modeling:

- Inertia is out of the equation. We no longer need to know the inertia metric $\flat$ (or the mass matrix $\mathbf{M}$).
- Since there is no dependency on metric, using an arbitrary local coordinate does not add any complication. Hence it is justified to assume $Q = \mathbb{R}^m$.

When $\mathbf{q} \in \mathbb{R}^m$, the force are thought of as elements in the dual space $\mathbf{f} \in (\mathbb{R}^m)^*$.

6.3.1 Reversible processes

A quasi-static process is called **reversible** if $\langle \frac{\partial R}{\partial \mathbf{v}} | \mathbf{v} \rangle = 0$ throughout the process.

In particular, in a reversible process, $\mathbf{f}^{\text{ext}} = \nabla U|_{\mathbf{q}}$ is a function of $\mathbf{q}$ and vice versa. That is, if we perform an experiment where the desired static position $\mathbf{q}$ is our independent variable, and the value of the required external force $\mathbf{f}^{\text{ext}}$ is the resulting dependent variable, then we can establish a function $\mathbf{f}^{\text{ext}} = F(\mathbf{q})$ from the experiment. Conversely, if we perform the same experiment but we let $\mathbf{f}^{\text{ext}}$ as the independent variable and resulting equilibrium position $\mathbf{q}$ as the dependent variable, then we would establish another function $\mathbf{q} = G(\mathbf{f}^{\text{ext}})$. These two functions are inverse function of each other: $F = G^{-1}$.

On the $\mathbf{q}$-$\mathbf{f}^{\text{ext}}$ space $T^*Q = \mathbb{R}^m \times \mathbb{R}^{m*}$, we can plot the function graph of F (equivalently G), which is an m dimensional surface in the $2m$-dimensional T^*Q.

We also know that the function F is the differential of a function $F = \nabla U$. This property, as a matter of fact, will ensure that its inverse function $G = F^{-1}$ is also the differential of a function $G = \nabla V$. We call the potential V the **Legendre transform** of the potential U.

Definition 6.4 Let $U : \mathbb{R}^m \to \mathbb{R}$ be a convex function. The **Legendre transform** or the **convex dual function** $V : \mathbb{R}^{m*} \to \mathbb{R}$ of U is given so that $dU : \mathbb{R}^m \to \mathbb{R}^{m*}$ and $dV : \mathbb{R}^{m*} \to \mathbb{R}^m$ are inverse function of each other. A common construction

is given by

$$V(\mathbf{f}) \coloneqq \sup_{\mathbf{q}\in\mathbb{R}^m} \langle \mathbf{f}|\mathbf{q}\rangle - U(\mathbf{q}) \tag{6.30}$$

6.3.2 Irreversible processes

In a irreversible process, $\mathbf{f}^{\text{ext}}$ and $\mathbf{q}$ are no longer functions of each other. The value of external force $\mathbf{f}^{\text{ext}}$ of a given position $\mathbf{q}$ (or the value of the position for a given external force) depends on the path we arrive at this pair of position and external force. This phenomenon of path-dependent multimodal position-force relation is called the **hysteresis effect**.

In a irreversible process, it is possible to schedule the external load $\mathbf{f}^{\text{ext}}$ such that the pair $(\mathbf{q}, \mathbf{f}^{\text{ext}})$ traces out a closed loop enclosing a nonzero area in the space T^*Q. Throughout the process, the external force $\mathbf{f}^{\text{ext}}$ is balanced with the internal force $\mathbf{f}^{\text{int}}$, which is decomposed into a conservative force arising from $\frac{\partial U}{\partial \mathbf{q}}$ and the non-conservative part. The total energy that is dissipated along a path γ in the space T^*Q is given by the integral

$$\text{Dissipated Energy} = \int_\gamma \left\langle \mathbf{f}^{\text{ext}} - \frac{\partial U}{\partial \mathbf{q}} \middle| d\mathbf{q} \right\rangle = \int_\gamma \langle \mathbf{f}^{\text{ext}}|d\mathbf{q}\rangle - \int_\gamma dU \tag{6.31}$$

When the path closes up into a closed loop, the second term vanishes $\oint_\gamma dU = \oint_{\partial\gamma} U = 0$ by Stokes' theorem. In that case, total dissipated energy is the enclosed area; let $\gamma = \partial\Omega$:

$$\oint_\gamma \langle \mathbf{f}^{\text{ext}}|d\mathbf{q}\rangle = \iint_\Omega \langle d\mathbf{f}^{\text{ext}} \wedge d\mathbf{q}\rangle. \tag{6.32}$$

■ **Example 6.3 — Perfect plasticity.** The perfect plasticity model, also known as the ideal plasticity model, is a simplified model for describing the behavior of materials under load. This model is particularly useful for analyzing plastic deformation in materials, such as metals, that exhibit both elastic and plastic behaviors.

Physically, perfect plasticity model assumes that a material can undergo elastic deformation up to a certain stress level called the yield stress (yield force). When the stress in the material exceeds the yield stress, it undergoes plastic (permanent) deformation without any increase in stress. In other words, the material can deform plastically without any additional load. The perfect plasticity model also assumes that there is no strain hardening, meaning the material does not strengthen as it undergoes plastic deformation.

A one-dimensional perfect plasticity model can be represented using a simple mechanical system consisting of a linear spring and a dry friction slider in series. This setup will help demonstrate how a material can undergo both elastic and plastic deformations. The linear spring represents the elastic behavior of the material. When the spring is stretched or compressed, it stores potential energy and follows Hooke's law ($f_{\text{e}} = -kx_{\text{e}}$), where f_{e} is the elastic force, k is the spring constant, and x_{e} is the stretch of the spring. The dry friction slider represents the

plastic behavior of the material. The slider is connected to the spring and moves on a surface with friction. When the applied force exceeds the static friction force, the slider starts to move, and plastic deformation occurs.

The one-dimensional perfect plasticity model can be demonstrated using the following process. Let the total displacement $x = x_\mathrm{e} + x_\mathrm{p}$ be the sum of the elastic displacement x_e and the plastic displacement x_p, which is how much the slider has moved. Let f_max be the maximal static friction.

1. Apply an external force (f^ext) to the coupled spring-slider system.
2. Calculate the spring force ($f_\mathrm{e} = -kx_\mathrm{e}$) based on the elastic displacement (x_e).
3. If $|f^\mathrm{ext}| < f_\mathrm{max}$, the system remains in the elastic region, and no plastic deformation occurs. The spring deforms elastically, $x_\mathrm{e} = f^\mathrm{ext}/k$ and the slider x_p does not move.
4. If $|f^\mathrm{ext}| = |f_\mathrm{e}|$ reaches f_max, the system enters the plastic region, and the slider starts moving.
5. When the external force is reduced, the system returns to elastic region, in which case the elastic displacement follows $x_\mathrm{e} = f^\mathrm{ext}/k$.
6. When the system if fully unloaded $f^\mathrm{ext} = 0$, the elastic displacement is zero but the total displacement $x = x_\mathrm{e} + x_\mathrm{p} = 0 + x_\mathrm{p}$ is nonzero x_p as a memory that the system has undergone plastic deformation.

This 1D elastoplasticity system can also be described using variational principle with the following setting:

- The total stretching is additively split into $x = x_\mathrm{e} + x_\mathrm{p}$.
- There is a set of internal variable, which is the plastic displacement x_p in this case.
- The potential energy is given by $U(x_\mathrm{e}, x_\mathrm{p}) = \frac{1}{2}kx_\mathrm{e}^2$.
- The dissipation function is given by $R(x_\mathrm{e}, x_\mathrm{p}; \dot{x}_\mathrm{e}, \dot{x}_\mathrm{p}) = f_\mathrm{max}|\dot{x}_\mathrm{p}|$.

■

7. Constrained Systems

Constrained dynamical systems are a class of systems in which the motion of particles or components is subject to certain restrictions or constraints, often arising from physical, geometrical, or mechanical considerations. We classify constraints into two categories similar to those for optimization problems: equality constraints and inequality constraints. **Equality constraints** limit the degrees of freedom of the system, dictating its behavior and motion. Examples of constrained dynamical systems include mechanical linkages, robotic manipulators, and particles moving on a curved surface or along a specified path. **Inequality constraints**, not necessarily limiting the degrees of freedom of the system, describe a set of admissible configurations. When the configuration reaches the boundary of the admissible set, collision and contact occurs.

7.1 Equality Constraints

In classical mechanics, equality constraints can be classified as holonomic or non-holonomic. **Holonomic constraints** are those that can be expressed as mathematical equations relating the generalized coordinates of the system, and often depend on the configuration of the system rather than the velocities. **Non-holonomic constraints**, on the other hand, involve the velocities of the system and may not be integrated into a simple equation relating the generalized coordinates.

7.1.1 Holonomic constraints

Given a configuration space Q, a holonomic contraint is described as

$$G(q) = \mathbf{0}_{\mathbb{R}^c} \quad \text{for some function } G\colon Q \to \mathbb{R}^c. \tag{7.1}$$

If Q is m dimensional, the constrained space is the $(m-c)$-dimensional zero-level set surface $C = \{q \in Q \mid G(q) = \mathbf{0}\}$, assuming $dG\colon TQ \xrightarrow{\text{linear}} \mathbb{R}^c$ is fullrank (rank-c)

at the zero level set. Geometrically, one can say that a holonomic constraint is a codimension-c submanifold $C \subset Q$.

For a non-dissipative system where the dynamics is described solely by an inertia metric and a potential energy, the subsequent dynamical system can be derived straightforwardly as the submanifold C has an induced inertia metric and potential energy inherited from Q.

If one represents C implicitly as the level set of a function G, then we may write the equation of motion for $q \in Q$ constrained by G with an additional **Lagrange multiplier** term as we minimize the action (5.33):

$$dS_q[\![\mathring{q}]\!] = \int_a^b \left(-\left\langle \flat_{q(t)} \overset{\nabla}{\ddot{q}}(t) \middle| \mathring{q}(t) \right\rangle - \langle dU_{q(t)} | \mathring{q}(t) \rangle - \langle \lambda | (dG_{q(t)})[\![\mathring{q}(t)]\!] \rangle \right) dt \tag{7.2}$$

$$= \int \left\langle -\flat_{q(t)} \overset{\nabla}{\ddot{q}}(t) - dU_{q(t)} - dG^*_{q(t)}\lambda \middle| \mathring{q}(t) \right\rangle dt \tag{7.3}$$

where $\lambda \in \mathbb{R}^{c*}$ is the Lagrange multiplier, and $dG^*_q \colon \mathbb{R}^{c*} \xrightarrow{\text{linear}} T^*_q Q$ is the adjoint of $dG_q \colon T_q Q \xrightarrow{\text{linear}} \mathbb{R}^c$. With least action principle, the equation of motion is given by

$$\begin{cases} \flat_q \overset{\nabla}{\ddot{q}} = -dU - dG^*_q \lambda \\ G(q) = 0 \end{cases} \tag{7.4}$$

The image of dG^*_q is exactly the annihilator of dG_q. Hence one calls $dG^*_q \lambda$ the **normal force** which keeps the state q on C.

7.1.2 Non-holonomic constraints

A non-holonomic constraint is described by assigning a lower dimensional feasible subspace $V_q \subset T_q Q$ of velocities at each position $q \in Q$. One may represent V_q by a linear equation

$$A_q \dot{q} = \mathbf{0}_{\mathbb{R}^c} \quad \text{for some } A_q \colon T_q Q \xrightarrow{\text{linear}} \mathbb{R}^c \tag{7.5}$$

A holonomic constrained system can be written in a non-holonomic form with $A_q = dG_q$.

In geometry, such an assignment of subspaces of tangent spaces is called a **distribution**. A distribution V is said to be **integrable** if there exists a family of surfaces, whose dimension equals to the dimension of V, that foliates the space Q such that V is everywhere tangent to the surface in the family. All one-dimensional distributions (direction fields) are integrable, and the corresponding foliation consists of the integral curves of the direction fields. A generic distribution of dimension greater or equal to 2 is non-integrable.

In fact, a distribution is (locally) integrable if and only if the distribution is closed under the Lie bracket on tangent vector fields in the distribution. That is, $\Gamma(V) \subset \Gamma(TQ)$ is a Lie subalgebra.

> ■ **Example 7.1 — Rolling ball.** A rolling ball moving on a surface without slipping presents an example of a nonholonomically constrained system. In this case, the constraint arises from the no-slip condition, which dictates that the point of the ball in contact with the surface has zero velocity relative to the surface. Under this

constraint, the 3D orientation of the ball have only 2 degrees of freedom among the 3-dimensional space of all 3D rotations. However, moving in an infinitesimal cycle using the 2 degrees of freedom can yield a net motion into the 3rd degree of freedom. ∎

∎ **Example 7.2 — Car.** The configuration of a car is four dimensional: The $\mathbb{R}^2$ center of mass, the $\mathbb{S}^1$ orientation, and the position of the steering wheel which is in an angular interval $[-a, a]$. At any moment, the allowed motion of the car is 2-dimensional: Driving back and forth and steering the wheel. This 2D distribution is non-integrable, giving rise to a nonholonomic system. One can effectively move the car sideway (parallel parking) by a sequence of maneuver within the constraint distribution. ∎

∎ **Example 7.3 — Quasistatic processes.** The above examples are all special cases of quasistatic process with a special type of friction. The friction is described by a dissipation function that is extremely steep (expensive) in certain direction, and very flat (cheap) in some other direction. For example, one could still directly move a car sideway by pushing it by brute force, but it is just much more expensive. In general, any dissipation function of this type defines a constraint on the velocity in a quasistatic process, and most likely such a constraint is non-holonomic (non-integrable). ∎

∎ **Example 7.4 — Locomotion.** An animal or a robot moves around in an environment by performing a cycle of motion of its movable components. The configuration space Q can be described as a fiber bundle Q over a base manifold B, where the base is the configuration space of movable component of the animal, and each fiber is the group of Euclidean motions (the translation and rotation of the animal). Suppose we are also given a dissipation function over Q. Now, for each motion executed in B, the corresponding motion in the fiber is defined by that with minimal dissipation. This defines a notion of "horizontal subspace" in the fiber bundle, also known as a connection. For each path in the base manifold, there is a unique lifted path in Q such that it is parallel to the horizontal subspace. If these horizontal subspaces are non-integrable, a cyclic motion on the base will lift to a spiral path that has a net Euclidean motion along the fiber. Geometrically, this is called the curvature of the connection. ∎

Nonholonomic constraints can be imposed in the incremental potential framework. One may also directly model these constraints by setting barrier functions in the potential U (as holonomic constraints) and in the Rayleigh dissipation function R (as non-holonomic constraints).

Another well-known approach is **Kane's method of quasivelocities**. Instead of using the coordinates in Q to express some basis for T_qQ, use another basis $\vec{e}_1, \ldots, \vec{e}_m \in T_qQ$ so that $\vec{e}_1, \ldots, \vec{e}_{m-c}$ is tangential to V. Any feasible velocity $\dot{q} \in V$ is parameterized as $\dot{q} = u_1\vec{e}_1 + \cdots u_{m-c}\vec{e}_{m-c}$. The list of coefficients $\mathbf{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_{m-c} \end{bmatrix}$ is called a quasivelocity. (It is called quasivelocity instead of a velocity since the basis $\vec{e}_i$ might not be integrable into a coordinate system; *i.e.* quasivelocity is not really

any velocity after some change of coordinates.) In terms of it, the equation of motion takes the form of $\mathbf{M}\overset{\nabla}{\dot{\mathbf{u}}} = \mathbf{f}$ with some induced inertia $\mathbf{M} \in \mathbb{R}^{(m-c)\times(m-c)}$ (possibly requiring Levi-Civita connection ∇ expressed in this basis), and the equation is coupled with $\dot{q} = u_1\vec{e}_1 + \cdots u_{m-c}\vec{e}_{m-c}$. The force $\mathbf{f}$ is given by the covector $-dU$ expanded in the dual basis of $\vec{e}_1, \ldots, \vec{e}_{m-c}$.

Typical examples for quasivelocities are the rates of changes in the movable joints in a locomotion system.

7.2 Inequality Constraints

Dynamical systems with inequality constraints are also called **impulsive system** where collision and contact are of primary interest. They naturally arise for the non-interpenetration condition. For example the condition can be written as $\phi \geq 0$ where ϕ is the distance function between moving objects. The inequality constraints define a set of admissible configurations $\Omega \subset Q$, $\Omega = \{q \in Q \,|\, \phi(q) \geq 0\}$.

In general the function ϕ is a map $\phi\colon Q \to \mathcal{V}$ for some c-dimensional vector space $\mathcal{V}$ in which a convex cone (apexed at the origin) is defined $\mathcal{C} \subset \mathcal{V}$. (For example $\mathcal{V} = \mathbb{R}$ and $\mathcal{C} = [0, \infty)$.) The cone $\mathcal{C}$ defines a polar cone $\mathcal{C}^\circ = \{\lambda \in \mathcal{C}^* \,|\, \langle\lambda|\vec{a}\rangle \leq 0, \forall\, \vec{a} \in \mathcal{C}\}$ in the dual space. (For example $[0, \infty)^\circ = (-\infty, 0] \subset \mathbb{R}^*$.) The induced Lagrange multiplier force that enforces the inequality constraint takes the form

$$-d\phi_q^*\lambda \in T_q^*Q, \quad \langle\lambda|\phi\rangle = 0. \tag{7.6}$$

The condition $\langle\lambda|\phi\rangle = 0$ is equivalent to the condition that only one of $\lambda \in \mathcal{C}^\circ$, $\phi \in \mathcal{C}$ are allowed to be in the interior of the cone. In the case of $\mathcal{C} = [0, \infty)$, this condition translates to that

$$\lambda \leq 0, \quad \phi \geq 0, \quad \lambda\phi = 0. \tag{7.7}$$

For each $q \in \Omega \subset Q$, let $\mathcal{A}_q \subset T_qQ$ be the cone of admissible velocities, which are $\vec{v} \in T_qQ$ so that $q + \epsilon\vec{v}$ stays in Ω for $\epsilon \to 0$. Concretely,

$$\mathcal{A}_q \coloneqq \{\vec{v} \in T_qQ \,|\, d\phi_q[\![\vec{v}]\!] \geq 0\}\,. \tag{7.8}$$

Note that the set of all possible $d\phi^*\lambda \in T^*Q$ (which is a nontrivial ray or cone at the boundary of Ω and otherwise zero) form the outward normal cone, which is the same as the polar cone of the cone $\mathcal{A}_q$ of admissible velocities

$$\mathcal{N}_q \coloneqq \{(d\phi_q)^*\lambda \,|\, \lambda \in \mathcal{C}^\circ\} = \mathcal{A}^\circ. \tag{7.9}$$

All possible values for **normal forces** or **normal impulses** are given by $-\mathcal{N}_q$.

If a point q at the boundary of Ω describes a contact configuration between two physical materials that experiences friction, then each normal force $f_\mathrm{n} = -d\phi_q^*\lambda \in T_q^*Q$ can also be accompanied by a tangential force f_t in the physical space.

In an example where $f_\mathrm{n}, f_\mathrm{t} \in \mathbb{R}^3$ we have

$$\langle f_\mathrm{n}, f_\mathrm{t}\rangle_{\mathbb{R}^3} = 0, \quad |f_\mathrm{t}|_{\mathbb{R}^3} \leq \mu|f_\mathrm{n}|_{\mathbb{R}^3} \tag{7.10}$$

where μ is the friction coefficient. The set of all $f = f_\mathrm{n} + f_\mathrm{t}$ that satisfy (7.10) and that f_n is oriented inward (as a covector) into the feasible domain form a cone called the **friction cone**.

In general, given an admissible set $\Omega \subset Q$, we assign a friction cone $\mathcal{F}_q$ (apexed at the origin) in T_q^*Q at each $q \in \partial\Omega$ subject to the following conditions

- $\mathcal{F}_q \supset -\mathcal{N}_q$, *i.e.* it contains the cone of normal forces;
- $\mathcal{F}_q$ is convex.

For q in the interior of Ω, the friction cone is simply $\mathcal{F}_q = \{0\} \in T_q^*Q$. The friction cone is the space where we can draw additional forces or impulses to redirect the momentum.

7.2.1 Simple Impact

Postulate 7.1 — Impact. We say that a collision occurs when the current state $(q, \vec{v} = \dot{q}) \in TQ$ satisfies $q \in \partial Q$ and $\vec{v} \notin \mathcal{A}_q$. When $(q^\mathrm{old}, \vec{v}^\mathrm{old})$ is at a collision state, the state is instantaneously replaced by some $(q^\mathrm{new}, \vec{v}^\mathrm{new})$ so that

- $q^\mathrm{new} = q^\mathrm{old}$; (we will call it q;)
- $\vec{v}^\mathrm{new} \in \mathcal{A}_q$;
- The impulse $J \coloneqq \flat(\vec{v}^\mathrm{new} - \vec{v}^\mathrm{old})$ belongs to the friction cone $J \in \mathcal{F}_q$;
- Energy non-increasing $|\vec{v}^\mathrm{new}|_\flat^2 \le |\vec{v}^\mathrm{old}|_\flat^2$; (other potential energies or external force do not make any effect here in this instantaneous collision;)

The feasible set of impulses J satisfying the energy non-increasing condition is the interior of an ellipsoid that passes through the origin and is centered at $-\flat\vec{v}^\mathrm{old}$. The condition that $\vec{v}^\mathrm{new} \in \mathcal{A}_q$ reduces the impulse set by intersecting the ellipsoid with a half space (in case $\mathcal{A}_q$ is a half space) separated by a plane passing through the center of the ellipsoid. Finally, the condition $J \in \mathcal{F}$ restricts this half ellipsoid to its intersection with the friction cone.

Additional energy limitation should also be employed, which is not explicitly listed above. For example, one can define a subset in the cone of normal forces $\mathcal{E} \subset -\mathcal{N}$ as the set of pure elastic impulses. Any impulse J in the feasible set is decomposed into $J = J_\mathrm{e} + J_\mathrm{d}$ where $J_\mathrm{e} \in \mathcal{E}$ is conservative and J_d is dissipative. The decomposition can be determined as the one that minimizes the dissipation $\langle J_\mathrm{d} | \frac{\vec{v}^\mathrm{old} + \vec{v}^\mathrm{new}}{2} \rangle$. Knowing the total dissipated energy, one also asserts that $|\vec{v}^\mathrm{new}|_\flat^2 \le |\vec{v}^\mathrm{old}|^2 - \langle J_\mathrm{d} | \frac{\vec{v}^\mathrm{old} + \vec{v}^\mathrm{new}}{2} \rangle$.

Heuristic approaches are often employed. For example, one may fix a restitution coefficient $0 \le e \le 1$ and assert that the normal component of the velocity is always scaled by e before and after the collision. This usually determines a normal force, which then gives rise to a horizontal friction force that shrinks the horizontal velocity or completely stops it (depending on whether the friction is smaller than the maximal static friction).

7.2.2 Contact

A more general type of contact is that two bodies may touch but not bouncing off from each other. For example, one body may be resting or sliding on another body, and it is the potential energy (such as gravity) that keep them stay in contact. In that case, the potential term and external foces should also be put into consideration.

Postulate 7.2

III Continuum Mechanics

IV Geometric Mechanics

V

Appendix

Index