



Fitzwilliam College

Summer School Journal

2024 Summer

Cambridge, 28th July - 10th August



◆ About the Journal

This is an annual journal publishing the best essays written by the top-performing students of the **FitzEd Summer School** which is the **official summer programme of Fitzwilliam College in the University of Cambridge**. The journal covers eleven fields: Mathematics, Physics, Computer Science, Microbiology, Elements of Mathematical Economics, Engineering, Chemistry, Medicine, Palaeobiology, Philosophy, and Nuclear Engineering.

◆ About the Programme

The FitzEd Summer School is the official Summer School Programme of Fitzwilliam College in the University of Cambridge. It is the **only summer programme in Cambridge where all courses are exclusively taught by academics from the University of Cambridge or one of its constituent colleges**. The same academics who lecture and supervise undergraduate students at Cambridge teach undergraduate-level content at the FitzEd Summer School. This programme is designed to **provide students with a flavour of undergraduate study at Cambridge**, and an opportunity to **explore topics beyond what is covered within the school curriculum**.

The core of Fitzwilliam's academic activities is a desire to retain 'the best of the old', while enthusiastically embracing 'the best of the new'. Fitzwilliam has always been characterised by discussion, debate and creativity of ideas and full participation should form a positive, rewarding and sustainable part of an academic course.

◆ List of Academic Course Instructors

DR ANDREA CHLEBIKOVA

DR ASHLEIGH L WISWMAN

DR OLIVER E DEMUTH

DR LAURA VAN HOLSTEIN

PROFESSOR MATTHEW J. MASON

DR AARON D'SA

DR SAEED KAYHANIAN

DR MILES STOPHER

DR VASILEIOS KOTSIDIS

DR JOAO RODRIGUES

MRS SERENA POVIA

DR STEPHEN SAWIAK

DR ANDREA GIUSTI

DR JOHN FAWCETT

DR ALEX CARTER

DR ASHRAF ZARKAN

Table of Contents

1 Mathematics for the Natural Sciences	Parseval's Theorem for Real Fourier Series HAOTIAN MA	6
	Show how to sum the first N natural numbers YANBIN ZANG	9
	Fourier Transform: Their Meaning and Their Use in the Solution of Differential Equations YURU CHEN	11
	Potential energy of a pendulum PEILING LI	14
2 Physics	The Understanding and Extension of EPR Paradox JIARUI HE	20
	Quantum Tunneling: A Review ELYAS ALBATTAT	26
3 Microbiology	The Pathogenicity and Clinical Relevance of Salmonella typhi MENGHAN XU	32
	The Pathogenicity and Clinical Relevance of Listeria monocytogenes YICHEN JEN	36
4 Elements of Mathematical Economics	Risk and reward: An exploration of optimal investments LINGLIN ZHOU	42
	Analyzing Investment Risks and Benefits Using Utility Theory and CRRA SHUOWEN HUANG	45
5 Engineering of Sustainable Vehicles	Gyroscopic Effect CHENGYU WANG	52
	Nuclear Propulsion and its Future Possibilities RUOXI WANG	57
6 Physical Chemistry	Quantum tunnelling and chemistry LUOFAN WU	62
	Explain how the presence of a magnetic field can affect reaction kinetics (radical pair mechanism) SZE WU WANG	66

7 Medicine	How does the body respond to haemorrhage? WAI WONG	69
	How Can the Different Gasses in the Anesthetic Circuit be Measured? Can all Methods be Used for all Gasses? PEILING LI	72
8 Computer Science	Overview of JPEG-1 and JPEG 2000 compression ERIC ZIMING LU	76
	Comparison between Ray Tracing and Radiosity CHANG LIU	85
9 Palaeobiology: Evolution and Behaviour	Mass-speciation and extinction events in Felidae WANZHANG HUANG	93
	The user variability in muscle creation of the shoulder musculature of Coelophysis MEIMEI XIE	96
10 Philosophy in Cambridge: Past and Present	Is Language Perfect? Could There Ever Be a Perfect Language? YIFEI LI	103
	Should impose restrictions on speech? If so,when and how? If not,why not? LIWEI LIN	105
11 Nuclear Engineering	Assessing Public Opinion on Nuclear, Benefits and The Future of Nuclear MALIK ALAMRI	108
	Nuclear Reactors for Medical Isotopes: An Overview FATIMAH AHMED ALABDULLAH	114



01

Mathematics for the Natural Sciences

Parseval's Theorem for Real Fourier Series

HAOTIAN MA

1. A Loose Definition of Parseval's Theorem

This essay discusses a specific property of Fourier series, known as Parseval's theorem. This essay focuses only on Fourier series with real coefficients. Hence, under this circumstance, Parseval's theorem states that the sum of squares of the Fourier coefficients of a function is π times the integral of the square of the function, expressed as:

$$\frac{1}{\pi} \int_{-\pi}^{\pi} f(x)^2 dx = \frac{a_0^2}{2} + \sum_{i=1}^{\infty} (a_i^2 + b_i^2)$$

2. Proof of Parseval's theorem

The proof is not hard in terms of concept, but it is complex since it includes a great number of calculations. The difficulty of this proof lies in deriving the general formula for $f(x)^2$, as there will be many products of sines and cosines, and then the second step is to integrate those trigonometric functions [1].

In general, the Fourier series of function of $f(x)$ is

$$f(x) = \frac{a_0}{2} + \sum_{i=1}^{\infty} (a_i \cos(ix) + b_i \sin(ix))$$

To integrate $f(x)^2$, we need first to derive its general formula:

$$f(x)^2 = \left[\frac{a_0}{2} + \sum_{i=1}^{\infty} (a_i \cos(ix) + b_i \sin(ix)) \right]^2 = \frac{a_0^2}{4} + a_0 \sum_{i=1}^{\infty} (a_i \cos(ix) + b_i \sin(ix)) + \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} [a_i a_j \cos(ix) \cos(jx) + 2a_i b_j \cos(ix) \sin(jx) + a_i a_j \sin(ix) \sin(jx)]$$

Even though the expression looks quite complex, after integrating, most terms will vanish, this is because the answers of following integrations are always 0:

$$\int_{-\pi}^{\pi} \sin(nx) \cos(mx) dx, \int_{-\pi}^{\pi} \sin(nx) dx, \int_{-\pi}^{\pi} \cos(nx) dx \quad (\text{when } n \neq 0)$$

And only when $n=m$, otherwise, the answers of these two integrations are π , otherwise, they are 0:

$$\int_{-\pi}^{\pi} \cos(nx) \cos(mx) dx, \int_{-\pi}^{\pi} \sin(nx) \sin(mx) dx$$

Thus, integrating $f(x)^2$:

$$\begin{aligned} \frac{1}{\pi} \int_{-\pi}^{\pi} f(x)^2 dx &= \frac{1}{\pi} \int_{-\pi}^{\pi} \frac{a_0^2}{4} dx + \frac{1}{\pi} a_0 \sum_{i=1}^{\infty} \int_{-\pi}^{\pi} a_i \cos(ix) + b_i \sin(ix) dx \\ &+ \frac{1}{\pi} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \int_{-\pi}^{\pi} a_i a_j \cos(ix) \cos(jx) dx + \frac{1}{\pi} 2 \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \int_{-\pi}^{\pi} a_i b_j \cos(ix) \sin(jx) dx \\ &+ \frac{1}{\pi} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \int_{-\pi}^{\pi} b_i b_j \sin(ix) \sin(jx) dx \end{aligned}$$

According to the integrals given below, the integrals in the first summation are all 0. The integrals in the third summation are 0. The integrals in the second and the fourth summations, when $i \neq j$, equals to 0, hence simplify the expression [2]:

$$\frac{1}{\pi} \int_{-\pi}^{\pi} f(x)^2 dx = \frac{1}{\pi} \left[\frac{a_0^2}{4} x \right]_{-\pi}^{\pi} + \frac{1}{\pi} [0]_{-\pi}^{\pi}$$

$$+ \frac{1}{\pi} \sum_{i=1}^{\infty} \int_{-\pi}^{\pi} a_i^2 \cos(ix)^2 dx + \frac{1}{\pi} 2 \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} 0 + \frac{1}{\pi} \sum_{i=1}^{\infty} \int_{-\pi}^{\pi} b_i^2 \sin(ix)^2 dx$$

furtherly simplify it:

$$\frac{1}{\pi} \int_{-\pi}^{\pi} f(x)^2 dx = \frac{1}{\pi} \frac{a_0^2}{4} 2\pi + \frac{1}{\pi} \sum_{i=1}^{\infty} a_i^2 \pi + \frac{1}{\pi} \sum_{i=1}^{\infty} b_i^2 \pi = \frac{a_0^2}{2} + \sum_{i=1}^{\infty} (a_i^2 + b_i^2)$$

3. Applica-on of Parseval's Theorem in the Basel Problem

One important application of Parseval's theorem is solving Basel problem. It was first posed by Pietro Mengoli in 1650 and solved by Leonhard Euler in 1734. The problem is named after a place that is called Basel, which was the hometown of Euler and the Bernoulli family who attempted to solve this problem, however, failed [3].

The Basel problem states that the sum of the reciprocals of squares of all positive integers converges to $\pi^2/6$:

$$\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{1}{1} + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \frac{1}{25} + \dots = \frac{\pi^2}{6}$$

The solution is based on the Fourier series of $f(x)=x$. This is because the coefficients of Fourier series have a format of $1/i$, suggesting a connection to $1/i^2$.

$$f(x) = \sum_{i=i}^{\infty} \frac{2 \times (-1)^{n+1}}{i} \sin(ix)$$

Hence the coefficient can be summarized as $a_i = 0$ and $b_i = \frac{2 \times (-1)^{n+1}}{i}$ for any positive integers i .

According to Parseval's theorem, the sum of squares of the Fourier coefficients of a function is π times as big as the integral of the square of the function:

$$\frac{1}{\pi} \int_{-\pi}^{\pi} f(x)^2 dx = \frac{1}{\pi} \int_{-\pi}^{\pi} x^2 dx = \frac{1}{\pi} \left[\frac{x^3}{3} \right]_{-\pi}^{\pi} = \frac{2}{3} \pi^2$$

$$\frac{a_0^2}{2} + \sum_{i=1}^{\infty} (a_i^2 + b_i^2) = \frac{0^2}{2} + \sum_{i=1}^{\infty} (0^2 + \left(\frac{2 \times (-1)^{n+1}}{i}\right)^2) = \sum_{i=1}^{\infty} \frac{4}{i^2}$$

Therefore:

$$\frac{2}{3} \pi^2 = \sum_{i=1}^{\infty} \frac{4}{i^2} \text{ and hence } \frac{1}{6} \pi^2 = \sum_{i=1}^{\infty} \frac{1}{i^2}$$

4. Another Applica-on of Parseval's Theorem

This application of Parseval's theorem relates to the concept of conservation of energy. As known, Fourier series are sometimes used for modulation and demodulation of voice signals, translating a specific sound wave into summation of numbers of trigonometric functions [4]. For sound waves, intensity is proportional to the square of the amplitude, and the amplitude for trigonometric function is represented by the coefficient in front of it. Expression of it is $I \propto a_i^2, b_i^2$. Therefore, in this case, Parseval's theorem shows that the intensity of the original sound wave is equal to the summation of intensities of all component trigonometric-function sound waves.

5. References

- [1]: Dr. Trefor Bazely. (2022, March 14). Parseval's Identity, Fourier Series, and Solving this Classic Pi Formula [Video]. YouTube. <https://www.youtube.com/watch?v=WPeU34jndSw>
- [2]: Wolfram Research, Inc. (n.d.). Parseval's Theorem -- from Wolfram MathWorld. <https://mathworld.wolfram.com/>

ParsevalsTheorem.html

[3]: Ayoub, Raymond (1974), "Euler and the zeta function", *Amer. Math. Monthly*, 81 (10): 1067–86

[4]: Real world application of Fourier series. (n.d.). Mathematics Stack Exchange.

<https://math.stackexchange.com/questions/579453/real-world-application-of-fourier-series>

Show how to sum the first N natural numbers

YANBIN ZANG

1. Question:

Show how to sum the first N natural numbers. Describe how the difference method can be used to sum series of the form $f(n+1)-f(n)$. Use it to find formulae for the sums of natural numbers squared, cubed and to the fourth.

$$\text{Sum}=1+2+3+\dots+N$$

$$2x\text{Sum}=(1+2+3+\dots+N)+(1+2+3+\dots+N)$$

$$2x\text{Sum}=(1+N)+(2+N-1)+(3+N-2)+\dots+(N+1)=(1+N)+(1+N)+(1+N)+\dots+(1+N)=(1+N) \times N$$

$$\text{Sum}=(1+N) \times N/2$$

This is the method of how to sum the first N natural numbers.

Difference method uses $f(n+1)-f(n)$ to represent a_n .

$$a_1=f(2)-f(1),$$

$$a_2=f(3)-f(2),$$

$$a_n=f(n+1)-f(n).$$

$$\text{Sum}=f(n+1)-f(n)+f(n)-f(n-1)+\dots+f(3)-f(2)+f(2)-f(1)$$

$$=f(n+1)-f(1)$$

In brief, the thinking is to represent series with functions of $f(n)$ then we can know the Sum.

Then we can try to use this method to calculate n^2 , n^3 and n^4

(1) $a_n = n^2$. Find Sum.

$$\text{Assume } n^2=f(n+1)-f(n)$$

$$\text{Make } f(n)=an^3+bn^2+cn+d.$$

(Why we assume an^3 ? a $(n+1)^3-an^3$ will make a of an^3 become zero. There leaves only $3n^2$ and so on. So, if we just assume an^2 , there will be no n^2 after calculation but the left side of the equation is n^2 . It does not paired.)

$$\text{Then, } a(n+1)^3+b(n+1)^2+c(n+1)+d-an^3-bn^2-cn-d=n^2$$

$$3an^2+(3a+2b)n+a+b+c=n^2$$

undetermined coefficient

$$3a=1$$

$$3a+2b=0$$

$$a+b+c=0$$

so we can get

$$a=1/3$$

$$b=-1/2$$

$$c=1/6$$

$$f(n)=1/3n^3-1/2n^2+1/6n$$

$$\text{Sum}=f(n+1)-f(1)$$

$$=1/3(n+1)^3-1/2(n+1)^2+1/6(n+1)-0$$

$$=1/3n^3+1/2n^2+1/6n$$

(2) $a_n = n^3$. Find Sum.



Assume $n^3 = f(n+1) - f(n)$

Make $f(n) = an^4 + bn^3 + cn^2 + dn + e$.

Then, $a(n+1)^4 + b(n+1)^3 + c(n+1)^2 + d(n+1) + e - an^4 - bn^3 - cn^2 - dn - e = n^3$

$$4a + (6a+3b) + (4a+3b+2c)n + a+b+c+d = n^3$$

undetermined coefficient

$$4a=1$$

$$6a+3b=0$$

$$4a+3b+2c=0$$

$$a+b+c+d=0$$

so we can get

$$a=1/4$$

$$b=-1/2$$

$$c=1/4$$

$$d=0$$

$$f(n) = 1/4n^4 - 1/2n^3 + 1/4n^2$$

$$\text{Sum} = f(n+1) - f(1)$$

$$= 1/4(n+1)^4 - 1/2(n+1)^3 + 1/4(n+1)^2 - 0$$

$$= 1/4n^4 + 1/2n^3 + 1/4n^2$$

(2) $a_n = n^4$. Find Sum.

Assume $n^4 = f(n+1) - f(n)$

Make $f(n) = an^5 + bn^4 + cn^3 + dn^2 + en + f$.

Then, $a(n+1)^5 + b(n+1)^4 + c(n+1)^3 + d(n+1)^2 + e(n+1) + f - a n^5 - b n^4 - c n^3 - d n^2 - e n - f = n^4$

$$5a n^4 + (10a+4b)n^3 + (10a+6b+3c)n^2 + (5a+4b+3c+2d)n + a+b+c+d+e = n^4$$

undetermined coefficient

$$5a=1$$

$$10a+4b=0$$

$$10a+6b+3c=0$$

$$5a+4b+3c+2d=0$$

$$a+b+c+d+e=0$$

so we can get

$$a=1/5$$

$$b=-1/2$$

$$c=1/3$$

$$d=0$$

$$e=0$$

$$f(n) = 1/5n^5 - 1/2n^4 + 1/3n^3$$

$$\text{Sum} = f(n+1) - f(1)$$

$$= 1/5(n+1)^5 - 1/2(n+1)^4 + 1/3(n+1)^3 - 1/30$$

$$= 1/5n^5 + 1/2n^4 + 1/3n^3$$

From above, we can know that the difference method is a convenient way to solve sums of different series. These are my thoughts of this topic.

Fourier Transform: Their Meaning and Their Use in the Solution of Differential Equations

YURU CHEN

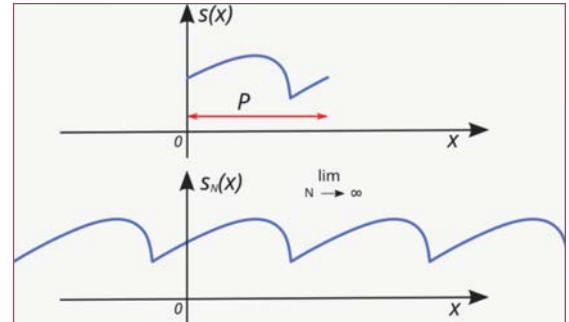
1. From FS to FT

Any periodic function can be split into sums of trigonometric functions

$$s_N(x) = \sum_{n=-N}^N C_n e^{2\pi i \frac{n}{P} x} \quad N \rightarrow +\infty, P \rightarrow +\infty$$

$$e^{i\theta} = \cos \theta + i \sin \theta$$

$$s_0(t) = \int_{-\infty}^{+\infty} s(x) e^{-2\pi i t x} dt$$



2. Basic Concepts

Fourier Transform	$\hat{f}(\xi) = \int_{-\infty}^{+\infty} f(x) e^{-2\pi i \xi x} dx$
Inverse Fourier Transform	$f(x) = \int_{-\infty}^{+\infty} \hat{f}(\xi) e^{2\pi i \xi x} d\xi$
Relationship	$f(x) \stackrel{\mathcal{F}}{\underset{\mathcal{F}^{-1}}{=}} \hat{f}(\xi)$

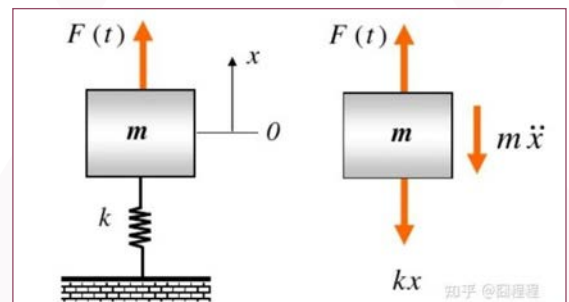
3. Differentiation Properties

$$\begin{cases} \hat{f}'(\xi) = \mathcal{F}\left\{\frac{d}{dx} f(x)\right\} = 2\pi i \xi \hat{f}(\xi) \\ \widehat{f^{(n)}}(\xi) = \mathcal{F}\left\{\frac{d^n}{dx^n} f(x)\right\} = (2\pi i \xi)^n \hat{f}(\xi) \\ \mathcal{F}\left\{\frac{d^n}{d\xi^n} \hat{f}(\xi)\right\} = (2\pi i x)^n f(x) \\ \mathcal{F}\{x^n f(x)\} = \left(\frac{i}{2\pi}\right)^n \frac{d^n}{d\xi^n} \hat{f}(\xi) \end{cases}$$

4. Example: Spring System

An object with mass m is held by a spring with coefficient k and a force F is applied, determine its motion pattern if added a small disturbance.

function: $m\ddot{x}(t) + kx(t) = F(t)$



5. Usual

$$m\lambda^2 + k = 0$$

$$\lambda = \pm \sqrt{\frac{k}{m}} i$$

$$x(t) = e^{0t} (A \cos \sqrt{\frac{k}{m}} t + B \sin \sqrt{\frac{k}{m}} t)$$

partial : $P(t)$

$$x(t) = A \cos \sqrt{\frac{k}{m}} t + B \sin \sqrt{\frac{k}{m}} t + P(t)$$

6. Fourier

$$m(2\pi i\xi)^2 X(\xi) + kX(\xi) = \mathcal{F}\{F(t)\}$$

$$X(\xi) = \frac{\mathcal{F}\{F(t)\}}{k - 4\pi^2 m \xi^2}$$

$$x(t) = \mathcal{F}^{-1}\{X(\xi)\}$$

assume $F(t) = \delta \implies \mathcal{F}\{F(t)\} = 1$

$$x(t) = \mathcal{F}^{-1}\{X(\xi)\} = \mathcal{F}^{-1}\left\{\frac{1}{k - 4\pi^2 m \xi^2}\right\}$$

$$= \int_{-\infty}^{+\infty} \frac{1}{k - 4\pi^2 m \xi^2} e^{2\pi i \xi t} d\xi$$

$$= \begin{cases} \frac{1}{\sqrt{rm}} \sin\left(\sqrt{\frac{k}{m}} t\right), & t > 0 \\ 0, & t < 0 \end{cases}$$

	Function	Fourier transform unitary, ordinary frequency	Fourier transform unitary, angular frequency	Fourier transform non-unitary, angular frequency	Remarks
	$f(x)$	$\hat{f}(\xi) \triangleq \hat{f}_1(\xi)$ $= \int_{-\infty}^{\infty} f(x) e^{-i2\pi\xi x} dx$	$\hat{f}(\omega) \triangleq \hat{f}_2(\omega)$ $= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-i\omega x} dx$	$\hat{f}(\omega) \triangleq \hat{f}_3(\omega)$ $= \int_{-\infty}^{\infty} f(x) e^{-i\omega x} dx$	Definitions
101	$a f(x) + b g(x)$	$a \hat{f}(\xi) + b \hat{g}(\xi)$	$a \hat{f}(\omega) + b \hat{g}(\omega)$	$a \hat{f}(\omega) + b \hat{g}(\omega)$	Linearity
102	$f(x - a)$	$e^{-i2\pi\xi a} \hat{f}(\xi)$	$e^{-ia\omega} \hat{f}(\omega)$	$e^{-ia\omega} \hat{f}(\omega)$	Shift in time domain
103	$f(x) e^{iax}$	$\hat{f}\left(\xi - \frac{a}{2\pi}\right)$	$\hat{f}(\omega - a)$	$\hat{f}(\omega - a)$	Shift in frequency domain, dual of 102
104	$f(ax)$	$\frac{1}{ a } \hat{f}\left(\frac{\xi}{a}\right)$	$\frac{1}{ a } \hat{f}\left(\frac{\omega}{a}\right)$	$\frac{1}{ a } \hat{f}\left(\frac{\omega}{a}\right)$	Scaling in the time domain. If $ a $ is large, then $f(ax)$ is concentrated around 0 and $\frac{1}{ a } \hat{f}\left(\frac{\omega}{a}\right)$ spreads out and flattens.
105	$\hat{f}_n(x)$	$\hat{f}_1(x) \xrightarrow{\mathcal{F}_1} f(-\xi)$	$\hat{f}_2(x) \xrightarrow{\mathcal{F}_2} f(-\omega)$	$\hat{f}_3(x) \xrightarrow{\mathcal{F}_3} 2\pi f(-\omega)$	The same transform is applied twice, but x replaces the frequency variable (ξ or ω) after the first transform.

7. Reference

[1] https://en.wikipedia.org/wiki/Fourier_transform

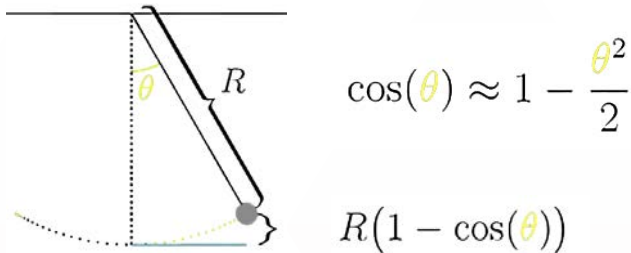
[2] https://zhuanlan.zhihu.com/p/390654546utm_psn=1804548506821738496

Potential energy of a pendulum

PEILING LI

1. Series expansion

An expression for how high the weight of the pendulum is above its lowest point:



Approximate a complicated function as a series expansion to make our analysis easier and faster

2. Complex Functions and Their Series Expansion

Complex function: A function whose range is in the complex number is said to be a complex function, or a complex-valued function.

$$w = f(z)$$

$$f(z) = u(x,y) + iv(x,y)$$

2.1. Complex functions

- Single-valued / Multiple-valued (Many-valued)
- Examples of complex functions

The hyperbolic sine and the hyperbolic cosine of a complex variable are defined as they are with a real variable; that is,

$$\sinh z = \frac{e^z - e^{-z}}{2} \quad \text{and} \quad \cosh z = \frac{e^z + e^{-z}}{2}.$$

The other four hyperbolic functions are defined in terms of the hyperbolic sine and cosine functions with the relations:

$$\begin{aligned} \tanh z &= \frac{\sinh z}{\cosh z} & \coth z &= \frac{\cosh z}{\sinh z} \\ \operatorname{sech} z &= \frac{1}{\cosh z} & \operatorname{csch} z &= \frac{1}{\sinh z}. \end{aligned}$$

2.2. Series expansion — Geometric series

- For rational functions, $\frac{p(z)}{q(z)}$ ($p(z)$ and $q(z)$ are polynomials, $q(z) \neq 0$)

$$\begin{aligned} \frac{1}{1-cw} &= \sum_{n=0}^{\infty} (c \cdot w)^n, \text{ for } |w| < \frac{1}{|c|} \\ \frac{1}{1-\frac{b}{w}} &= \sum_{n=0}^{\infty} \left(\frac{b}{w}\right)^n, \text{ for } |w| > |b| \end{aligned}$$

2.3. Series expansion — Geometric series — Examples

• Ex-1 $f(z) = \frac{1}{3z-2}$ around $z_0 = 0$

$$\begin{aligned} f(z) &= -\frac{1}{2-3z} \\ &= -\frac{1}{2} \cdot \frac{1}{1-\frac{3}{2}z} \\ &= -\frac{1}{2} \left(1 + \frac{3}{2}z + \left(\frac{3}{2}z\right)^2 + \left(\frac{3}{2}z\right)^3 + \dots \right) \\ &= -\frac{1}{2} \sum_{n=0}^{\infty} \left(\frac{3}{2}z\right)^n \end{aligned}$$

• Ex-2 $f(z) = \frac{1}{(1-z)^2}$

$$= \frac{1}{(1-z)} \cdot \frac{1}{(1-z)}$$

• Ex-2 $f(z) = \frac{1}{(1-z)^2}$

$$\begin{aligned} &= \left(\frac{1}{1-z} \right)' \\ &= (1 + z + z^2 + z^3 + \dots)' \\ &= 1 + 2z + 3z^2 + \dots + nz^{(n-1)} + (n+1)z^n \\ &= \sum_{n=0}^{\infty} (n+1)z^n \end{aligned}$$


2.4. Series expansion — Taylor series

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 + \dots$$

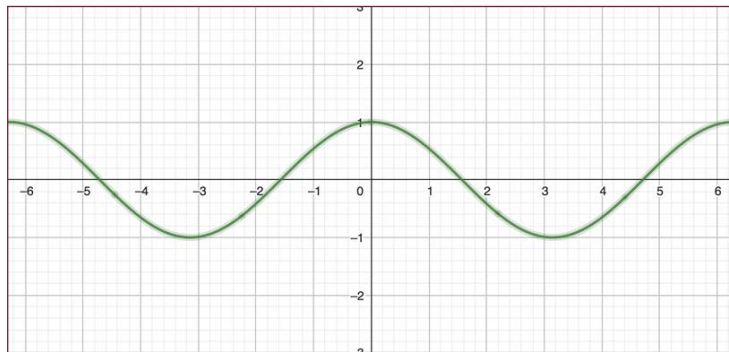


$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x-a)^n.$$

Free to choose

• $\cos(x)$ 

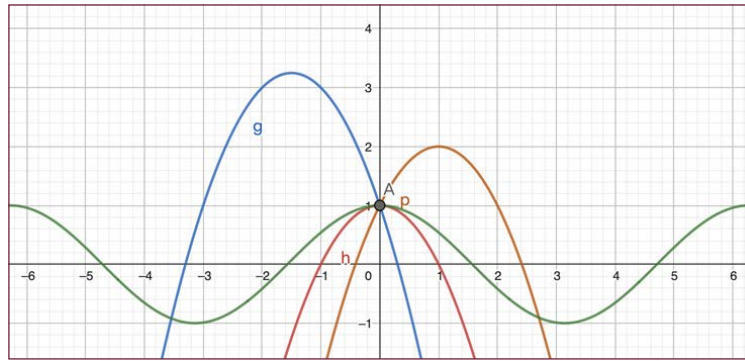
• $P(x) = c_0 + c_1x + c_2x^2 + \dots$



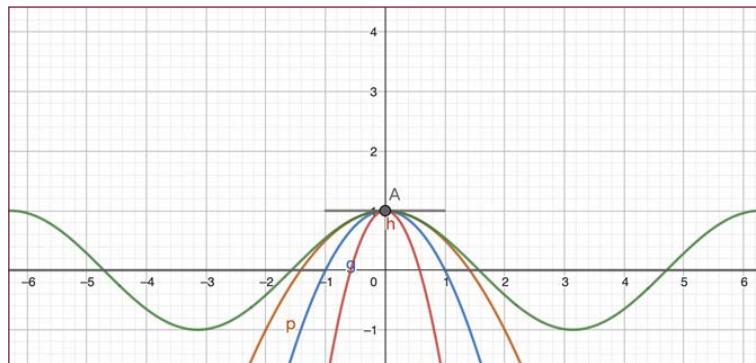
• $x = 0: \cos(x) = 1$

• $P(x) = c_0 + c_1x + c_2x^2$

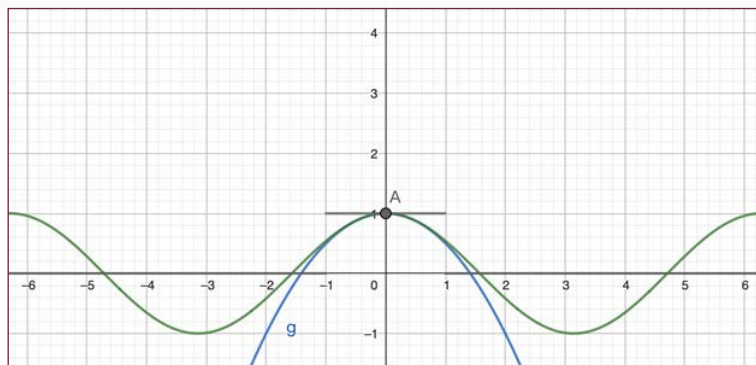
• $P(0) = c_0 + c_1 \cdot 0 + c_2 \cdot 0 = 1 \rightarrow c_0 = 1$



- $\frac{d(\cos)}{dx}(0) = -\sin(0) = 0$
- $\frac{dP}{dx}(x) = c_1 + 2c_2x$
- $\frac{dP}{dx}(0) = c_1 + 2c_2 \cdot 0 = 0 \rightarrow c_1 = 0$



- $\frac{d^2(\cos)}{dx^2}(0) = -\cos(0) = -1$
- $\frac{d^2(P)}{dx^2}(x) = 2c_2$
- $\frac{d^2(P)}{dx^2}(0) = 2c_2 = -1 \rightarrow c_2 = -\frac{1}{2}$



- $P(x) = c_0 + c_1x + c_2x^2 + c_3x^3 + \dots$
- $\frac{dP}{dx}(x) = 0 \cdot c_0 + 1 \cdot c_1 + 2 \cdot c_2x + 3 \cdot c_3x^2 + \dots$
- $\frac{d^2(P)}{dx^2}(x) = 0 \cdot c_0 + 0 \cdot 1 \cdot c_1 + 1 \cdot 2 \cdot c_2 + 2 \cdot 3 \cdot c_3x + \dots$
- $\frac{d^3(P)}{dx^3}(x) = 0 \cdot c_0 + 0 \cdot 1 \cdot c_1 + 1 \cdot 2 \cdot c_2 + 1 \cdot 2 \cdot 3 \cdot c_3 + \dots$

Factorial

- $P(0) \rightarrow c_0$
- $\frac{dP}{dx}(0)/1! \rightarrow c_1$
- $\frac{d^2(P)}{dx^2}(0)/2! \rightarrow c_2$
- $\frac{d^3(P)}{dx^3}(0)/3! \rightarrow c_3$

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n$$

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x - a)^n.$$

If $f(z)$ is analytic throughout a disk $|z - z_0| < R$, then

$$f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(z_0)}{n!} (z - z_0)^n$$

2.5. Series expansion — Taylor series — Maclaurin series

$$f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(z_0)}{n!} (z - z_0)^n$$

A Taylor series with centre $z_0 = 0$, which is referred to as Maclaurin series

$$f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} z^n$$

Some important Maclaurin series that you may want to remember...

$$\begin{aligned} \frac{1}{1-z} &= \sum_{n=0}^{\infty} z^n, & |z| < 1; \\ e^z &= \sum_{n=0}^{\infty} \frac{z^n}{n!} & |z| < \infty; \\ \sin z &= \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{n!} & |z| < \infty; \\ \cos z &= \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{n!} & |z| < \infty; \\ \sinh z &= \sum_{n=0}^{\infty} \frac{z^{2n+1}}{n!} & |z| < \infty; \\ \cosh z &= \sum_{n=0}^{\infty} \frac{z^{2n}}{n!} & |z| < \infty; \end{aligned}$$

2.6. Series expansion — Laurent series

If $f(z)$ is analytic throughout an annulus domain $R_1 < |z - z_0| < R_2$, centered an z_0 , then



$$f(z) = \sum_{n=0}^{\infty} a_n(z - z_0)^n + \sum_{n=1}^{\infty} \frac{b_n}{(z - z_0)^n},$$

analytic part

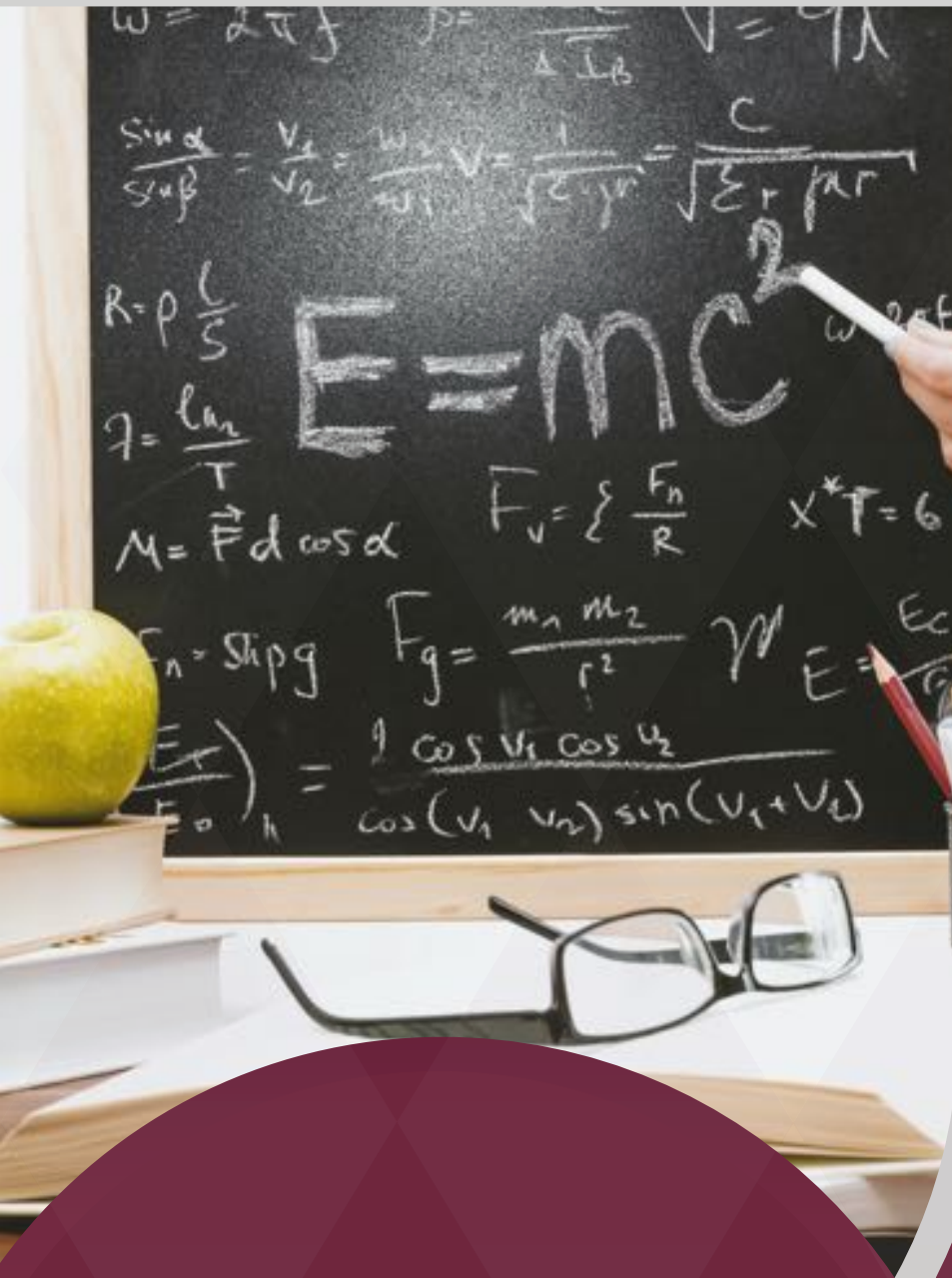
principal part

$$a_n = \frac{1}{2\pi i} \oint_C \frac{f(z)dz}{(z - z_0)^{n+1}}, \quad n = 0, 1, 2, \dots$$

$$b_n = \frac{1}{2\pi i} \oint_C \frac{f(z)dz}{(z - z_0)^{-n+1}}, \quad n = 1, 2, \dots$$

3. Reference

- [1] https://www.bilibili.com/video/BV1Gx411Y7cz/?share_source=copy_web
- [2] [https://math.libretexts.org/Bookshelves/Analysis/Complex_Analysis_-_A_Visual_and_Interactive_Introduction_\(Ponce_Campuzano\)/02%3A_Chapter_2/2.01%3A_Complex_functions](https://math.libretexts.org/Bookshelves/Analysis/Complex_Analysis_-_A_Visual_and_Interactive_Introduction_(Ponce_Campuzano)/02%3A_Chapter_2/2.01%3A_Complex_functions)
- [3] [https://math.libretexts.org/Bookshelves/Analysis/Complex_Analysis_-_A_Visual_and_Interactive_Introduction_\(Ponce_Campuzano\)/05%3A_Chapter_5/5.01%3A_Series](https://math.libretexts.org/Bookshelves/Analysis/Complex_Analysis_-_A_Visual_and_Interactive_Introduction_(Ponce_Campuzano)/05%3A_Chapter_5/5.01%3A_Series)
- [4] <https://youtu.be/RC15R-ktnUI?si=OJKzdxFkrJwitR1N>
- [5] <https://b23.tv/Js2dfCE>
- [6] [https://math.libretexts.org/Bookshelves/Analysis/Complex_Analysis_-_A_Visual_and_Interactive_Introduction_\(Ponce_Campuzano\)/05%3A_Chapter_5/5.02%3A_Taylor_Series](https://math.libretexts.org/Bookshelves/Analysis/Complex_Analysis_-_A_Visual_and_Interactive_Introduction_(Ponce_Campuzano)/05%3A_Chapter_5/5.02%3A_Taylor_Series)
- [7] [https://math.libretexts.org/Bookshelves/Analysis/Complex_Analysis_-_A_Visual_and_Interactive_Introduction_\(Ponce_Campuzano\)/05%3A_Chapter_5/5.03%3A_Laurent_Series](https://math.libretexts.org/Bookshelves/Analysis/Complex_Analysis_-_A_Visual_and_Interactive_Introduction_(Ponce_Campuzano)/05%3A_Chapter_5/5.03%3A_Laurent_Series)
- [8] <https://mathworld.wolfram.com/ComplexFunction.html>



02

Physics

The Understanding and Extension of EPR Paradox

JARUI HE

1. Introduction

In 1935, the well-known paper *Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?* written by Einstein, Podolsky, Rosen was released, [1] and brought out the famous debate about the incompleteness in quantum mechanics. Through the hard work of numerous physicists in decades, this argument is almost getting the answer finally. This essay is generated to share the understanding of EPR paradox and discuss Bell's theory and relevant experiments developed by EPR paradox.

2. EPR Paper

1. Completeness in Physics

For the first step, the distinction between objective reality and physical concepts is given in the paper, with the emphasis on the fact that the former does not depend on specific theories. Then, two criteria for judging the success of physical theories are given: correctness and completeness. [2]

The focus of the rest of the paper is to question the completeness of quantum mechanics. The reality condition is given, which means that if the value of a physical quantity can be predicted with certainty without disturbing the system, then there must be a physical reality element corresponding to this physical quantity, that is to say, this physical quantity has reality. However, it should be noted that a definite value of a physical quantity is a sufficient and unnecessary condition for the physical quantity to have reality, that is, a value of a physical quantity without certainty can also have reality, and certainty and reality are different. Reality can also include probability, for example, when a coin is tossed, the two sides of the coin are probabilistic, but the physical quantity of the two sides of the coin is real, because the coin toss is not a quantum mechanical phenomenon. The reason for the probability of heads and tails on a coin toss is that not all the information of the coin toss is available, and if all the information is available, you must know both sides of the coin.

Therefore, reality refers to whether a physical quantity can be predicted with certainty in nature, not whether it can be predicted in a theory or in an experiment, because the failure to predict in these cases may also be due to incomplete theory or insufficient experimental information. [3] Next, the condition of completeness is given in the paper, which means that if a theory is complete, it must have physical quantities corresponding to the elements of physical reality, that is, these physical quantities must have reality, that is, they can be predicted essentially. If it is a complete theory, it must be able to give accurate predicted values of physical quantities with reality. [4]

3. EPR experiment

In order to clarify, The EPR experiment mentioned here is the original one based on the description of EPR paper instead of the EPR experiment about spin particles designed by Bohm [5]. Einstein supposed there were 2 systems interacting each other from $t=0$ to $t=T$, after that they no longer interact in any way. He further assumes that the states of the two systems are known before time $t=0$. Thus, the state of the composite system at the subsequent time can be calculated by means of Schrodinger equation. [6]

As the result of this process, non-commutative operators cannot have definite values at the same time. The explanation of this phenomenon is given: (1) quantum mechanics is incomplete in describing physical reality by means of wave function; Or (2) if the operator corresponding to two mechanical quantities is not commutative, the two mechanical quantities do not have simultaneous reality. (1) The incompleteness of quantum mechanics means that the physical quantity corresponding

to the commutation operator can essentially have a definite value at the same time, that is, these physical quantities have simultaneous reality, but because of the incompleteness of the description of quantum mechanics, it cannot give a definite value, similar to the use of incomplete theory to describe the flip of a coin, can not predict the front and back of the coin. (2) When the operators corresponding to two mechanical quantities are not commutative, the two mechanical quantities do not have simultaneous reality.

This interpretation means that the two physical quantities cannot by nature be determined simultaneously; they do not have simultaneous reality. In addition, Einstein used proof by contradiction to rule out other possibilities if neither (1) nor (2) were the cause of the phenomenon. If other causes if is possible , then these two quantities have simultaneous reality, that is to say, they are essentially predictable, then if quantum mechanics is complete, it must be possible to give a definite value for these two quantities, and this is inconsistent with what is observed in quantum mechanics, thus indicating that (1) or (2) is the cause of this phenomenon.[7]

Assuming that quantum mechanics is complete, that is, the cause of the above phenomena is not (1), but should be (2). Einstein then demonstrated that quantum mechanics cannot be explained by assuming that it is complete (2). The proof experiment is the following: First, the whole wave packet of the two particles is constructed, and then the particles are separated to prevent any interaction between them. Through quantum mechanics, it is concluded that as long as the momentum and coordinates of the first particle are measured, the momentum and coordinate values of the second particle can be accurately predicted. Moreover, since the two particles do not interact, the measurement of the first particle will not affect the second particle.

Therefore, we know the momentum and coordinate values of the second particle without disturbing the second particle, and according to the reality condition, it shows that the momentum and coordinates have reality, and they are the same physical reality (the second particle), which also shows that they have simultaneous reality, thus contradicting (2), and then draw the conclusion that the assumption that quantum mechanics is complete is wrong. So quantum mechanics is incomplete and some hidden variables could be existed. [8]

4. Conclusion for The Experiments

From Einstein's words on last part of the paper. He himself gave the starting point for overturning the above incomplete proof of quantum mechanics: the coordinates and momentum of the second particle are not determined at the same time. Einstein took into account the fact that the two particles do not interact, so they still have simultaneous reality even if the coordinate momentum is not determined at the same time, because he believed that the reality of the physical quantity of the second particle does not depend on the measurement of the first particle, that is, the state of the second particle is not affected by the first particle with which it does not interact.[9] Which could be interpreted as the two particles states were determined at the beginning.

5. Evaluation of EPR

The ideas given in the paper from the experiment are innovative but contain some flaws. One of the flaws is that the states conceived in the experiment are non-physical, which indicates the states are not able to achieve through experiment . According to the words of a physicist who is highly respected in China, Hongyi Fan ,

Einstein did not find the corresponding state vector from its wave function, neither the normalization coefficient of this state. It is not enough to write the wave function, because the wave function is just some representation of the corresponding quantum state. Fan used the integral theory within ordered operator (which is the theory Fan invented himself [10]) to construct a complete orthogonal entangled state representation [11] , which can be correctly derived and normalized to a singular Delta function, which shows that the state in EPR is non-physical, that is, the state can not

be realized experimentally. Thus, Einstein's "spooky action at a distance" does not exist when it comes to measuring the coordinates and momentum of two particles.[12]

6. Bohm's experiments for EPR

Because of the flaws like the one mentioned above in the original experiment, Bohm came up with his interpretation and improved experiment (EPRB experiment) in 1951[13]. In keeping with the essence of the EPR thought experiment, he replaced the continuous variables (momentum and position) with discrete variables (spin), and discussed the case of atoms with total spin 0. If you separate two atoms while keeping the total spin of the molecule constant, if you measure atom A in the x direction, you will get the spin state \uparrow , and if you measure atom B in the x direction, you will get the result \downarrow . Such measurements are always 100% correlated. However, due to the uncertainty principle, if you choose to measure B in other directions, such as y or z, then its spin state measurement results are unpredictable and probabilistic (here x, y, and z can be any three directions). Using discrete variable to replace continuous variable is not only more intuitive, but also easier to do mathematical processing, and EPRB experiment is also easier to realize.[14]

In 1952, Bohm published two consecutive articles in the Physical Review in which he proposed a hidden variable interpretation of quantum mechanics[15][16]. He believed that in the quantum world particles still move along a precise continuous trajectory, but this trajectory is not only determined by the usual forces, but also by a more subtle quantum potential. The quantum potential is generated by the wave function, which guides the motion of particles by providing dynamic information about the whole environment. It is its existence that leads to the strange motion behavior of micro-particles that is different from that of macroscopic objects. The most striking thing about Bohm's theory is its treatment of measurement. In this theory, the properties of a quantum system do not belong to the system itself; its evolution depends on both the system and the measuring instrument. Therefore, the statistical distribution of the measurement results regarding the hidden variables will vary with the experimental setup. It is this holistic feature that guarantees that Bohm's hidden variable theory has exactly the same predictions as quantum mechanics.

However, it has also led to a deeply uncomfortable result. [17] According to Bohm's theory, although it recovers the trajectory of the particle, it is a trajectory that can never be seen, and the hidden variables introduced in the theory – the determined position and velocity of the particle are in principle immeasurable. One can never know the actual trajectories of particles, and measurements of them will always yield results consistent with quantum mechanics.

Additionally, the other physical reality wave function assumed by Bohr's theory is also an undetectable hidden variable, because physical measurements of individual particles generally yield only a definitive result about the particle's properties.

7. Bell's Inequality

The famous physicist Bell was inspired by Bohm's discussion of EPRB experiments. To provide practical experimental evidence for hidden variables, he considered separating two entangled electrons far enough apart and measuring the spins of the two particles, A and B, separately. Bell theorized that it was impossible to build a detector that could measure the spin of a particle in multiple directions at the same time, so the particles in both places could only be measured in one direction, but the direction of measurement was no longer fixed in the same direction, but the direction of each person was chosen independently and randomly. Under the dual assumption of reality and locality, Bell established Bell's inequality by analyzing the correlation between hidden variables and particles in two cases of quantum mechanics: $|P(x,y) - P(x,z)| \leq 1 + P(y,z)$, [18] where $P(x,y)$ represents the average value of particle A along the x direction and particle B along the z direction. The same is true for $P(x,z)$ and $P(y,z)$. At the same time, he proposed Bell's theorem: no localized hidden variable theory can replicate all the predictions of quantum mechanics.[19] He further explained: "If the hidden variable theory were local, it would not be compatible with quantum mechanics, and if it were consistent with quantum mechanics, it would not be local. That's what the theory says." [20] Bell's inequality is strictly true in the classical world, but if the microscopic world is indeed as described

by quantum mechanics, then the inequality no longer holds. Therefore, by experimentally testing whether Bell's inequality is true, we can know whether Einstein and Bohr are right or wrong.

8. Experiments for Bell's Inequality

There were plenty of physicists made effects for proving Bell's inequality. For instance, the first official experiment for Bell's inequality carried out by John Francis Clauser [21] which was doubted to be inaccurate because the distance between the two particles was not enough to show the speed of coordination in the particles was faster than light speed.

This essay will mainly talk about the experiment done by Alain Aspect. Aspect was the first to design a way to avoid localized vulnerabilities. He sent two entangled photons to each end of a huge room at a distance of 12 meters, so that the contact between the two entangled photons took at least 40 nanoseconds, which is longer than the time to measure and obtain the result. Of course, in addition to ensuring sufficient distance, it is also necessary to have a rapidly changing experimental setup, so as to avoid possible hidden variables.[22] In 1957, Bohm had assumed that the direction of measurement of entangled particles was still changing during flight, which Bell thought was extremely important for the experimental setup. [23] In all early experimental measurement schemes, the measurement direction selected for each experiment was set in advance and remained unchanged. In order to quickly change the experimental Settings, Aspect used acousto-optic modulation technology to change the light path during the experiment, so that it changed periodically according to the frequency of 50 MHz, so as to achieve the purpose of rapidly changing the measurement direction, and realized the polarization measurement of changing direction for the first time. Of course, the measurement process uses consistent measurements.[23] Aspect, through this setting, ensures that there is no information exchange of hidden variables that do not exceed the speed of light between the two measurement points A and B during each measurement process, thus closing the localization vulnerability.

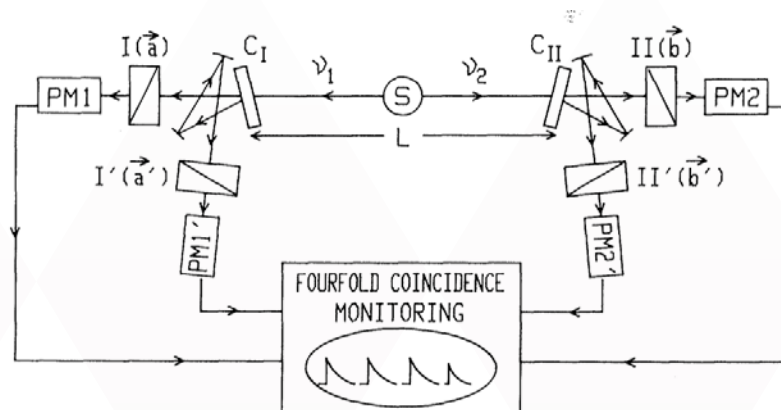


fig.1 Timing experiment with optical switches. Each switching device (C_i , C_{ii}) is followed by two polarizers in two different orientations. Each combination is equivalent to a polarizer switched fast between two or orientations.[25]

Moreover, Aspect uses more efficient entangled photon sources and two-channel detection systems to improve detection efficiency,[26] resulting in more accurate and convincing data. Based on the above important experimental improvements, the experimental results finally prove that the Bell inequality is not valid, and for the first time provide relatively reliable experimental evidence for violating the Bell inequality.

Although he tried to close the vulnerability as much as possible, it now seems that due to the technology at the time, the experimental scheme was not perfect, and only closed the localized vulnerability.

However, through the development of technology, more and more precise experiments had been carried out, and most of the results had disproved Bell's inequality successfully[27], which indicated that the incorrectness of EPR paper.

9. Conclusion

After almost a hundred year , the debate bought out by EPR seems to nearly come to end. It is no doubt that the quantum-theoretic predictions were obeyed at last.[28] However, David Mermin believed most physicists would agree that the announcement of EPR is greater than things like Bell's theory.[29] Just like the phrase Abraham Pais used to describe EPR,“the most profound discovery of science”,[30] the presentation of EPR paper provided an extensive boost to the development of quantum mechanics, including Bell's inequality, using the spirit of questioning.

10. Reference List

- [1] Einstein, A., Podolsky, B., & Rosen, N. (1935). Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Physical Review*, 47(10), 777–780. <https://doi.org/10.1103/physrev.47.777>
- [2] Einstein, A., Podolsky, B., & Rosen, N. (1935). Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Physical Review*, 47(10), 777–780. <https://doi.org/10.1103/physrev.47.777>
- [3] Einstein, A., Podolsky, B., & Rosen, N. (1935). Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Physical Review*, 47(10), 777–780. <https://doi.org/10.1103/physrev.47.777>
- [4] Einstein, A., Podolsky, B., & Rosen, N. (1935). Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Physical Review*, 47(10), 777–780. <https://doi.org/10.1103/physrev.47.777>
- [5] Bohm, D. (1951). *Quantum Theory* Prentice-Hall. Englewood Cliffs, NJ
- [6] Einstein, A., Podolsky, B., & Rosen, N. (1935). Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Physical Review*, 47(10), 777–780. <https://doi.org/10.1103/physrev.47.777>
- [7] Fan, H.Y., (2003). Operator ordering in quantum optics theory and the development of Dirac's symbolic method. *Journal of Optics B: Quantum and Semiclassical Optics*, 5(4), p.R147.
- [8] Einstein, A., Podolsky, B., & Rosen, N. (1935). Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Physical Review*, 47(10), 777–780. <https://doi.org/10.1103/physrev.47.777>
- [9] Einstein, A., Podolsky, B., & Rosen, N. (1935). Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Physical Review*, 47(10), 777–780. <https://doi.org/10.1103/physrev.47.777>
- [10] Einstein, A., Podolsky, B., & Rosen, N. (1935). Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Physical Review*, 47(10), 777–780. <https://doi.org/10.1103/physrev.47.777>
- [11] Fan, H. Y., Lu, H. L., & Fan, Y. (2006). Newton-Leibniz integration for ket-bra operators in quantum mechanics and derivation of entangled state representations. *Annals of Physics*, 321(2), 480–494.
- [12] Team, D. (2021). 百密一疏的 EPR 原始论文 (范洪义作). [online] Sciencenet.cn. Available at: <https://blog.sciencenet.cn/blog-3385349-1308055.html>
- [13] Bohm, D. (1951). *Quantum Theory* Prentice-Hall. Englewood Cliffs, NJ
- [14] Bohm, D. (1951). *Quantum Theory* Prentice-Hall. Englewood Cliffs, NJ
- [15] Bohm, D. (1952). A Suggested Interpretation of the Quantum Theory in Terms of 'Hidden' Variables. I. *Physical Review*,



85(2), pp.166–179. doi:<https://doi.org/10.1103/physrev.85.166>.

[16] Bohm, D. (1952). Reply to a Criticism of a Causal Re-Interpretation of the Quantum Theory. *Physical Review*, 87(2), pp.389–390. doi:<https://doi.org/10.1103/physrev.87.389.2>.

[17] Bohm, D. (1952). A Suggested Interpretation of the Quantum Theory in Terms of ‘Hidden’ Variables. I. *Physical Review*, 85(2), pp.166–179. doi:<https://doi.org/10.1103/physrev.85.166>.

[18] Bell, J.S. (1964). On the Einstein Podolsky Rosen paradox. *Physics Physique физика*, 1(3), pp.195–200. doi:<https://doi.org/10.1103/physicsphysiquefizika.1.195>.

[19] Bell, J.S. (1964). On the Einstein Podolsky Rosen paradox. *Physics Physique физика*, 1(3), pp.195–200. doi:<https://doi.org/10.1103/physicsphysiquefizika.1.195>.

[20] Bell, J.S. (1964). On the Einstein Podolsky Rosen paradox. *Physics Physique физика*, 1(3), pp.195–200. doi:<https://doi.org/10.1103/physicsphysiquefizika.1.195>.

[21] Freedman, S.J. and Clauser, J.F. (1972). Experimental Test of Local Hidden-Variable Theories. *Physical Review Letters*, [online] 28(14), pp.938–941. doi:<https://doi.org/10.1103/physrevlett.28.938>.

[22] Aspect, A., Dalibard, J. and Roger, G. (1982). Experimental Test of Bell’s Inequalities Using Time- Varying Analyzers. *Physical Review Letters*, 49(25), pp.1804–1807. doi:<https://doi.org/10.1103/physrevlett.49.1804>.

[23] Bell, J.S. (1964). On the Einstein Podolsky Rosen paradox. *Physics Physique физика*, 1(3), pp.195–200. doi:<https://doi.org/10.1103/physicsphysiquefizika.1.195>.

[24] Aspect, A., Dalibard, J. and Roger, G. (1982). Experimental Test of Bell’s Inequalities Using Time- Varying Analyzers. *Physical Review Letters*, 49(25), pp.1804–1807. doi:<https://doi.org/10.1103/physrevlett.49.1804>.

[25] Aspect, A., Dalibard, J. and Roger, G. (1982). Experimental Test of Bell’s Inequalities Using Time- Varying Analyzers. *Physical Review Letters*, 49(25), pp.1804–1807. doi:<https://doi.org/10.1103/physrevlett.49.1804>.

[26] Aspect, A., Dalibard, J. and Roger, G. (1982). Experimental Test of Bell’s Inequalities Using Time- Varying Analyzers. *Physical Review Letters*, 49(25), pp.1804–1807. doi:<https://doi.org/10.1103/physrevlett.49.1804>.

[27] Giustina, M., Versteegh, M.A. M., Wengerowsky, S., Handsteiner, J., Hochrainer, A., Phelan, K., Steinlechner, F., Kofler, J., Larsson, J.-Å., Abellán, C., Amaya, W., Pruneri, V., Mitchell, M.W., Beyer, J., Gerrits, T., Lita, A.E., Shalm, L.K., Nam, S.W., Scheidl, T. and Ursin, R. (2015). Significant-Loophole-Free Test of Bell’s Theorem with Entangled Photons. *Physical Review Letters*, 115(25). doi:<https://doi.org/10.1103/physrevlett.115.250401>.

[28] Mermin, N.D. (1985). Is the Moon There When Nobody Looks? Reality and the Quantum Theory. *Physics Today*, 38(4), pp.38–47. doi:<https://doi.org/10.1063/1.880968>.

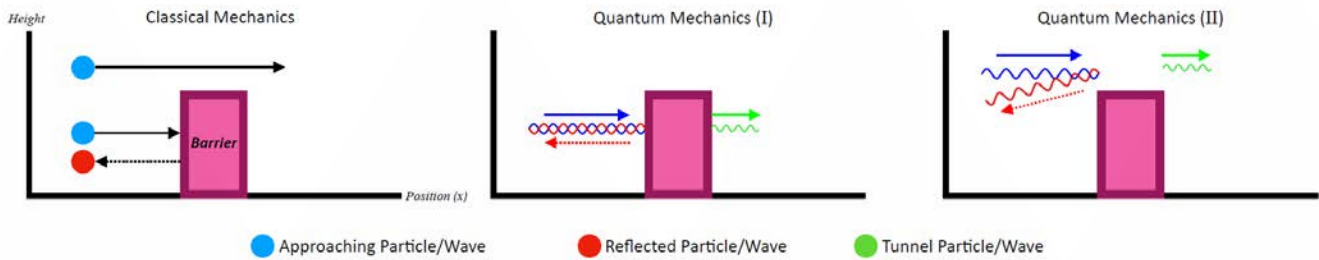
[29] Mermin, N.D. (1985). Is the Moon There When Nobody Looks? Reality and the Quantum Theory. *Physics Today*, 38(4), pp.38–47. doi:<https://doi.org/10.1063/1.880968>.

[30] A. Pais, "Subtle is the Lord..." *The Science and the Life of Albert Einstein*, Oxford U. P., New York (1982) p. 456.

Quantum Tunneling: A Review

ELYAS ALBATTAT

GRAPHICAL ABSTRACT



1. ABSTRACT

According to classical mechanics, if a particle has less energy than the height V_0 the region inside the barrier is forbidden, thus, the particle is reflected based on its trajectory. In this paper, however, we discuss the non-zero probability that a particle could end on the other side of a barrier while the potential energy of the particle is less than the potential energy of the barrier, utilizing the quantum wavefunction associated with that free particle. Moreover, the non-intuitive reflection of waves that possess higher potential energy than the barrier is also discussed. This paper reviews and analyzes both mathematically and physically this effect and proposes applications and roles of this phenomenon in physics and technology.

2. INTRODUCTION

2.1 Newtonian Mechanics

In classical mechanics, a wall will prevent a ball from continuing its path if the wall is higher than the trajectory of the ball, bouncing back and abiding by the physical rules of the respective wall, in which the region on the other side of the wall is classically forbidden. As the particle does not have enough energy, it is not able to surpass the barrier, hence, reflecting and there is zero-probability that the particle transmissions through the barrier.

2.2 Quantum Mechanics

However, in the quantum scale, where the uncertainty principle states that the position of a particle is not accurately and precisely known but has a possibility of existing in one position as opposed to the other. If the first scenario of the wall and the ball is applied to the quantum scale where the sizes are considerably smaller, and we shoot a particle onto a barrier that has higher energy than that particle, there will be a probability, no matter how small it is that the particle is going to end up on the other side of the barrier. The continuous wavefunction of a particle approaching the barrier will display an exponential decay inside the barrier. If the wavefunction remained continuous at the other side of the barrier leaving a finite probability of the particle to tunnel through the barrier. Moreover, if the wavefunction has more energy than the barrier, there is a non-intuitive possibility that the particle will be reflected even though it can go over the barrier.

$$V(x) = \begin{cases} 0, & x < 0 \\ V_0, & 0 \leq x \leq L \\ 0, & x > L \end{cases} \quad (\text{Eq. 1})$$

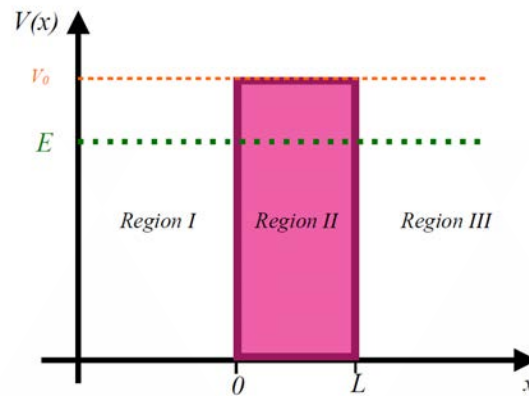


Figure 1: The x-axis is separated into three regions with respect to the potential barrier of height V_0 establishing three different wave-functions (eq. 1). The vertical represent energy, if $E > V_0$ the particle will not encounter the barrier, while if $E < V_0$ the particle will encounter the barrier. In region I, represents the first wavefunction from (eq. 1), where an approaching wave packet coexist with the re-reflected waves. Region II hosts the waves that have not been reflected in the potential barrier $x = 0$ moving at a constant potential $V(x) = +V_0$ tunneling through to region III in which $x > L$, the particle moves past the potential barrier even if the energy of the approaching wave (E) is smaller than the potential barrier.

This was first proposed theoretically in 1927 by Friedrich Hund, following the work of the newly published Schrödinger equation, a year later, it was analyzed by George Gamow to calculate the approximate half-life of Polonium, in which the alpha decay proceeded by quantum tunneling through the Coulomb barrier, an electrostatic interaction between nuclei, attractive at close proximities, but repulsive at further distances. Trapping alpha particles in, while keeping other atomic nuclei out of reach. The energy of the Coulomb barrier is approximately 26 million eV. In classical physics, this value is an insurmountable threshold. Making this objective of classical mechanics insufficient for applications that require overcoming this barrier, such as in nuclear fusion, atoms need to overcome the repulsive energy (Coulomb barrier noted as the potential wall) to be fused.

2.3 Schrödinger Equation

Depending on the finite values (dimensions) of the barrier and on the energy E , the probability of the approaching wave tunneling to the other end of that barrier can change. The probability can be found with the boundary-value problem for the time-independent Schrödinger equation for an individual particle along the wavefunction. Eq. 2 displays the general formula of the equation:

$$-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} \psi(x) + V(x)\psi(x) = E\psi(x) \quad (\text{Eq. 2})$$

In this equation, we presume that $E < V_0$, so that we can isolate the second partial derivative by multiplying by $\frac{-2m}{\hbar^2}$ and solve for the function $\psi(x)$ for regions I and III; then define $-k_0^2 = \frac{-2mE}{\hbar^2}$ and move it to the other side:

$$\frac{\partial^2}{\partial x^2} \psi(x) + k_0^2 \psi(x) = 0 \quad (\text{Eq. 3})$$

Supposing that it is continuous and for the x values for its first derivative. Additionally, satisfy a solution that gives a probabilistic interpretation such that $|\psi(x)|^2 = \psi^*(x) \cdot \psi(x)$ is the density of the probability. And since we divided the x -axis into three regions with the boundaries defined by the multifunction potential (eq. 1), we conclude $\psi_I(x)$ to be the probability in region I where $x < 0$, also with $\psi_{II}(x)$ by $0 \leq x \leq L$, and with $\psi_{III}(x)$ we get the solution of region III by $x > L$, thus, the resulting Schrödinger equations of the regions respectively would be expressed as:

$$-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} \psi_I(\mathbf{x}) = E \psi_I(\mathbf{x}); -\infty < x < 0 \quad (\text{Eq. 3.1})$$

$$-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} \psi_{II}(\mathbf{x}) + V_0 \psi_{II}(x) = E \psi_{II}(\mathbf{x}); -0 \leq x < L \quad (\text{Eq. 3.2})$$

$$-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} \psi_{III}(\mathbf{x}) = E \psi_{III}(\mathbf{x}); L < x < \infty \quad (\text{Eq. 3.3})$$

By substituting into the original equation and differentiating, we can infer that in re-gions I and III the solutions are in these gen-eral forms:

$$\psi_I(x) = Ae^{+ik_0x} + Be^{-ik_0x} \quad (\text{Eq. 4.1})$$

$$\psi_{III}(x) = Fe^{+ik_0x} + Ge^{-ik_0x} \quad (\text{Eq. 4.2})$$

Where $\frac{\sqrt{2mE}}{\hbar}$ is a wave number and the complex exponent denotes oscillations, so as the positive exponents of coefficients A and F, the wave travels to the right. While negative exponents for coefficients B and G, the wave travels to the left. And if B = 0 the probabilistic density of the approaching wave is $P(x) = \psi \cdot \psi^* = (Ae^{ik_0x})(Ae^{-ik_0x})$ the exponents result to zero, leaving $P(x) = A^2$ which is essentially the mathematical justifi-cation of why the particle only travels to the right.

$$e^{\pm ik_0x} = \cos kx \pm i \sin kx \quad (\text{Eq. 5})$$

So in region I, the approaching wave can be expressed as $|\psi_{\text{app}}(x)|^2 = |A|^2$ in the same way that the reflected wave is $|\psi_{\text{re}}(x)|^2 = |B|^2$ and the amplitude of the ransmitted wave is $|\psi_{\text{tra}}(x)|^2 = |F|^2$. The square of the wave amplitude is directly pro-portional to the wave intensity, according to the theory of waves. If one desires to find how much of the approaching wave (incident wave) tunnels through the barrier, it can be found by computing the square of the ampli-tude of the trans-mitted wave. The tunneling probability is the ratio of the amplitude of the trans-mitted wave ($|F|^2$) to the amplitude of the approaching wave ($|A|^2$).

However, in region II the terms of (Eq. 3.2) can be modeled to

$$\frac{\partial^2}{\partial x^2} \psi_{II}(x) + k_{in}^2 \psi_{II}(x) = 0 \quad (\text{Eq. 6})$$

As the wave function inside the barrier $k_{in}^2 = \frac{2m(V_0 - E)}{\hbar^2}$ considering the parameter k_{in} is a real number/realis a localized function it requires (I) to occupy a narrow domain. And (II) Stops abruptly. Since k^2 is positive because $E < V_0$. Where $f(x) = e^x + e^{-x}$ models the mathematical ex-pression of a barrier; the particle has a small chance of entering the barrier until it reaches a penetration depth. The probability density will attenuate as the particle position distribu-tion reaches the barrier. (Eq. 6) is the second-order differential equation; meaning we can write a solution for the wavefunction as a linear combination of real exponentials because it is the only region the particle cannot travel freely, due to being inside the barrier. Region II result is non-oscillatory, as opposed to the other regions, subtle curtailment of $\psi_{II(in)}(x)$:

$$\psi_{II(in)}(x) = Ce^{-k_{in}x} + De^{+k_{in}x} \quad (\text{Eq. 7})$$

In this case, however, we are not able to split them into left and right since they are real exponentials and are not traveling waves.

Nevertheless, if $E > V_0$ the particle goes over the barrier, it would be expressed in complex exponentials because it is a free, non-decaying wavefunction similar to the other regions. So the resultant wavefunction can be:

$$\psi_{II(over)} = He^{ik_{over}x} + Ke^{-ik_{over}x} \quad (\text{Eq. 8})$$

In which the exponentials are complex, thus can be modeled as constant H goes to the right and constant K goes to the left. However, because there is no hindrance that the particle encounters, we can consider $K = 0$ meaning that the particle will not be reflected.

The condition of the wavefunction implies the following: (I) it must be continuous. And (II) it must be normalizable. Meaning that the integral of the wave function times its conjugate equals one in the expression:

$$\int_{-\infty}^{\infty} \psi^*(x) \cdot \psi(x) dx = 1 \quad (\text{Eq. 9})$$

To examine the boundaries $x = 0$ and $x = L$ to ensure that they follow the conditions of a wavefunction, taking the function of the continuity of the particle's path. $\psi_I = \psi_{II(in)}$ by substituting $x = 0$ we obtain $A + B = C + D$ and since continuous functions have continuous derivatives: $\frac{\partial}{\partial x}\psi_I = \frac{\partial}{\partial x}\psi_{II(in)}$ when we substitute $x = 0$ we obtain

$$ik_0A - ik_0B = k_{in}C - k_{in}D \quad (\text{Eq. 10.1})$$

When examining the boundaries $x = L$ within the consecutive functions $\psi_{II(in)} = \psi_{III}$ we obtain:

$$Ce^{k_{in}a} + De^{-k_{in}a} = Fe^{ik_0a} \quad (\text{Eq. 10.2})$$

And its partial derivative $\frac{\partial}{\partial x}\psi_{II(in)} = \frac{\partial}{\partial x}\psi_{III}$ results:

$$k_{in}Ce^{k_{in}a} - k_{in}De^{-k_{in}a} = ik_0Fe^{ik_0a} \quad (\text{Eq. 10.2.1})$$

This is the solution to Schrödinger's equations, resulting in four equations with five variables.

2.4 Transmissibility & Reflectivity Factors

To obtain the transmissibility factor, we need to compare the state of the particle going right in ψ_I and ψ_{III} , since we defined the right part of the equation (eq. 4.1) and (eq. 4.2) respectively we can define it as

$$\begin{aligned} \psi_I(x) &= Ae^{ik_0x} + Be^{-ik_0x} \equiv \psi_{I(right)} + \psi_{I(left)} \\ \psi_{III}(x) &= Fe^{ik_0x} \equiv \psi_{III(right)} \end{aligned}$$

So the probability density function of finding a particle traveling right in (region I) would be expressed as:

$$P_{I(right)} = \psi_{I(right)}^* \cdot \psi_{I(right)} = A^* \cdot A$$

The probability of finding a particle traveling left in (region I) would be similarly written as $P_{I(left)} = B^* \cdot B$ and following the same analogy, finding a particle traveling right in (region III) is $P_{III(right)} = C^* \cdot C$. So the transmission coefficient is given by:

$$T = \left(1 + \frac{\sinh^2(k_{in}a)}{4\eta(1-\eta)}\right)^{-1} \quad (\text{Eq. 11})$$

Where $\eta = \frac{E}{V_0}$. The reflection coefficient is:

$$R = 1 - T \quad (\text{Eq. 12})$$

So if the energy of the particle is significantly lower than V_0 the expected reflectivity is larger than the transmissibility and vice versa.

3. APPLICATIONS

3.1 SCANNING TUNNELING MICROSCOPE

The physical principle of the scanning tunneling microscope (STM) utilizes the quantum tunneling effect in a system composed of the STM and a metal. When current is applied to the needle-like STM tip, electrons tunnel from the STM tip to the conducting surface of the metal, or vice versa. As we found, the tunneling probability is exponentially dependent on the distance, hence, the morphological surface can be plotted by keeping the current constant and measuring the height of the tip, obtaining atomic resolution.

3.2 TUNNEL DIODES

Essentially, a tunnel diode is a p-n junction device that showcases negative resistance. In other words, the current through it decreases when the voltage is increased. Tunnel diodes make use of the wave nature of electrons that allows them to tunnel through a potential barrier, whereas in Newtonian mechanics this effect is diminished. In a diode, when the semiconductors of the p and n regions are highly doped, the depletion region becomes extremely thin (10nm - 20nm), and with the quantum tunnel effect, there is a finite probability that the electrons would tunnel from the conduction band of the n region to the valence band of the p region, meanwhile, no energy loss is observed, meaning that it would be efficient.

3.3 ALPHA DECAY

One of the seminal contributions was due to Gamow, who in 1928 proposed that radioactive decay represents a manifestation of quantum tunneling. Noticing that isotopes of elements such as thorium, uranium, and bismuth decay radioactively by emitting an alpha particle, or helium nucleus. An interesting discrepancy then arose: while the kinetic energies of the alpha particles emitted by these isotopes vary only within a narrow range, their respective half-lives are very different. To resolve this discrepancy, Gamow introduced a nuclear model of a spherical potential well that holds the alpha particle, very similar to the finite potential barrier. The emission was then explained by quantum mechanical tunneling. Most importantly, the probability of tunneling depends rather sensitively on the energy of the particle; if the energy of alpha particles is higher than normal, then they meet a narrower potential barrier since the amplitude decreases exponentially with distance from the nucleus, which will, in turn, increase the probability of escape, hence decreasing the half-life value. The above quantum mechanical interpretation gives an appropriate understanding of the wide range of half-lives among isotopes despite their relatively uniform energies of emission of alpha particles.

4. CONCLUSION

In summary, the relation and probability of the tunnel phenomenon have been proven, moreover, by using the time independent Schrödinger's equation we proved the probability of a particle tunneling through a barrier.

In addition, the transmissibility and reflectivity factors have been considered and derived to final equations where it showed an inverse relationship.

At the end, we have mentioned applications of where this phenomenon occurs.

5. REFERENCES

- [1] Merzbacher, E. (2002). The early history of quantum tunneling. *Physics Today*, 55(8), 44-49.
- [2] Razavy, M. (2013). *Quantum theory of tunneling*. World Scientific.
- [3] Tomsovic, S. (1998). *Tunneling in complex systems (Vol. 5)*. World scientific.
- [4] Sun, J. P., Haddad, G. I., Mazumder, P., & Schulman, J. N. (1998). Resonant tunneling diodes: Models and properties. *Proceedings of the IEEE*, 86(4), 641-660.
- [5] Ling, S. J., Sanny, J., & Moebs, W. (2016). *University physics (Vol. 3)*. OpenStax College, Rice University.



03

Microbiology

The Pathogenicity and Clinical Relevance of *Salmonella typhi*

MENGHAN XU

1. Introduction

Salmonella typhi (*S. typhi*) is an intracellular pathogen (Zhang et al., 2008), thus survive and reproduce in the cytoplasm of the host cell, thereby avoiding clearance by the host's immune system (Jiang et al., 2021). *S. typhi* causes typhoid fever, which is a life-threatening systematic disease and often accompanied by lots of complications, including electrolyte imbalance as well as hematological and neurological complications (Saba Shahid et al., 2021). Humans are the only known natural host of *S. typhi*, and transmission occur through contaminated water and food. Thus, many developing regions around the world are still struggling with the public health problems of *S. typhi* transmission (Typhoid, 2023). Statistically, 9.2 million cases of typhoid fever are estimated to occur worldwide each year (CDC, 2024). This essay aims to give a comprehensive summary of *S. typhi* by discussing its structure, pathogenicity, treatment and existing vaccines. The essay will also attempt to give a prediction of the future development in dealing with *S. typhi* based on the latest studies and technologies.

2. The structure and pathogenicity of *Salmonella typhi*

S. typhi is a Gram-negative, obligate anaerobe that belongs to the serogroup D within subspecies I of the genus *Salmonella*. phylogenetic analysis indicates that this bacterium is highly monomorphic and started infecting human population relatively recently (Galán, 2016).

The antigens of *S. typhi* are O, H, and Vi (Figure 1), and their function is to increase its chance for survival and infection in the host. The “O” antigen inhibits phagocytosis and displays structural variations to “escape” from being targeted by B cells in the human’s immune system (Kintz et al., 2017). The H antigen is a flagellar antigen that also inhibits phagocytosis. Flagella are elongated structures on the surface of bacteria that help *S. typhi* to move and get through mucosal barriers, such as the intestinal epithelial cell layers, therefore entering the host cells. The movement of the flagella can also trigger signaling pathways in the host cell, therefore affecting the behavior of the host cell and even lead to cell damage in some cases (L. Li et al., n.d.) . The Vi antigen is a polysaccharide antigen within the capsule, which is a unique structure acting as a protection film to protect the bacteria against the host immune system and allow survival in extreme environments (Szu, 2013).

There are three types of toxins in *S. typhi*, including (i) cytotoxin, which inhibits protein synthesis in the host, (ii) endotoxin, which is associated with fever, and (iii) enterotoxin, which is the typhoid toxin associated with diarrhea (Figure 1) (Al-Khafaji et al, 2020). The typhoid toxin is an A2B5 type protein consisting of two catalytic subunits PltA and CdtB and a pentamer delivery platform composed of PltB proteins (Figure 2) (Song et al, 2013), thereby highlighting the unique combination that contributes to the functional diversity of typhoid toxin (Liu et al., 2022).

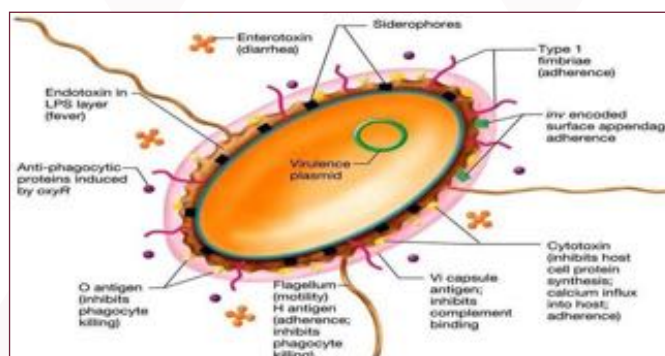


Figure 1. The structure of *Salmonella typhi* showing the three antigens (O, H & Vi) and three toxins (endotoxin, cytotoxin & enterotoxin) (Al-Khafaji et al, 2020)

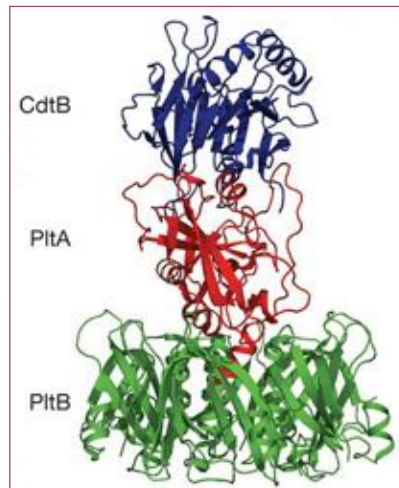


Figure 2. The structure of the typhoid toxin in *Salmonella typhi* showing the two catalytic subunits PltA and CdtB and the delivery platform composed of PltB pentamer (Song et al, 2013)

S. typhi has a circular chromosome carrying prophage genes as well as genes for distinct pathogenicity islands (SPIs) that encode effector proteins to promote bacterial virulence factors and facilitate infection (Figure 3) (Boyd et al, 2012; Zhang et al., 2008). *S. typhi* relies on two SPIs, namely SPI-1 and SPI-2, to encode Type III secretion systems (T3SS) for invasion and intracellular replication. SPI-1 T3SS is primarily responsible for promoting bacterial invasion and entry into host cells, while SPI-2 T3SS is involved in survival and replication within host cells (L. Li et al., 2022.).

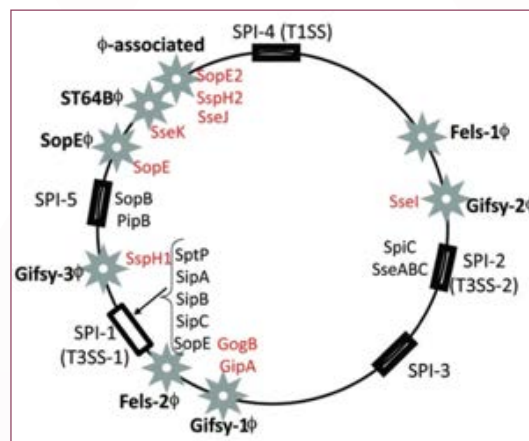


Figure 3. the chromosome of *Salmonella typhi* showing the location of prophage (stars) and pathogenicity islands (black boxes) (Boyd et al, 2012)

3. Treatment and prevention of *Salmonella Typhi* infections

The most common treatment of *S. typhi* infections are antibiotics, including fluoroquinolones, cephalosporins and macrolides. Fluoroquinolones inhibit bacterial binary fission, while cephalosporins and macrolides can inhibit cell wall and protein synthesis, respectively. However, *S. typhi* is increasingly developing antibiotic resistance leading to multi-drug resistance (MDR) against these antibiotics, thereby leading to increased treatment challenges (Typhoid Fever, 2023).

In addition to antibiotics, vaccines can also help in preventing the spread of *S. typhi* greatly by triggering the host immune system to produce antibodies against *S. typhi* antigens. There are three types of vaccines available to prevent *S. typhi* infections, including (i) Ty21a, an oral vaccine that contains live attenuated *S. typhi* modified so that it retains the ability to replicate and stimulate an immune response in human but no longer or has less possibility able to cause typhoid fever, (ii) Vi-PS, a capsular polysaccharide vaccine that contains the polysaccharide antigen of *S. typhi*, and (iii) TCV, a conjugated vaccine that contains the Vi polysaccharide antigen of *S. typhi* along with the tetanus toxoid antigen (Ndezure et al., 2023);

Milligan R et al., 2018; “Typhoid Vaccines: WHO Position Paper,” 2018; Szu, 2013; Sharon M. Tennant & Myron M. Levine, 2015).

In terms of the cross-protection ability of these vaccines. The TCV vaccine is primarily used to target *S. typhi* and does not provide cross protection against other *Salmonella* serotypes (P. Li et al., 2018), while the Ty21a vaccine has shown some degree of cross-protection against *S. paratyphi B* (Jayaum S Booth et al., 2020). Thus, the low protection rate and the lack of cross protection are main challenges faced by current *S. typhi* vaccines.

Although the accessibility and affordability of *S. typhi* vaccines have improved, vaccination availability and uptake are still a problem in many developing regions around the world (“Typhoid Vaccines: WHO Position Paper,” 2018). This is particularly problematic as people in these developing areas have a higher possibility of getting *S. typhi* infection due to poor hygiene practices. On another note, bacterial vaccine market is facing issues related to the lack of profit and investment (Ulmer et al., 2006), which makes the focus mainly on developing new oral vaccines as they tend to be more profitable, easier to administer, and capable of activating mucosal immunity, which is particularly important for gastrointestinal diseases such as typhoid (Hans Van der Weken et al., 2020). However, injectable vaccines are more quickly taken up in the circulation (Omidian & Chowdhury, 2023), thereby potentially capable of providing higher protection.

3. Conclusion

Salmonella typhi is a Gram-negative bacterium that can cause typhoid fever with serious systemic effects, which is an important health issue especially in developing countries. Humans are the only host for *S. typhi*, and the release of *S. typhi* toxins is the main cause of symptoms. Despite the presence of vaccines and antibiotic treatments, challenges such as antibiotic resistance and limited vaccine effectiveness and cross-protection still exist. The development of new oral vaccines and strategies to overcome antibiotic resistance problems are crucial for global health issues regarding *S. typhi* infections. Thus, future work should focus on vaccine and antibiotic development.

4. References

- [1] Al-Khafaji, N. S. K., Al-Bayati, A. M. K., & Al-Dahmashi, H. O. M. (2021). Virulence Factors of *Salmonella Typhi*. In A. Lamas, P. Regal, & C. M. Franco (Eds.), *Salmonella spp.*
- [2] Boyd E F, Carpenter MR & Chowdhury N. (2012). Mobile effector proteins on phage genomes. *Bacteriophage*. 2(3):139-148. doi: 10.4161/bact.21658.
- [3] CDC (2024, April 25). About Typhoid Fever and Paratyphoid Fever. Centers for Disease Control and Prevention.
- [4] Galán, J. E. (2016). Typhoid toxin provides a window into typhoid fever and the biology of *salmonella typhi*. *Proceedings of the National Academy of Sciences of the United States of America*, 113(23), 6338–6344. <https://doi.org/10.1073/pnas.1606335113>
- [5] Hans Van der Weken, Eric Cox, & Bert Devriendt. (2020). *Advances in Oral Subunit Vaccine Design*.
- [6] Jayaum S Booth, Eric Goldberg, Robin S Barnes, Bruce D Greenwald, & Marcelo B Sztein. (2020). Oral typhoid vaccine Ty21a elicits antigen-specific resident memory CD4+ T cells in the human terminal ileum lamina propria and epithelial compartments. *J Transl Med*.
- [7] Jiang, L., Wang, P., Song, X., Zhang, H., Ma, S., Wang, J., Li, W., Lv, R., Liu, X., Ma, S., Yan, J., Zhou, H., Huang, D., Cheng, Z., Yang, C., Feng, L., & Wang, L. (2021). *Salmonella Typhimurium* reprograms macrophage metabolism via T3SS effector SopE2 to promote intracellular replication and virulence. *Nature Communications*, 12(1). <https://doi.org/10.1038/s41467-021->

21186-4

- [8] Kintz, E., Heiss, C., Black, I., Donohue, N., Brown, N., Davies, M. R., Azadi, P., Baker, S., Kaye, P. M., & Woude, M. van der. (2017). *Salmonella enterica* serovar Typhi lipopolysaccharide O-antigen modification impact on serum resistance and antibody recognition. *Infection and Immunity*, 85(4). <https://doi.org/10.1128/IAI.01021-16>
- [9] Li, L., Dan, G. U., Xin'an, J., & Zhiming, P. (2022). Role of pathogenicity islands in *Salmonella* during persistent infection. *Microbiology China*. <http://journals.im.ac.cn/wswxtbcn>
- [10] Li, P., Liu, Q., Luo, H., Liang, K., Han, Y., Roland, K. L., Curtiss, R., & Kong, Q. (2018). Bi-valent polysaccharides of VI capsular and O9 O-Antigen in attenuated *Salmonella* Typhimurium induce strong immune responses against these two antigens. *Npj Vaccines*, 3(1). <https://doi.org/10.1038/s41541-017-0041-5>
- [11] Liu, X., Chen, Z., Jiao, X., Jiang, X., Qiu, J., You, F., Long, H., Cao, H., Fowler, C. C., Gao, X., Richard, E., & Brennan, G. (2022). Molecular Insights into the Assembly and Functional Diversification of Typhoid Toxin. <https://journals.asm.org/journal/mbio>
- [12] Milligan R, Paul M, Richardson M, & Neuberger A. (2018). Vaccines for preventing typhoid fever.
- [13] Ndezure, E. et al. (2023). Drugs, Vaccines and Druggable Targets of *Salmonella* Typhi. <https://doi.org/10.20944/preprints202311.0732.v1>.
- [14] Omidian, H., & Chowdhury, S. D. (2023). Advancements and Applications of Injectable Hydrogel Composites in Biomedical Research and Therapy. In *Gels* (Vol. 9, Issue 7). Multidisciplinary Digital Publishing Institute (MDPI). <https://doi.org/10.3390/gels9070533>
- [15] Saba Shahid, Marvi Mahesar, Nida Ghouri, & Saba Noreen. (2021). A review of clinical profile, complications and antibiotic susceptibility pattern of extensively drug-resistant (XDR) *Salmonella* Typhi isolates in children in Karachi. *BMC Infectious Diseases*, 21, 1 900.
- [16] Sharon M. Tennant, & Myron M. Levine. (2015). Live attenuated vaccines for invasive *Salmonella* infections. 33.
- [17] Song J, Gao X, & Galán G E (2013). Structure and function of the *Salmonella* Typhi chimaeric A(2)B(5) typhoid toxin. *Nature*. 499 (7458): 350-354. doi: 10.1038/nature12377.
- [18] Szu, S. C. (2013). Development of Vi conjugate - A new generation of typhoid vaccine. In *Expert Review of Vaccines* (Vol. 12, Issue 11, pp. 1273-1286). <https://doi.org/10.1586/14760584.2013.845529>
- [19] Typhoid. (2023). World Health Organization.
- [20] Typhoid fever. (2023). Mayo Clinic.
- [21] Typhoid vaccines: WHO position paper. (2018). World Health Organization, 3.
- [22] Ulmer, J. B., Valley, U., & Rappuoli, R. (2006). Vaccine manufacturing: Challenges and solutions. In *Nature Biotechnology* (Vol. 24, Issue 11, pp. 1377-1383). <https://doi.org/10.1038/nbt1261>
- [23] Zhang, X.-L., Jeza, V. T., & Pan, Q. (2008). *Salmonella* Typhi: from a Human Pathogen to a Vaccine Vector (Vol. 5).

The Pathogenicity and Clinical Relevance of *Listeria monocytogenes*

YICHEN JEN

1. Introduction

In Halifax, Nova Scotia, an outbreak of listeriosis was linked to the consumption of cabbage that had been contaminated with sheep manure containing *Listeria monocytogenes* (*L. monocytogenes*) (Schlech WF, 1983). Since then, the significance of *L. monocytogenes* in food industry is gradually being recognized. This essay begins by discussing how the specific properties of *L. monocytogenes* enables it to survive in variable environmental conditions such as low temperatures. We will then describes the core virulence determinants and pathogenesis of *L. monocytogenes*. Subsequently, we will explores the antimicrobial options and the rise of antibiotic resistance in *L. monocytogenes*.

2. Properties of *L. monocytogenes*

L. monocytogenes is a Gram-positive, rod-shaped facultative intracellular pathogen that can be found in the soil, groundwater, and feces of animals and humans (Stea, Purdue, Jamieson, Yost, & Hansen, 2015). Diseases caused by *L. monocytogenes* include listeriosis, bacteremia, meningitis or meningoencephalitis, pregnancy-associated infections manifesting as miscarriage or neonatal sepsis, and foodborne infections (Koopmans, Brouwer, Vázquez-Boland, & van de Beek, 2023).

3. Environmental Survival of *L. monocytogenes*

L. monocytogenes is a resilient microorganism that can easily adapt to fluctuating environments such as low and high temperature, low and high pH levels, high hydrostatic pressure, ultraviolet light, presence of heavy metals and biocides. The movement of *L. monocytogenes* is possible between 22° C and 28° C, but is restricted above 30° C. The bacterium can grow in temperatures ranging from -0.4° C to 45° C, with 37° C being the optimum (Allerberger, 2003). When a typical bacterial cell is exposed to low temperatures, the concentration of unsaturated fatty acids increases to prevent the loss of cytoplasmic contents (Bucur, 2018). In contrast, *L. monocytogenes* responds in a different way by increasing the accumulation of potent osmolytes glycine betaine and carnitine from the environment through a chill-activated transport system. These osmolytes contribute to the survival and growth of *L. monocytogenes* at lower temperatures (Zeisel, 2003).

4. Core Virulence Determinants of *L. monocytogenes*

In the genome of *L. monocytogenes* strain EGD-e, only 10 of the 2,853 coding sequences are regulated by the positive regulatory factor A (PrfA) transcriptional regulator, whose products play crucial roles in listerial intracellular parasitism (Scotti M. , Monzó, Lacharme-Lora, Lewis, & Vázquez-Boland, 2007).

The PrfA-regulated transcriptional units (PrfA regulon) are arranged in distinct chromosomal regions (Fig. 1), of which four lie in a discrete 10-kb region called the *Listeria* pathogenicity island 1 (LIPI-1), a pathogenicity island essential for *Listeria* virulence (Vázquez-Boland, Domínguez-Bernal, González-Zorn, Kreft, & Goebel, 2001). The LIPI-1 includes (i) *hly* that encodes the pore-forming toxin listeriolysin O (LLO), from the cholesterol-dependent cytolysin family (Schnupf & Portnoy, 2007), (ii) *plcB* and *plcA* encoding two phospholipase C, which cooperates with LLO to promote bacterial release from the phagocytic vacuole (Wei, Zenewicz, & Goldfine, 2005), (iii) *mpl* that encodes a metalloprotease involved in the post-secretional processing of pro-PlcB into an active phospholipase (Scotti M. , Monzó, Lacharme-Lora, Lewis, & Vázquez-Boland, 2007), and (iv) *actA* that encodes the actin-polymerizing surface protein ActA, required for actin-based intracellular motility and cell-to-cell spread (Kocks, et al., 1992).

The remaining three PfrA-regulated transcriptional units are located at different points on the *L. monocytogenes* chromosome, and include (i) *inIAB* locus encoding internalins A (InIA) and B (InIB) to regulate the invasion of non-phagocytic cells (Gaillard, Berche, Frehel, Gouln, & Cossart, 1991), (ii) *inIC* encoding small, secreted internalin homologue that is required for full virulence in mice (Engelbrecht, et al., 1996), and (iii) *hpt* that encodes a hexose phosphate transporter required for rapid bacterial growth in the host cell cytosol (Chico-Calero I, 2002). The seven PfrA-regulated transcriptional units are shown in Fig. 1.

5. Pathogenesis of *L. monocytogenes*

There are four steps in the *L. monocytogenes* intracellular infection cycle, including (i) host cell invasion, (ii) escape from the phagocytic vacuole, (iii) rapid intracellular proliferation, and (iv) actin-based motility and cell-to-cell spread (Fig. 2) (Koopmans, Brouwer, Vázquez-Boland, & van de Beek, 2023).

First, host cell invasion happens when *L. monocytogenes* transmits through contaminated food and enters non-phagocytic cells such as enterocytes, fibroblasts, hepatocytes or endothelial cells. Entry is achieved by the cooperation of internalins A and B (InIA and InIB) and the actin-polymerising protein ActA. InIA binds to E-cadherin, a junctional protein expressed by various cell types (Braun & Cossart, 2000), while InIB recognizes the tyrosine kinase receptor Met, the natural ligand for hepatocyte growth factor (HGF) (Shen, Naujokas, Park, & Ireton, 2000). InIA and InIB block endocytic recycling machinery of the gC1qR complement component C1q receptor and host cell surface glycosaminoglycans to stimulate cytoskeletal remodeling and bacterial internalization with the activation of class I phosphoinositide 3-kinase (PI3-K) (Bierne & Cossart, 2002).

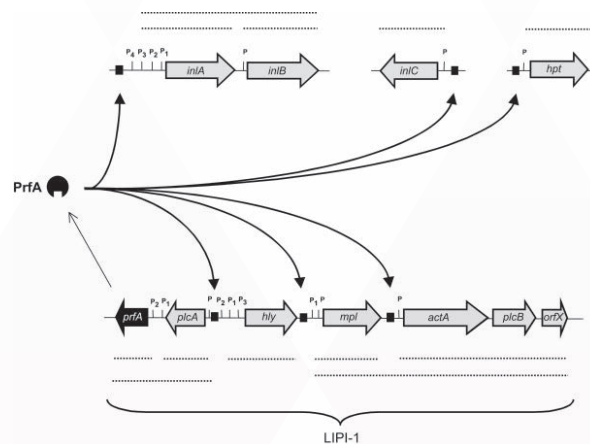


Fig. 1. The core seven PfrA-regulated transcriptional units of *L. monocytogenes*.

The PrfA regulon include, (top) the (i) *inIAB* locus encoding internalins A & B, (ii) *inIC* encoding internalin homologue, (iii) *hpt* encoding a hexose phosphate transporter, and (bottom) the LIPI-1 10-kb region containing (iv) *plcA* and *plcB* encoding two phospholipase C, (v) *hly* encoding the pore-forming toxin listeriolysin O, (vi) *mpl* encoding a metalloprotease, and (vii) *actA* encoding the actin-polymerizing surface protein ActA. (Scotti M. , Monzó, Lacharme-Lora, Lewis, & Vázquez-Boland, 2007).

Second, escaping from the phagocytic vacuole happens after internalization, where other PrfA-regulated virulence factors like LLO, the phospholipases PlcA and PlcB, and the activating metalloprotease Mpl damage phagosomal membrane, allowing the bacteria to escape into the host cell's cytoplasm.

Third, rapid intracellular proliferation occurs once the bacteria reach the cytosol, with the help of Hpt transporter and nutrient-rich environment of host cell (Chico-Calero, 2002).

And finally, actin-based motility and cell-to-cell spread occur during *L. monocytogenes* multiplying, as the bacteria

populate the host and secret virulence factors like InlC, then utilize ActA to enable movement within the host cell and spread to adjacent cells by forming actin tails (Kocks C. G., 1992).

There four steps in the *L. monocytogenes* intracellular infection cycle are highlighted in Fig. 2.

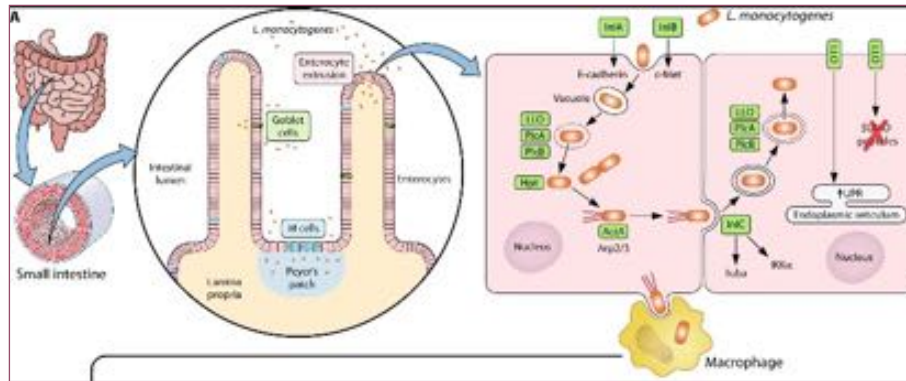


Fig. 2. The four steps in the *L. monocytogenes* intracellular infection cycle.

The steps of *L. monocytogenes* infection include (from left to right): (i) transmission through contaminated food and host cell invasion, mediated by InlA and B, (ii) escaping from the phagocytic vacuole with the help of LLO and PlcA and B, (iii) intracellular proliferation mediated by Hpt, and (iv) actin-based motility and cell-to-cell spread mediated by ActA and InlC. (Koopmans, Brouwer, Vázquez-Boland, & van de Beek, 2023).

6. Treatment of *L. monocytogenes* infection

The β -lactam antibiotics penicillin and aminopenicillins, such as ampicillin or amoxicillin, are the first-choice treatments, even though they tend to be bacteriostatic against intracellular *L. monocytogenes*. The binding of β -lactams to key penicillin-binding-proteins (PBPs) can effectively kill *L. monocytogenes* since β -lactams has been shown to reduce the production of listeriolysin O (LLO) (Nichterlein T, 1996). In addition, the combination of β -lactam antibiotics with aminoglycosides is sometimes used by lacks strong evidence to support its efficacy.

From 1926 to 2007, a study in France found that *L. monocytogenes* showed no resistance to 23 different antibiotics (Morvan A, 2010). However, more recent studies indicate that *L. monocytogenes* has developed resistance to certain antibiotics, including erythromycin, chloramphenicol, rifampin, gentamicin, and cotrimoxazole....

7. Conclusion

According to the information provided by CDC, multiple cases of *Listeria* infection in 12 states in the United States in 2024 have resulted in at least two deaths and the hospitalization of 28 people. Consequently, the impact of human listeriosis on society still remains significant. Moreover, there are still unanswered questions related to why some healthy young patients without risk factors contract neuroinfection (Koopmans MM, 2013), and it is expected that scientists will increasingly focus on the pathophysiological mechanisms of listeriosis in order to develop better treatments and answer these pertinent questions.

8. References

- [1] Allerberger, F. (2003). *Listeria*: growth, phenotypic differentiation and molecular microbiology. *FEMS Immunology & Medical Microbiology*, Volume 35, Issue 3, 183–189.
- [2] Bierne, H., & Cossart, P. (2002). InlB, a surface protein of *Listeria monocytogenes* that behaves as an invasin and a growth

factor. *J Cell Sci* 115, 3357–3367.

- [3] Braun, L., & Cossart, P. (2000). Interactions between *Listeria monocytogenes* and host mammalian cells. *Microbes Infect* 2, 803–811.
- [4] Bucur, F. I.-G. (2018). Resistance of *Listeria monocytogenes* to stress conditions encountered. *Front. Microbiol.* 9, 2700.
- [5] Chico-Calero I, S. M.-Z.-B. (2002). Hpt, a bacterial homolog of the microsomal glucose- 6-phosphate translocase, mediates rapid intracellular proliferation in *Listeria*. *Proc Natl Acad Sci USA* 99, 431–436.
- [6] Chico-Calero, I. S.-Z.-B. (2002). *Listeria monocytogenes* hpt gene, encodes a hexose phosphate transporter required for the intracellular growth, and virulence. *Journal of Bacteriology*, 84–92.
- [7] Engelbrecht, F., Chun, S., Ochs, C., Hess, J., Lottspeich, F., Goebel, W., & Sokolovic, Z. (1996). A new PrfA-regulated gene of *Listeria monocytogenes* encoding a small, secreted protein which belongs to the family of internalins. *Mol Microbiol* 21, 823–837.
- [8] Gaillard, J.-L., Berche, P., Frehel, C., Gouin, E., & Cossart, P. (1991). Entry of *L. monocytogenes* into cells is mediated by internalin, a repeat protein reminiscent of surface antigens from gram-positive cocci. *Cell* 65, 1127–1141.
- [9] Kocks, C. G. (1992). *Listeria monocytogenes*-induced actin assembly requires the actA gene product, a surface protein. *Cell*, 521–531.
- [10] Kocks, C., Gouin, E., Tabouret, M., Berche, P., Ohayon, H., & Cossart, P. (1992). *L. monocytogenes*-induced actin assembly requires the actA gene product, a surface protein. *Cell* 68, 521–531.
- [11] Koopmans MM, B. M. (2013). *Listeria monocytogenes* sequence type 6 and increased rate of unfavorable outcome in meningitis: epidemiologic cohort study. *Clin Infect Dis* 57, 247–253.
- [12] Koopmans, M., Brouwer, M., Vázquez-Boland, J., & van de Beek, D. (2023). Human Listeriosis. *Clin Microbiol Rev* 36, e00060–19.
- [13] Morvan A, M. C.-B. (2010). Antimicrobial resistance of *Listeria monocytogenes* strains isolated from humans in France. *Antimicrob Agents Chemother* 54, 2278–2731.
- [14] Nichterlein T, D. E. (1996). Subinhibitory concentrations of beta-lactams and other cell-wall antibiotics inhibit listeriolysin production by *Listeria monocytogenes*. *Int J Antimicrob Agents* 7, 75–81.
- [15] Schlech WF, L. P. (1983). Epidemic Listeriosis — Evidence for Transmission by Food. *The New England Journal of Medicine*. 308, 203–6.
- [16] Schnupf, P., & Portnoy, D. A. (2007). Listeriolysin O: a phagosome-specific lysin. *Microbes and Infection* 9, 1176–1187.
- [17] Scotti, M., Monzó, H. J., Lacharme-Lora, L., Lewis, D. A., & Vázquez-Boland, J. A. (2007). The PrfA virulence regulon. *Microbes and Infection* 9, 1196–1207.
- [18] Scotti, M., Monzó, H., Lacharme-Lora, L., Lewis, D., & Vázquez-Boland, J. (2007). The PrfA virulence regulon. *Microbes and Infection*, 1196–1207.
- [19] Shen, Y., Naujokas, M., Park, M., & Ireton, K. (2000). InlB-dependent internalization of *Listeria* is mediated by the Met receptor tyrosine kinase. *Cell* 103, 501–510.



[21] Stea, E. C., Purdue, L. M., Jamieson, R. C., Yost, C. K., & Hansen, L. T. (2015). Comparison of the Prevalences and Diversities of *Listeria* Species and *Listeria monocytogenes* in an Urban and a Rural Agricultural Watershed. *Appl Environ Microbiol* 81, 3812–3822.

[22] Vázquez-Boland, J. A., Domínguez-Bernal, G., González-Zorn, B., Kreft, J., & Goebel, W. (2001). Pathogenicity islands and virulence evolution in *Listeria*. *Microbes and Infection*, 3, 571–584.

[23] Wei, Z., Zenewicz, L. A., & Goldfine, H. (2005). *Listeria monocytogenes* phosphatidylinositol-specific phospholipase C has evolved for virulence by greatly reduced activity on GPI anchors. *Proc Natl Acad Sci USA* 102, 12927–12931.

[24] Zeisel, S. H. (2003). Concentrations of choline-containing compounds and betaine in common foods. *J. Nutr.* 133, 1302–1307.



04

Elements of Mathematical Economics

Risk and reward: An exploration of optimal investments

LINGLIN ZHOU

1. Introduction

In finance, finding the best investments is about balancing risk and reward carefully. This essay will explore how this balance works, focusing on the Capital Asset Pricing Model (CAPM) as a guide. First, I will define what risk and reward mean when it comes to investing. Then, I will discuss CAPM and how it helps people make investment choices. I will also look at what people criticize about CAPM and its limits to get a full picture. The main idea here is that while CAPM helps us understand how risk affects returns, we must consider its limits and how it works in real-life situations.

2. MainBody

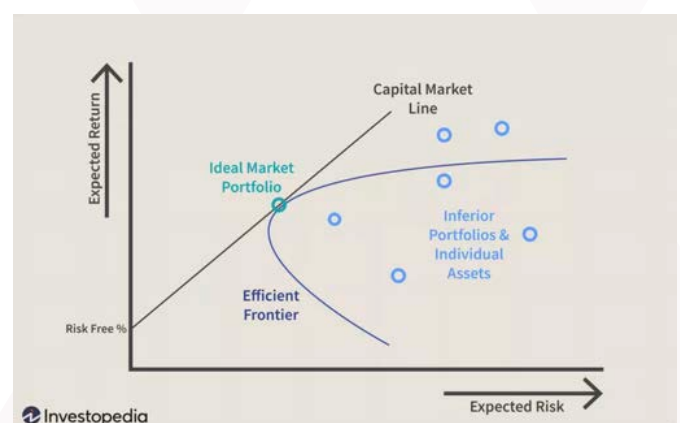
Investing inherently involves making choices under uncertainty. Risk, defined in financial terms as the chance that an outcome or investment's actual gains will differ from an expected outcome or return, is a central consideration for investors (Investopedia, 2024). And reward tempts investors with the promise of potential gains. It's important to weigh the risks against the potential rewards to make the most of investments. Modern portfolio theory, developed by Harry Markowitz, gives us a way to achieve the balance using math.

Markowitz's idea says investors can build portfolios to get the most rewards for a certain amount of risk, or to reduce risk while aiming for specific rewards. The CAPM uses the principles of modern portfolio theory to determine if a security is fairly valued. Using the CAPM to create a portfolio is meant to assist investors in controlling their risk. The formula for calculating the expected reward is " $ER_i = R_f + \beta_i (ER_m - R_f)$ ", where ER_i is the expected return of investment, R_f is the risk-free rate, β_i is the beta of the investment, and $(ER_m - R_f)$ is the market risk premium. If an investor were able to use the CAPM to perfectly optimize a portfolio's return relative to risk, it would exist on a curve called the efficient frontier. The "efficient frontier," an important part of this theory, shows the portfolios that give the best expected return for a set risk level, or the lowest risk for a set return level. This shows how investors must decide: higher rewards typically require taking on more risk.

The efficient frontier is the set of optimal portfolios that offer the highest expected return for a defined level of risk or the lowest risk for a given level of expected return. It is found using math methods like quadratic programming and variance-covariance matrices. Investing spreads out across assets that have returns that are not strongly linked, which helps lower the total risk of the portfolio without giving up potential gains. Diversification is key to managing risk in investment strategies. Portfolios on the efficient frontier aim to maximize returns for a given level of risk. The mix of investments in a portfolio determines its returns.

Risk is measured by the standard deviation of a security. Investors want to load up a portfolio with high-yielding assets that have a cumulative standard deviation that is less than the sum of the standard deviations of the individual securities. The less synchronized the securities (lower covariance), the lower the standard deviation. If this mix of optimizing the return versus risk paradigm is successful, then that portfolio should line up along the efficient frontier line (Ganti,2024).

(Investopedia)



The graph above is a visual representation of the efficient portfolio frontier. It plots the expected risk on the x-axis and return (measured as standard deviation) on the y-axis. The resulting curve shows all possible portfolios that can be constructed with a given set of assets, with each point on the curve representing a different portfolio.

On the horizontal axis, higher expected risk levels are to the right. On the vertical axis, higher returns are towards the top. The efficient frontier is the curve that includes all the portfolios offering the best expected return for each level of risk.

Portfolios above the curve offer higher expected returns than those on the efficient frontier, while those below it offer lower expected returns. Generally, the further a portfolio is from the curve, the greater its risk compared to its expected return (Option Alpha, 2023).

In the real world, CAPM aims to create the best mix of investments for a client who is okay with moderate risk. Here's how it works:

First, the analyst looks at all the different investments that are available. They use past data or predictions to guess how much money each investment might make. Then, they figure out how much each investment's returns move along with the whole market using a method called regression analysis. Next, they use CAPM formula which is $ER_i = R_f + \beta_i (ER_m - R_f)$ to figure out what they expect each investment to earn.

After that, they put all the different investments on a chart. The x-axis represents expected returns and the y-axis represents risk (measured as standard deviation). Based on how much risk the client wants to take and what they want to get from their investment, the analyst picks the mix of investments that fits best. This mix keeps in mind what CAPM says will happen.

Every so often, the analyst looks at how well the investments are doing compared to what CAPM said would happen. If the market changes, they might need to change the mix to keep the right level of risk and money-making chances.

CAPM can be used as a tool to evaluate the reasonableness of future expectations or to conduct comparisons. For investors, there are many benefits that CAPM model can bring. First of all, the model is quite simple and convenient. The calculations are reliable (Nirmal bang). Also, CAPM helps investors in constructing diversified portfolios by quantifying the relationship between individual asset returns and market returns. Diversification reduces unsystematic risk (specific to individual assets) while allowing investors to target a desired level of systematic risk (related to the market as a whole).

However, there are several problems with CAPM when we look at how things work in real life. First, the idea of a risk-free rate doesn't match reality. Regular people can't borrow or lend money at the same low interest rates that governments do. Investments always involve some risk, so there's no such thing as zero risk. This means actual investment returns might be lower than what CAPM predicts (Nirmal bang). Furthermore, the risk-free rate changes a lot. CAPM uses the interest rates on short-term government bonds to figure out this rate, but these rates can change quickly, sometimes in just a few days. Also, to use CAPM, investors need to calculate something called beta. Beta measures how much a particular investment moves compared to the whole stock market. But finding the exact beta value is really hard and takes a lot of time. So, often people use a rough estimate instead, which speeds up calculations but makes them less accurate. Additionally, CAPM assumes there's a "market portfolio" that includes every type of asset, whether they're easy to invest in or not. But in reality, not all assets are available to investors, and some aren't tradable. Also, CAPM assumes everyone can borrow and lend at the risk-free rate, but this isn't true for everyone. Borrowing usually costs more than lending, and not everyone has the same access to financial markets.

So, while CAPM is a good starting point to understand how risk and return are connected, it has many challenges when it comes to putting it into practice in the real world.

For example, think about a tech startup that has a high beta, meaning it's more unpredictable than the overall market.

According to CAPM, this should mean it deserves a higher expected return. However, if the startup operates in a unique market with strong barriers to entry and has special technology, its actual risk level—and thus the return it requires—could be quite different from what CAPM predicts. of the gC1qR complement component C1q receptor and host cell surface glycosaminoglycans to stimulate cytoskeletal remodeling and bacterial internalization with the activation of class I phosphoinositide 3-kinase (PI3-K) (Bierne & Cossart, 2002).

3. Conclusion

In summary, The Capital Asset Pricing Model (CAPM) helps investors understand how risk and return are connected. It helps them figure out if an investment will give enough return for its risk level. By using CAPM, investors can decide how to mix different investments in their portfolios, aiming to balance risk and return for the best possible investment results. While CAPM gives a good starting point for understanding how risk and return relate, applying it in real situations is tricky because there are many factors to consider.

4. References

- [1] Elton, E. J. (2014). Modern portfolio Theory and investment Analysis. *The Journal of Finance*, 37(5), 1317. <https://doi.org/10.2307/2327857>
- [2] Ganti, A. (2024, June 22). Efficient frontier: what it is and how investors use it. Investopedia. <https://www.investopedia.com/terms/e/efficientfrontier.asp>
- [3] Henry, S., Plessis .K.D., Hysmith, R. (2023, May 31). Efficient frontier. Option Alpha. <https://optionalpha.com/learn/efficient-frontier>
- [4] James, C. (2024, May 16). Risk: what it means in investing, how to measure and manage it. Investopedia. <https://www.investopedia.com/terms/r/risk.asp>
- [5] Kenton, W. (2024, July 01). Capital asset pricing model (CAPM): definition, formula, and assumptions. Investopedia. [https://www.investopedia.com/terms/c/capm.asp#:~:text=Key%20Takeaways,to%20the%20market%20\(beta\)](https://www.investopedia.com/terms/c/capm.asp#:~:text=Key%20Takeaways,to%20the%20market%20(beta)).
- [6] Markowitz, H. M. (1952). Portfolio Selection. *Journal of Finance* (pp.77-91).
- [7] Nirmal Bang. (n.d.). Capital asset pricing model. Retrieved August 7, 2024, from <https://www.nirmalbang.com/knowledge-center/capital-asset-pricing-model.html>

Analyzing Investment Risks and Benefits Using Utility Theory and CRRA

SHUOWEN HUAN

1. Introduction

Nowadays, investment plays an important role in people's lives. People save money through investment, and companies earn profits through investment. Investment is like a gamble; it includes two elements: risk and benefit. In "Strategy Execution," Simons notes that competitive risk is a challenge you must constantly monitor and address (as cited in Gibson, 2023). Without risk, companies cannot operate. Risk promotes competition and profit but also involves a significant degree of uncertainty, such as bankruptcy. This paper uses utility theory and CRRA (Constant Relative Risk Aversion) to provide a clear framework for analyzing risk and assessing whether an investment can be beneficial. The paper is structured as follows: Section 2 demonstrates the basic concepts of utility theory (Luque, 2023) and constant

relative risk aversion; We explore the relationship between the coefficient of relative risk aversion with different people and identify the effect of increasing the amount of investment by people with different attitudes to risk on utility in Section 3; Section 4 Calculating the expected utility between Investment Yum! Brands and Investment MCD. But this is not all, we will compare the expected utility of A and B to determine how to choose the appropriate investment option based on risk aversion. Finally, the paper will summarize how CRRA helps in evaluating investment risk.

2. Section snippets

2.1 Concepts of utility theory and constant relative risk aversion

Basis of utility theory

People have different buying intentions when they buy goods. This phenomenon is because the satisfaction and utility brought by goods are different. The utility (satisfaction) people get from a product varies from person to person. For example, buying a new dress or a computer is not the same satisfaction you get. This may depend on one's economic level or other external factors. Utility theory studies the logic behind people obtaining different levels of satisfaction from the consumption of goods and services (Luque, 2023), two types of methods can measure utility, one of them is called the Cardinal approach. Two types of methods can measure the utility, one is called the Cardinal approach and another is called the Ordinal approach, please refer to Figure 1. The Ordinal approach is that consumers can't get satisfaction from buying something, This means that satisfaction is not measurable (Qualitative), it is ranked according to a person's preferences. On the contrary, the Cardinal approach shows that satisfaction can be measured and that assumptions can be expressed numerically. For this article, we mainly use the Cardinal approach.

Figure 1 Breakdown of utility theory approaches and their components

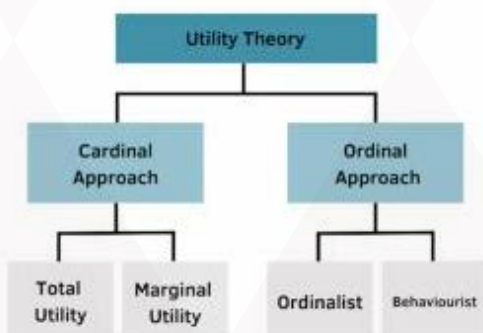


Figure 1 shows two more subdivisions of the cardinal approach, where in the 19th-century economists set total utility and marginal utility as the amount of measurable output. Total utility refers to the complete amount of satisfaction gained. Marginal utility refers to the satisfaction gained from an extra unit consumed (Pettinger, 2024).

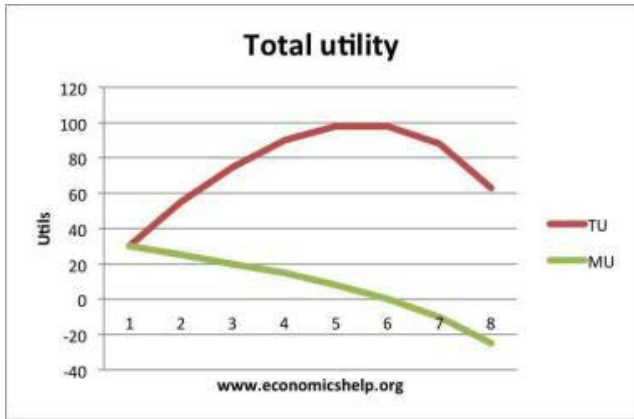


Figure 2 The graph of total utility and marginal utility (Luque, 2023)

The horizontal coordinate in Figure 2 is the number of cakes and the vertical coordinate is the utility unit. The red line represents the total utility and the green line represents the marginal Utility. The Total quality showed an increasing trend when eating one to five cakes, and subsequently decreased. Marginal utility generally shows a declining trend. Marginal satisfaction with eating cake is a diminishing function. As people consume more of a commodity, the satisfaction of each commodity brings decreases. For example, from the figure, at the beginning you are very hungry, the first piece of cake can bring great satisfaction, but as the number of cakes eaten increases, the satisfaction of the cake will decrease.

2.2 Explaining risk aversion

In economics, people often have different perceptions of risk when they invest. Risk aversion refers to people's tendency to prefer certainty over uncertainty, and they will tend to avoid Risk so these people are also known as risk averters. Risk-averse people often demand higher compensation to accept the extra risk, which is called the Risk Premium. Conversely, people who like Risk are said to be risk-seeking. A neutral attitude toward risk is risk-neutral.

2.3 Utility function to represent people 's attitudes towards risk

We can use the utility function to represent these people's preferences and satisfaction. Figure 3 depicts Risk aversion, risk-neutral, risk-loving plots, based on utility. The utility function of the Risk averter is usually concave. The way to judge a concave utility function is to connect two points on the curve, the line connecting the two points, which is inside the curve is the concave utility function, convex utility function, and vice versa. In the Risk-Averse plot, the slope of the curve represents marginal utility, which decreases with increasing wealth. A utility function with a negative second derivative implies risk aversion. As wealth increases, the slope of the tangent lines of the utility function is smaller and smaller. This means "speed" is decreasing and the second derivative is negative (Moreira, 2022). In contrast, Risk-Seeking is a convex utility function, which means that marginal utility increases as wealth increases. The utility functions for risk-neutral are linear, in which case they are based only on the expected return and do not consider the Risk.

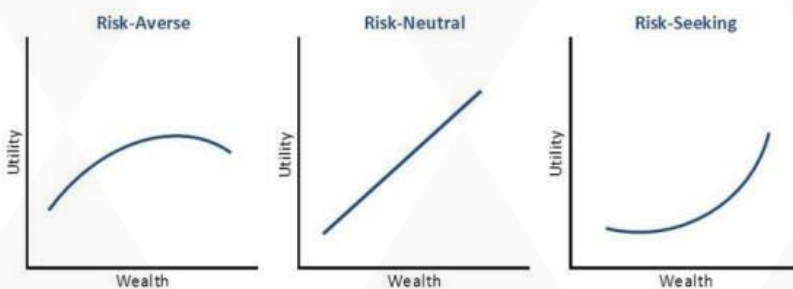


Figure 3 Risk-aversion, risk-neutral, risk-loving plots, based on utility

2.4 A basic introduction to constant relative risk aversion (CRRA)

The CRRA function is also known as the Power Utility Function and CRRA is independent of wealth. Changes in wealth do not affect changes in risk aversion. The CRRA utility function ("CRRA-utility", 2011) is:

$$u(W) = \frac{W^{1-\gamma} - 1}{1-\gamma}$$

Where W is wealth, γ is the coefficient of relative risk aversion, it's a constant. You can view this as where a and b are constants, $y = \frac{W^a}{a} - b$. You don't have to worry if $\gamma = 1$, $u(W)$ is undefined. We can substitute $u(W) = \ln(W)$. Solve the case of $0/0$ or ∞/∞ by using L'Hopital's rule (L'Hopital, 1696). So let's take the derivative of $u(W)$. $u'(W) = \frac{(1-\gamma)W^{-\gamma}}{1-\gamma} = W^{-\gamma}$. This simplifies to $W^{-\gamma}$. Secondly, the limit value is recalculated, and the original limit is converted to the limit after taking the derivative of the numerator and denominator, $\lim_{\gamma \rightarrow 1} \frac{W^{-\gamma-1}}{1-\gamma} = \lim_{\gamma \rightarrow 1} W^{-\gamma}$. The limit of $W^{-\gamma}$ is equal to $\lim_{\gamma \rightarrow 1} \frac{1}{W} = \frac{1}{W}$. By integrating back to the original formula, we can find that when $\gamma = 1$, $u(W) = \ln(W)$.

The second derivative is $u''(W) = -\gamma W^{-\gamma-1}$. Coefficient of relative risk aversion could be the second derivative divided by the first derivative and then times -1 and W , so $\gamma = \frac{-u''(W)}{u'(W)} W = \frac{-(-\gamma W^{-\gamma-1})}{W^{-\gamma}} W$. The reason for the negative sign is that if it is a concave utility function, the second derivative is negative and the first derivative is positive.

3. The relationship and impact of CRRA

3.1 Determining gamma for a specific group

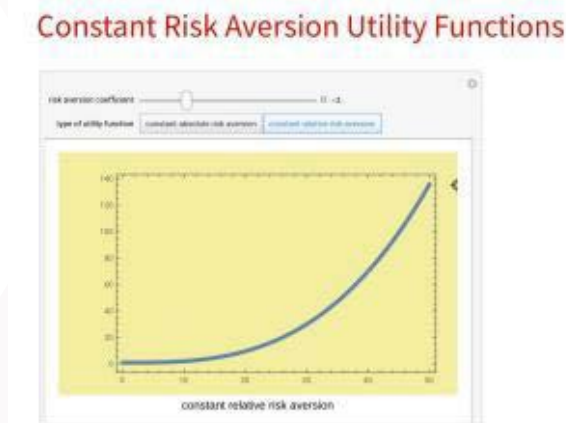
The Coefficient of Relative Risk Aversion (CRRA) takes on different values for individuals with varying degrees of risk aversion. If you are risk-averse, then $\gamma > 0$, please refer to Figure 4 (Chandler, 2011).

Figure 4 Graph with gamma greater than 0



X-axis represents wealth and y-axis represents utility. If you are risk-seeking, then $\gamma < 0$ (please refer to figure 5).

Figure 5 Graph with gamma less than 0



If you are risk-neutral, then $\gamma = 0$ (please refer to figure 6).

Constant Risk Aversion Utility Functions

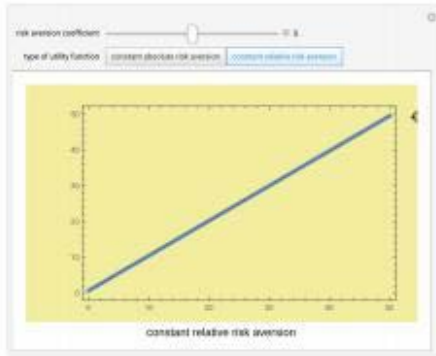


Figure 6 Graph with gamma equal to 0

Why does gamma have different results in different intervals? We can deduce that by taking the second derivative of the CRRA formula, $u(W) = -\gamma W^{-\gamma-1}$, which represents the speed at which the utility function rises. When $u''(W)$ is negative $-\gamma W^{-\gamma-1} < 0$, it means that as wealth (W) increases, the speed at which the utility function rises decreases, making the utility function concave. It follows that $\gamma > 0$, the investor is risk-averse. On the contrary, if $u''(W)$ were positive $-\gamma W^{-\gamma-1} > 0$, it would mean that the speed at which the utility function rises increases with wealth, making the utility function convex. So $\gamma < 0$, the investors are risk-seeking.

Moreover, $-\gamma W^{-\gamma-1} = 0$ it means that the rate of change of marginal utility does not change as wealth changes, so when $\gamma = 0$ the investor is risk-neutral.

3.2 The effect of increasing the amount of investment

If the investment amount increases, what impact does it have on utility? First, let's start with the CRRA formula itself.

$u(W) = \frac{W^{1-\gamma} - 1}{1-\gamma}$, total wealth before investment is W_0 . Increased amount of investment is ΔW , and W_1 is total wealth after investment. $W_1 = W_0 + \Delta W$. Plug W_1 and W_0 into the CRRA formula $u(W_0) = \frac{W_0^{1-\gamma} - 1}{1-\gamma}$ and $u(W_1) = \frac{(W_0 + \Delta W)^{1-\gamma} - 1}{1-\gamma}$ change of total utility (Δu) is equal to $u(W_1) - u(W_0)$,

$$\Delta u = u(W_1) - u(W_0) = \frac{(W_0 + \Delta W)^{1-\gamma} - 1}{1-\gamma} - \frac{W_0^{1-\gamma} - 1}{1-\gamma}$$

For example, suppose $W_0 = \$100$ and the increased investment $\Delta W = \$50$. Three situations: $\gamma = 2$ (risk-averse), $\gamma = -2$ (risk-seeking), $\gamma = 0$ (risk-neutral). The result is worked out, check out Figure 7.

	ΔU
$\gamma = 2$ (Risk-averse)	3.33×10^{-3}
$\gamma = -2$ (Risk-seeking)	791666.6667
$\gamma = 0$ (Risk-neutral)	50

Figure 7 The calculation of increasing the amount of investment

For risk-averse individuals, the increase in utility is smaller (because they are more concerned with risk). For risk-seeking individuals, the increase in utility is larger (because they enjoy risk). For risk-neutral individuals, the increase in utility is equal to the amount of investment.

4. The understanding of Expected utility and calculation

4.1 Concept of Expected Utility

Expected utility is the expected value of your future actions to you. It is used to analyze how to make decisions in unknown

situations. Expected Utility is calculated by summing the utilities of different outcomes weighted by their respective probabilities.
$$\text{Expected Utility} = \sum_i p_i \cdot U(C_i)$$
 In actual decision-making, decision-makers usually evaluate the Expected Utility, and the higher the Expected Utility the higher the overall expected satisfaction or benefit.

5. Calculation

Suppose you want to invest in a catering company. Investment project Yum! Brands would have the following results: W1 = 1200000RMB Probability1 = 0.5, W2 = 2400000RMB Probability2 = 0.5. McDonald's (MCD) : W3 = 1000000RMB Probability3 = 0.6, W4 = 2800000RMB Probability4 = 0.4. The calculations are divided into three scenarios (see Figure 7) for people with different levels of risk aversion.

Case 1: It is known that when $\gamma > 0$, the utility function is concave. So investors are risk-averse. Suppose $\gamma = 2$. Project Yum! Brands: $u(W1) = 1 - (1/1200000)^\gamma$, $u(W2) = 1 - (1/2400000)^\gamma$, $E[u(W12)] = (u(W1) + u(W2))0.5$. Project MCD: $u(W3) = 1 - (1/1000000)^\gamma$, $u(W4) = 1 - (1/2800000)^\gamma$, $E[u(W34)] = u(W3) \times 0.6 + u(W4) \times 0.4$. Compare the Expected utility results with $E[u(W12)] > E[u(W34)]$. So when you're risk-averse, invest in Yum! Brands get higher overall expected satisfaction and earnings.

Case 2: Given $r < 0$, the utility function is convex. So investors are risk-seeking. Suppose $r = -2$. Project Yum! Brands: $u(W1) = (1200000^{-2})^{-1/3}$, $u(W2) = (2400000^{-2})^{-1/3}$, $E[u(W12)] = 5.67 \times 10^{17} \times 0.6 + 4.608 \times 10^{18} \times 0.4$. Project MCD: $u(W3) = (1000000^{-2})^{-1/3}$, $u(W4) = (2800000^{-2})^{-1/3}$, $E[u(W34)] = u(W3)0.6 + u(W4)0.4$, $E[u(W12)] < E[u(W34)]$ it seems that when you would be a risk-seeking, Overall satisfaction and returns from investing in MCD are higher.

Case 3: Given $r = 0$, the utility function is linear and the investor is risk-neutral. Project Yum! Brands: $\gamma = 0$, $u(C1) = 1199999$, $u(C2) = 2399999$, $E[u(W12)] = 1299999$. Project MCD: $u(C3) = 999999$, $u(C4) = 2799999$, $E[u(W34)] = 1719999$. Draw the conclusion $E[u(W12)] > E[u(W34)]$. Without considering the risk, invest in Yum! Brands are a better choice.

	$u(W_1)$	$u(W_2)$	$E[u(W_{12})]$	$u(W_3)$	$u(W_4)$	$E[u(W_{34})]$	Result
Situation 1: risk-averse Gamma=2	0.99999 91667	0.99999 95833	0.99999 9375	0.99999 9	0.99999 96429	0.99999 92572	$E[u(W_{12})] > E[u(W_{34})]$
Situation 2: risk-seeking Gamma=-2	5.76×10^{17}	4.608×10^{18}	2.592×10^{18}	1.125×10^{18}	7.31733×10^{18}	3.6015×10^{18}	$E[u(W_{12})] < E[u(W_{34})]$
Situation 3: risk-neutral Gamma=0	1199999	2399999	1299999	999999	2799999	1719999	$E[u(W_{12})] > E[u(W_{34})]$

Figure 7 The table of three groups of people

6. Conclusion

This article mainly explains that utility theory measures the satisfaction or happiness derived from decisions. There are two main approaches: the Cardinal approach and the Ordinal approach. The Cardinal approach quantifies utility with numerical values, while the Ordinal approach ranks preferences without assigning numerical values. Furthermore, the utility function for risk-averse individuals is concave, for risk-seeking individuals it is convex, and for risk-neutral individuals it is linear. In the CRRA formula, we can use the second derivative to prove that when γ is greater than 0, it indicates risk aversion; when it is less than 0, it indicates risk-seeking; and when it is equal to 0, it indicates risk neutrality. Additionally, increasing the investment amount results in a smaller utility increase for risk-averse individuals compared to the other two types. We also explored that Expected Utility represents the anticipated satisfaction from various outcomes, each with a specific probability. It is calculated as the weighted average of the utility values of all possible outcomes. Using Expected Utility,

we evaluated the investments in MCD and Yum! Brands for individuals with different risk preferences found that both risk-averse and risk-neutral investors would prefer investing in Yum! Brands due to higher satisfaction and lower risk.

7. References

- [1] Elton, E. J. (2014). Modern portfolio Theory and investment Analysis. *The Journal of Finance*, 37(5), 1317. <https://doi.org/10.2307/2327857>
- [2] Ganti, A. (2024, June 22). Efficient frontier: what it is and how investors use it. Investopedia. <https://www.investopedia.com/terms/e/efficientfrontier.asp>
- [3] Henry, S., Plessis .K.D., Hysmith, R. (2023, May 31). Efficient frontier. Option Alpha. <https://optionalpha.com/learn/efficient-frontier>
- [4] James, C. (2024, May 16). Risk: what it means in investing, how to measure and manage it. Investopedia. <https://www.investopedia.com/terms/r/risk.asp>
- [5] Kenton, W. (2024, July 01). Capital asset pricing model (CAPM): definition, formula, and assumptions. Investopedia. [https://www.investopedia.com/terms/c/capm.asp#:~:text=Key%20Takeaways,to%20the%20market%20\(beta\).](https://www.investopedia.com/terms/c/capm.asp#:~:text=Key%20Takeaways,to%20the%20market%20(beta).)
- [6] Markowitz, H. M. (1952). Portfolio Selection. *Journal of Finance* (pp.77-91).
- [7] Nirmal Bang. (n.d.). Capital asset pricing model. Retrieved August 7, 2024, from <https://www.nirmalbang.com/knowledge-center/capital-asset-pricing-model.html>



05

Engineering of Sustainable Vehicles

Gyroscopic Effect

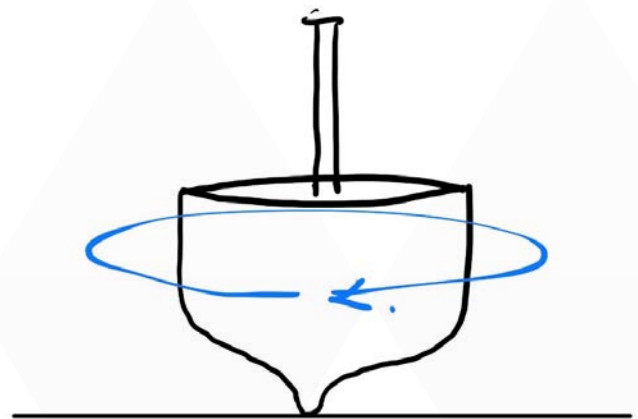
CHENGYU WANG

1. Abstract

As the one of the most well-known toys, you probably played spinning top before. But have you ever thought about how does it work? In this essay, we will first learn the basics concepts in rotational dynamics. By analysing the two examples of spinning tops, we will learn the Gyroscopic effect, and also study its application on motorcycles. After understanding this essay, you will be able to solve simple problems for the motion of spinning objects. Also, you will learn the gyroscopic effect and the principles supporting the spinning top works. Finally, you will know how gyroscopic effect can be applied into motorcycles to improve its stability and steering performance. As one of the most important principles in rotational dynamics, understanding the gyroscopic effect will be extremely beneficial for your further study in physics and engineering. Looking at the future, the gyroscopic effect has the huge potential to be applied on the medical devices, personal transportation and renewable energy.

2. Introduction

Gyroscopic effect is one of the most fascinating effects in Engineering. Have you have thought about why the spinning top can rotate for a while can keep not falling? Through the analysis of the gyroscopic effect, this essay will present the principles behind this phenomenon.



The main objectives of this essay are to introduce the basic principles of gyroscopic effect and to understand its applications in motorcycles.

3. Rotational Dynamics

Rotational Dynamics is the study of the motion and force of spinning objects which rotates around an axis. Comparing with the linear dynamics, many basic concepts are quite different, so we need introduce them before the analysis of gyroscopic effect.

In linear dynamics, we use displacement to measure the change in position. However, in rotational dynamics, we will use angular displacement to measure the angle which the object rotates through. There is also angular velocity and angular acceleration, where:

$$\omega = \frac{d\theta}{dt}$$

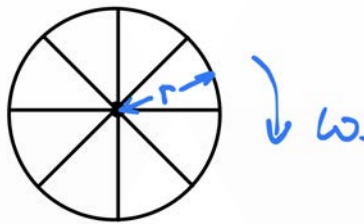
$$\alpha = \frac{d^2\theta}{dt^2}$$

In addition, mass is used to measure the resistance to change in the motion in linear dynamics. The moment of inertia is used to measure the resistance to change in the angular speed in rotational dynamics.

$$I \equiv mr^2$$

(r is the distance from the axis of rotation to the object)

(For example, for the r of a wheel:)



Force is the cause of change in motion in linear dynamics. In rotational dynamics, torque is the cause of change in the angular speed.

$$\tau = F \times r = I\alpha$$

Last but not least, in linear dynamics, momentum is a quantity which always conserves in an isolated system. Thus, there is angular momentum which also conserves when no external torque is applied in rotational dynamics, because the rate of change in angular momentum equals to the sum of all external torques applying on it.

$$L = r \times p = I\omega$$

$$\frac{dL}{dt} = \sum \tau$$

If you are interested in rotational dynamics, I highly recommend this video which can be helpful for you to learn more about and also deepen your understanding of these basic concepts:

<https://www.youtube.com/watch?v=yYysCEZHCro&t=722s>

4. Simple Rotational Motion

Now, we can go back to analyse the motion of spinning top. As shown in the diagram below, initially, we apply a force which causes to the torque applying on the spinning top. So due to the torque, the spinning top starts to rotate. We can calculate the torque applied on the spinning top:

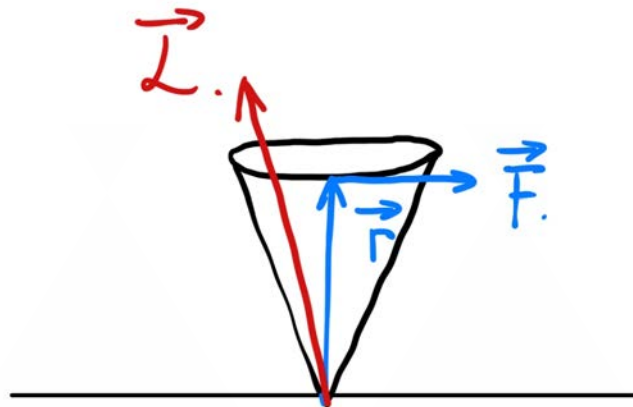
$$\tau = F \times r$$

In this case, the torque caused by the weight is zero because the direction of the weight, and displacement from the centre of rotation to the centre of mass are same.

From the previous section, we learn that:

$$dL = \tau \cdot dt$$

In this formula, since angular momentum L and torque τ are vectors, and time is a scalar instead, so the direction of angular momentum L is exactly equal to the direction of the torque τ.



Notice that because of the rule of cross product of vectors, the direction of torque and angular momentum is perpendicular to both of displacement and force, which is into the page.

Then, if we neglect the air resistance and any other resistance, because there is no other external torque acting on the spinning top, in this ideal situation, the angular momentum is conserved. Because:

$$L = I\omega$$

For this spinning top, its moment of inertia is a fixed constant. So, in this case, the angular velocity stays unchanged all time which means that the spinning top will always keep rotating with same angular speed.

However, in the real world, because of the effect of air resistance and friction, the angular velocity will gradually fall to zero.

In this example, the key point is that the rotational axis never changes in this process, which shows that the spinning object has the tendency to maintain its axis of rotation. Moreover, if the angular speed caused by the external torque of the object is larger, the angular momentum also increases which further strengthens the tendency.

5. Precessional Motion

After understanding the reason why the spinning top can keep steady, we can now focus a more complicated problem:

As shown in the diagram below, at this time, the spinning top is initially oblique. The angle between the axis of rotation and dotted vertical line is θ . We still act a force on it which generates a torque and make it rotate.

However, in this case, we can no longer ignore the torque caused by the weight. We decompose the weight into two forces: w_1 and w_2 . Again, the torque caused by the force w_1 is zero because the direction of the weight w_1 , and displacement from the centre of rotation to the application of force w_1 are same. The magnitude of the other force can be calculated in this way:

$$F = w \sin\theta = mg \sin\theta$$

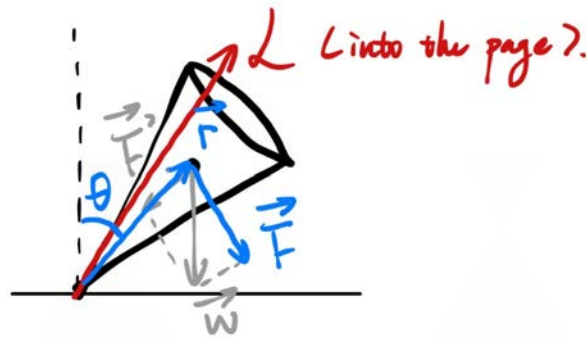
So, the torque caused by the weight equals to:

$$\tau = F \times r$$

Because:

$$\frac{dL}{dt} = \tau$$

This torque results in a change in change in angular momentum. The direction of change in angular momentum is equal to the direction of torque, and their directions are into the page again here.



The change in angular momentum will further cause a change in angular velocity of the rotational axis, which directs into the page. This makes the rotational axis rotate around the dotted vertical line anticlockwise, and then the central of rotation gradually becomes closer and closer to the vertical dotted line. Finally, the rotational axis will coincide with the dotted vertical line.

Physically, this kind of motion of which the axis of the spinning object moves caused by the external torque is known as precessional motion.

In this example, after external torque acting on the spinning object, we found that the axis of rotation moves as the result of this. However, through the analysis, we know that the axis of rotation became stable again after a while. So, from this example, we can get the conclusion that as a spinning object experiencing an external torque which causes to the movement of the rotational axis, the rotational axis has the tendency to become static again with time.

6. Gyroscopic Effect

These two examples above have fully shown the gyroscopic effect.

Gyroscopic effect is the tendency of a rotating object to maintain the direction of its axis of rotation (shown in example 1) and keep the axis of the rotation static (shown in example 2).

7. Application

The gyroscopic effect deeply affects the design of the motorcycle. I will analyse its application on motorcycles from the perspective of stability and steering.

In the first place, from the first example, we found that the axis of rotation tends to be unchanged when no external torque due to the effect of angular momentum. So, when the wheel is rotating in a high angular velocity, the motorcycle also tends to become stable, because of the high angular momentum.



In addition, from the perspective of steering, from the second example we learnt that the rotating object has the tendency to make its rotational axis become vertical to the ground. When cyclers change the direction of motorcycle, they will lean the body of motorcycle. In this situation, gyroscopic effect would help the motorcycle become vertical to the ground again after changing the direction.



8. Conclusions

After learning the basic concepts in rotational dynamics, through two examples, we learnt that gyroscopic effect describes the spinning object has the tendency to maintain its rotational axis and the tendency becomes stronger when angular velocity is greater. Also, when external torque is applied on the body, the rotational axis of the body attempts to become static. Finally, the gyroscopic effect shows why the motorcycles have stability and ability to lean.

9. References

- [1] Ingenuity, I. (2020, 6 8). What is Gyroscopic Effect? | Gyroscopic Effect on Airplane - Youtube. Retrieved from Youtube: <https://www.youtube.com/watch?v=3WZNE8EDKMw>
- [2] Mauritiu, c. f. (2010, 11 18). File:Spinning top (5448672388).jpg. Retrieved from wikipedia commons: [https://commons.wikimedia.org/wiki/File:Spinning_top_\(5448672388\).jpg](https://commons.wikimedia.org/wiki/File:Spinning_top_(5448672388).jpg)
- [3] Pröβdorf, S. (2022, 9 24). File:2022-09-24 Motorsport, IDM, Finale Hockenheimring 1DX 3890 by Stepro.jpg. Retrieved from wikipedia commons: https://commons.wikimedia.org/wiki/File:2022-09-24_Motorsport,_IDM,_Finale_Hockenheimring_1DX_3890_by_Stepro.jpg
- [4] Romain9247. (2008, 7 14). File:1200gsfr.jpg. Retrieved from wikipedia commons: <https://commons.wikimedia.org/wiki/File:1200gsfr.jpg>
- [5] SciTech, T. (2016, 5 29). Gyroscopic precession -- An intuitive explanation. Retrieved from Youtube: <https://www.youtube.com/watch?v=n5bKzBZ7XuM>
- [6] Shorts, S. (2019, 5 10). Rotational Dynamics - A-level Physics (Engineering). Retrieved from youtube: <https://www.youtube.com/watch?v=yYysCEZHCro&t=722s>
- [7] Wikipedia. (2024, 8 7). Bicycle and motorcycle dynamics. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Bicycle_and_motorcycle_dynamics

Nuclear Propulsion and its Future Possibilities

RUOXI WANG

1. Abstract

Nuclear energy is a new energy source, emitting a tremendous amount of energy with nearly zero pollution, combining the advantages of clean, green, and high energy density. Furthermore, vehicles powered by nuclear energy are predicted to be sustainable as they are durable and environmentally friendly. This report first discusses the powering process of nuclear submarines and their pros and cons. As nuclear submarines are pioneers in nuclear-powered vehicles, their propulsion system brings several helpful ideas to the future manufacturing of other nuclear vehicles. Then, this report gives an ideal model of the nuclear-powered car which might appear in the future by shrinking the structure of nuclear submarines and learning from the experience gained in that design and manufacturing process. However, barriers from different aspects arose while designing the ideal model, including technical limitations, the possibility of accidents, financial difficulties, and social controversy. Consequently, this report concludes that developing nuclear propulsion systems in daily-used vehicles depends quite a bit on other factors. Overall, this report is a simple overview of the nuclear energy generating process and the current use of nuclear propulsion systems. The ideal model presented and the problem evaluations might be quite helpful in the actual designing and manufacturing process of future nuclear vehicles.

2. Introduction

Nuclear energy is already used as an environmentally friendly energy source. It also seems attractive and theoretically capable of being used in daily vehicle propulsion systems, as the process of generating energy through nuclear fission and fusion has zero harmful gas emission and much greater energy produced per unit of fuel mass. However, several practical obstacles remain. This essay aims to provide details about current uses and attempts at nuclear-powered vehicles, analyse their pros and cons, and discuss the ideal model of nuclear-powered cars and the possibilities to achieve this goal.

3. About nuclear energy generating process

This paragraph will briefly introduce the nuclear energy generation process to help better understand its use. Nuclear power uses controlled nuclear reactions, nuclear fission, to obtain energy [1]. The whole process is shown in Fig 1. Nuclear fission involves breaking down massive unstable nuclei, such as Uranium 235, the most widely used nuclear fissile material. During the breaking, a large amount of energy is released, and this energy can transform into the heat energy of water and then the kinetic energy of vehicles.

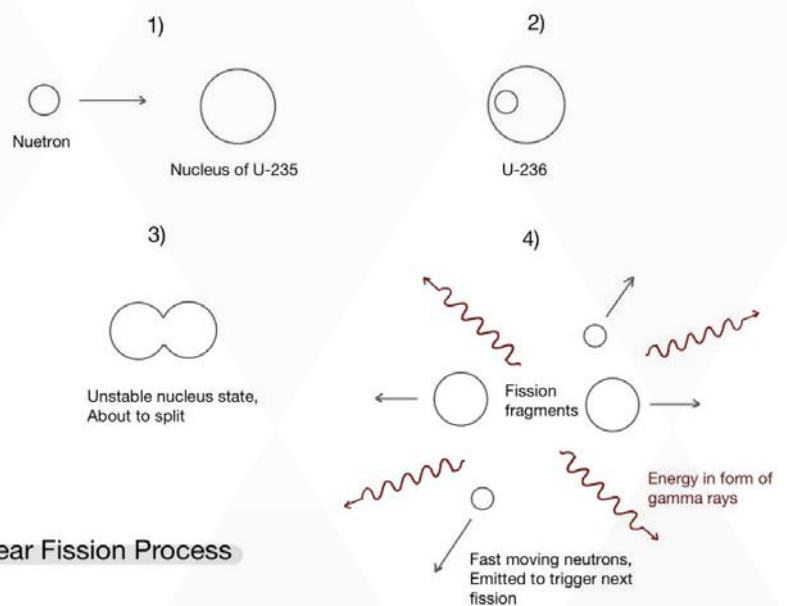
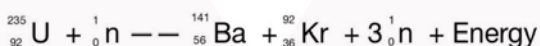


Fig. 1. (up) Shows process sketch of a typical Uranium-235 fission [2]

Fig. 2. (bottom) Shows equation of a typical Uranium-235 fission



Nuclear power will be desirable for vehicles in the future mainly because of its high energy density and low harmful gas emissions. According to Einstein's formula, $E=mc^2$, even a minimal loss in a material's mass can generate an extremely large amount of energy. That is much more efficient than traditional fossil fuels as the vehicle only needs a few fuels to support the long journey and does not need refuelling for a very long time. By calculation, uranium's lifetime fuel costs were \$3.17 in 2012, lower than the cost of one gallon of gasoline [3]. Along with the zero greenhouse gas emissions during the reaction process, nuclear power can be seen as a perfect energy source for the environment in the future.

4. Existing example of nuclear-powered vehicles: submarine

There are now a few examples of nuclear-powered vehicles, including nuclear submarines. A nuclear submarine is mainly powered by its internal nuclear reactor, where the nuclear fission happens, and the most-used reactor in a submarine is a pressurised water reactor (PWR). Here is the basic structure of a propulsion system in a nuclear submarine.

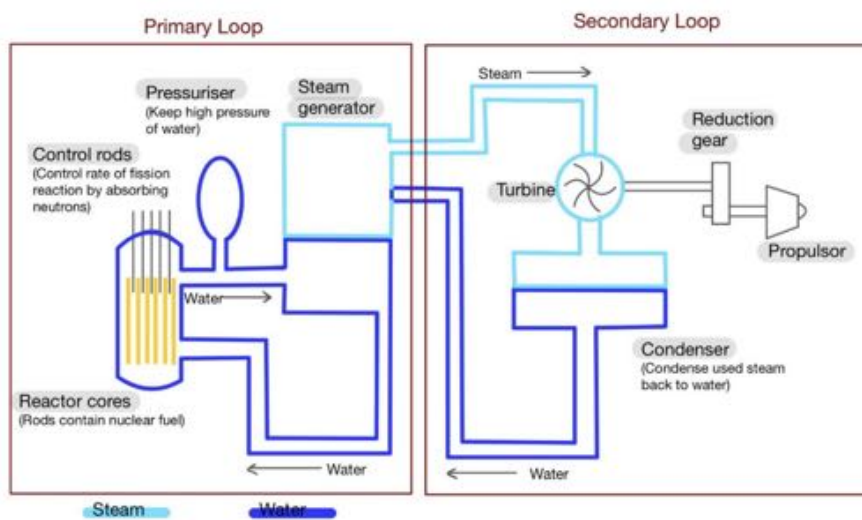


Fig. 3. Basic structure of nuclear propulsion system in a submarine

During the energy-generating process through nuclear fission, the nuclear fuel inside the reactor cores undergoes nuclear fission, releasing a large amount of heat energy absorbed by pressurised water in the primary loop. Water keeps in a liquid state due to the high pressure, carries the energy into the secondary loop in liquid state, then vaporises in steam generator, pushes the turbine to rotate, converting the heat energy of water to kinetic energy, and finally drives the propulsor of the submarine, power it to move [4]. Then the steam turns back to liquid through a condenser, back to the first loop and repeats the process.

The nuclear-powered submarine takes advantage of the predicted advantages of nuclear power: Firstly, it has high endurance and energy density as the fuel it carries only at one time can support 20-year travel [5], much higher than the duration of the conventional submarine (two weeks). Secondly, it has high speed, almost twice the speed of typical submarines [6], as the reactor has high power and efficiency. Thirdly, there is nearly no greenhouse gas emissions, as no carbon dioxide is produced. Although nuclear submarines are only used for military purposes due to their complexity and high manufacturing cost, their successful propulsion system still presents potential possibilities for using nuclear energy for other daily vehicles.

5. An ideal model for nuclear propulsion system for a car

This part considers the propulsion system of a nuclear-powered car as a scaled-down model of a nuclear submarine. By

referencing the structure of the submarine, it gives some characteristics of an ideal nuclear-powered car model.

- Fast speed and flexibility in changing speed by using efficient nuclear power process. Additionally, a streamlined shape can help to obtain this goal from aerodynamic perspectives.
- Need a much smaller but strong and efficient nuclear reactor to suit the size of a car.
- Zero greenhouse gas emissions.
- Minor sound provides, within 100 dB, the criteria of "noise" [7]. This point is important as cars are used in large number in cities and other environments.

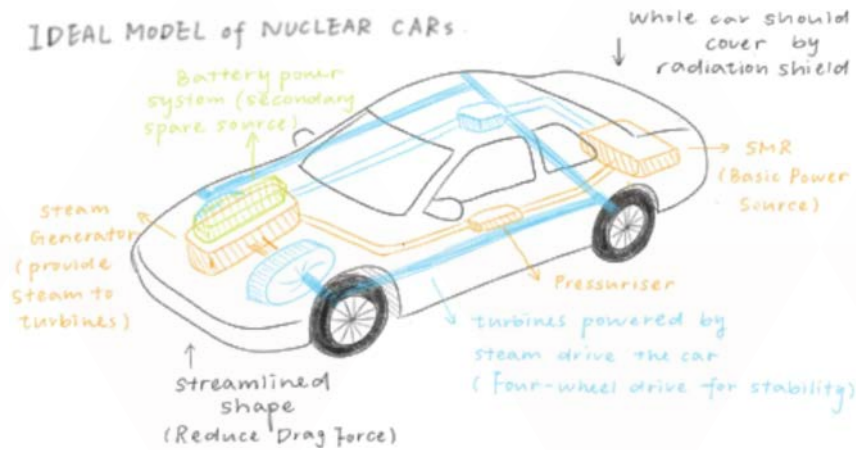


Fig. 4. Ideal model of nuclear propulsion system in a car

6. Practical barriers to reach ideal nuclear propulsion

Despite the advantages, there are still many inevitable obstacles that need to be considered to achieve the ideal vehicle discussed above:

- It is hard to find smaller nuclear reactors to suit a small volume of daily vehicles now. It is apparent that the reaction inside a nuclear reactor is very complex from above, so few countries aim to manufacture small modular reactors (SMRs); for example, Rolls-Royce in Britain has recently announced to continue to invest in SMRs project to produce reactors with the length of about a few metres [8]. However, using it in daily vehicles might need an even smaller reactor. Only a few countries are now determined to commercialise SMRs for daily use [9] as the number of people needed is too small. However, the cost is equal to that of large reactors.
- Other disposals besides gases, such as radioactive waste products produced in the fission process, also contaminate the environment; these by-products can severely harm people's health as they can ionise living cells in the body and damage DNA. So, future materials used in cars must withstand more aggressive situations, limiting the leakage of radioactive disposals.



Fig. 5. Chernobyl nuclear disaster (1986) [10]



Fig. 6. Fukushima nuclear disaster (2011) [11]



- Concerns about the safety of nuclear-powered vehicles are dominant. People might not easily accept the wide use of nuclear cars, as any accidents with nuclear reactions are destructive, such as those at Chernobyl (1986) and Fukushima (2011). Also, after the American nuclear submarine Hartford accident, traces of Pu-239, anthropogenic pollutants, were detected in several algal species in Sardinia, Italy [12]. So, as nuclear accidents have severe and prolonged effects, vehicles that are used daily should only use this kind of energy to power if they are 100% safe in any severe accidents, which is hard to attain. Furthermore, the social attitude toward nuclear-powered cars is not optimistic due to their environmental impacts.
- Manufacturing innovative vehicles with nuclear reactors requires substantial financial support, which is usually hard to find due to the long development time and uncertain success.
- The large amount of water used in the coolant system makes these cars unsuitable for regions with limited water availability.

Finding methods to cope with these barriers is an essential first step in achieving extensive use of nuclear power in cars.

7. Conclusion

This report widely discusses the principle and current usage of nuclear energy and also brings some predictions about its further applications in the transportation field. Undoubtedly, using nuclear energy instead of fossil fuels in cars can help reduce pollution and increase efficiency. However, this idea has several limitations. So, whether nuclear energy can be used for vehicle propulsion is still questioned; it strongly depends on whether manufacturers can quickly remove the current limits. However, nuclear energy can still be seen as an energy source with potential.

8. References

- [1]: Wikipedia. (2007). Nuclear power. https://en.wikipedia.org/wiki/Nuclear_power
- [2]: Based on USNRC Technical Training Center. Reactor Concepts Manual. <https://www.nrc.gov/reading-rm/basic-ref/students/for-educators/02.pdf>
- [3]: Claire Durkin. (2012). Nuclear powered Passenger Vehicles. PH241. <http://171.67.100.116/courses/2012/ph241/durkin1/>
- [4]: (2023). 核潜艇动力分析 . <https://b23.tv/laSf5Pn>
- [5], [6]: Wu Xie. (2018). 核动力潜艇比常规动力潜艇有什么优势和缺点? China Science Communication. <https://m.gmw.cn/baijia/2018-07/12/29825231.html>
- [7]: Noise limiter. <https://noiselimiters.co.uk/buy/images/slvls.jpg>
- [8]: Tracey Honney. (2024). Rolls-Royce seeks SMR business investors. Nuclear Engineering International. <https://www.neimagazine.com/news/rolls-royce-seeks-smr-business-investors/>
- [9]: Wikipedia. (2010). Small modular reactor. https://en.wikipedia.org/wiki/Small_modular_reactor
- [10]: graph from The Legacy of the Chernobyl Nuclear Disaster. <https://www.worldatlas.com/history/the-legacy-of-the-chernobyl-nuclear-disaster.html>
- [11]: graph from Fukushima Five Years After Nuclear Disaster. https://www.nytimes.com/interactive/2016/world/asia/japan-fukushima-anniversary.html?_r=0
- [12]: Cristaldi Mauro. (2006). Nuclear Powered Submarines as Hazards for the Marine Environment. <https://iris.uniroma1.it/handle/11573/97992>



06

Physical Chemistry

Quantum tunnelling and chemistry

LUOFAN WU

1. Introduction

Quantum tunnelling is a phenomenon where objects, or mostly microscopic objects pass through a potential barrier with insufficient kinetic energy. This occurs due to the particle-wave duality of objects, the wave function of the particle can pass over the barrier that is classically impossible.

Quantum tunnelling has been used to explain a range of phenomena that are impossible in classical physics, such as fusion and alpha decay. Nowadays, tunnelling is used in fields such as scanning tunnelling microscopy, semiconductors, and superconductors. It also plays roles in chemical reactions that involve light atoms such as hydrogen.

2. Background information

Friedrich Hund first proposed tunnelling effect in 1927, he applied Schrödinger equation to a phenomenon of particles tunneling through a barrier with insufficient energy.[1][2] Also, George Gamow used this phenomenon to explain alpha decay, and he used Schrödinger equation to find out the relationship between the half-life of the particle and the energy of emission, which directly depends on the probability of tunnelling.[3] Later, in 1959, Leo Esaki invented Esaki diode by discovering the tunnel effect of electrons in semiconductors,[4]

3. Mechanism Wavefunction

The mechanism of quantum tunnelling can be understood by the system below (Fig 1)

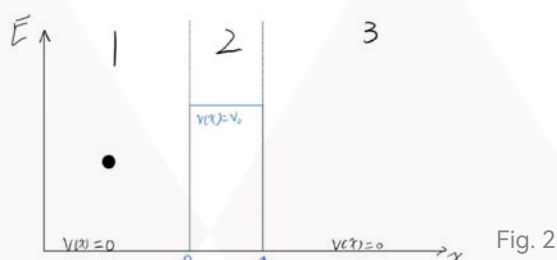


and time-dependent Schrödinger equation [5]:

$$-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} \psi(x) + \hat{V}(x) \psi(x) = E \psi(x)$$

In the system shown in Fig1, the horizontal axis is displacement of particle and vertical axis is the energy. The blue area in the middle represents a finite potential barrier with potential energy V_0 .

The system consists of a free electron that only moves in one dimension, at the left of the barrier, and it carries a lower potential energy than the barrier.



This system can be split into three regions:

Region 1: $x < 0$, where potential energy $V(x)$ is 0

Region 2: $0 < x < a$, where potential energy $V(x)$ is V_0

Region 3: $x > a$, where potential energy $V(x)$ is 0

After solving the Schrödinger equation we can discover that the wave function $\Psi(x)$ at region 1 and 3 are all complex exponentials, which describes a wave that is stationary and unaltered by a potential; while in region 2 the wave function is real exponential, represents a wave that has non-oscillate probability density. The wave function of a particle incident to the potential barrier is shown in Fig 3. The amplitude of the wave shows the likelihood of finding the particle at that position.

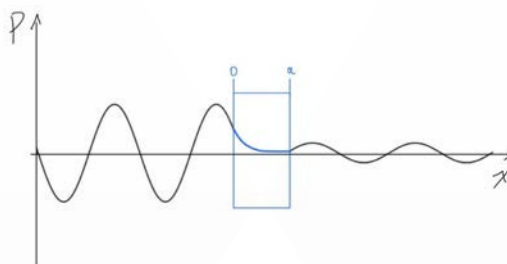


Fig. 3

Before entering the barrier, the particle shows a wave behaviour. As it enters the barrier, it becomes an evanescent wave that experiences an exponential decay in the barrier. After passing through, the particle resumes its wave behaviour on the other side of the barrier, with a reduced amplitude. This phenomenon demonstrates that the particle has tunnelled through the barrier, even not meet the classical energy requirement to do so. Since the amplitude drops below zero, the real probability of finding the particle requires further treatment.

Probability and factors affect it

The real physical probability of the appearance of the particle can be calculated by:

$$|\Psi(x)|^2$$

As the wavefunction has become probability:



Fig. 4

In Fig4, it is shown that the probability of finding the electron at region 1 and 3 respectively consists of constant values. In region 2 it experiences an exponential decay where the probability is defined as the following equation[6]:

$$P = e^{\frac{-4x\pi}{h}\sqrt{2m(V-E)}}$$

According to the equation, as the length of the barrier or the mass of the particle increases, the probability will decrease. It never drops to zero but can be small enough to be considered negligible.

4. Relation to chemical reactions

Chemical reactions have an energy barrier called activation energy, this must be achieved to convert reactants to products. Chemical kinetics indicates that activation energy can be more easily passed by increasing the temperature or pressure or

adding the catalyst.

Tunnelling can provide an alternative mechanism for reaction, which does not require increase reactant energy to proceed the reaction. According to P. Zuev et al. (2003),

who carried out ring expansion of 1-methylcyclobutylfluorocarbene at a temperature of 8K, the actual rate of reaction is 152 orders of magnitude greater than pass through the activation energy normally [7].

Fig 5 below shows the formation of a new five-membered ring.

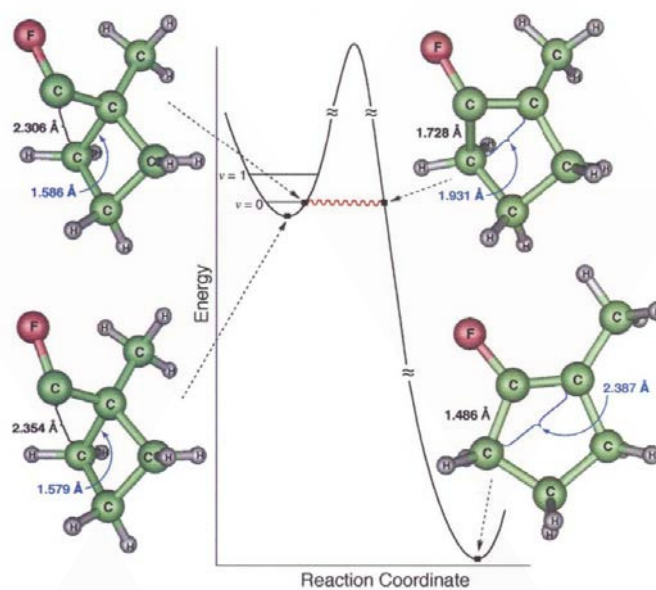


Fig. 5

The reactant molecule at the lowest energy vibrates as a whole and causes a change in distance between atoms, the molecule subsequently tunnels through and forms the molecule at the top right of the figure, which is the product at a higher vibrational level. This molecule has new bond forms according to atom distance, and the product can be formed by its vibrational relaxation. Overall, this greatly increases the rate of formation of the product.

Another paper published by N. Scrutton et al. (1999) about protein dynamics in driving enzymatic H-tunnelling showed the relationship between enzyme catalysis and hydrogen tunnelling. In the paper, Scrutton proved that H-transfer by a process of ground-state tunnelling occurred at the enzyme-catalyzed reaction, where it overcomes a range of difficulties and increases the efficiency of catalysis based on vibrationally enhanced ground-state tunnelling theory in the paper [8].

5. Conclusion

Quantum tunnelling is a result of particle-wave duality, objects (mainly microscopic particles) can pass through a barrier with a probability that decays exponentially according to the mass of the particle and thickness of the barrier. This phenomenon causes a great increase in the rate of chemical reaction, by allowing the reactants to form products without reaching activation energy.

6. Bibliography

- [1] F. Hund, Z. Phys. 40, 742 (1927).
- [2] F. Hund, Z. Phys. 43, 805 (1927).

[3] G. Gamow, Z. Phys. 51, 204 (1928).

[4] L. Esaki, New Phenomenon in Narrow Germanium Para-Normal-Junctions, Phys. Rev., 109, 603-604 (1958)

[5] Professor Dave Explains (2020). Unpacking the Schrödinger Equation. YouTube. Available at: <https://www.youtube.com/watch?v=l4fFG2utivw> [Accessed 5 Aug. 2024].

[6] Engel, T. and Hehre, W.J. (2010). Quantum Chemistry & Spectroscopy.

[7] Zuev, P.S. et al. (2003) Carbon Tunneling from a Single Quantum State, Jstor. Available at: https://www.jstor.org/stable/pdf/3833607.pdf?casa_token=sFlhXSONibgAAAAA:aULzll1-9lyC36quy6sxpmdN9HoicOKqYegUchJwP45rVJp3JnsLoqZjat9HH_SJAuN0zzDilRs7povAljKtX-s9xeMZcS2A2HA2gDNewdVEWssBt38 (Accessed: 08 August 2024).

[8] Scrutton, N.S., Jaswir Basran and Sutcliffe, M.J. (1999). New insights into enzyme catalysis. Ground state tunnelling driven by protein dynamics. European journal of biochemistry, 264(3), pp.666-671. doi:<https://doi.org/10.1046/j.1432-1327.1999.00645.x>.

Explain how the presence of a magnetic field can affect reaction kinetics (radical pair mechanism)

SZE WU WANG

Reaction kinetics is a crucial area in physical chemistry that can be applied to many fields in everyday life to control the rate of chemical reactions. From photosynthesis to combustion in car engines (Wong, 2020), it can be affected by various factors such as temperature, concentration of reactants, and the presence of catalysts. Magnetic field, however, is often overlooked despite the alterations it could cause in chemical processes. In fact, even the Earth's magnetic field, which is hundreds of times weaker than a refrigerator magnet can play an essential role in affecting reaction kinetics. This essay will demonstrate such phenomenon via the radical pair mechanism: the underlying mechanism that allows animals to navigate.

A chemical interaction is likely to be significant when its energy exceeds (Hore, 2019), the energy of each molecule's thermal motion, where k_B is Boltzmann's constant, and T is temperature. In room temperature (298K), $k_B T$ equals 2.479 kJ/mol (Milo et al, 2010), and the energy of a molecule's interaction with the geomagnetic field is almost ten million times smaller than such value, making it thermodynamically insignificant. However, this does not apply to radical pair mechanism, which involves a highly non-equilibrium state where even the smallest energy applied would be significant.

Electrons possess the property of spin angular momentum, which causes the formation of magnetic fields around them. Paired electrons have opposite spins, hence produces opposite magnetic moments which cancel each other out (Walling, 2024). Radicals, however, have unpaired electrons, which makes them paramagnetic. Radical pairs require two radicals to be created simultaneously and can exist in a singlet ($S=0$, antiparallel spins) or triplet state ($S=1$, parallel spins). Atomic nuclei with an odd number of protons or neutrons could also have spin and are hence magnetic. This allows hyperfine interactions to occur, meaning the nuclear spins interact with electron spin in radicals. According to Wigner's rule, spin should be conserved during chemical reactions. For instance, for a pair of methyl radicals, it would be thermodynamically favourable to recombine and form ethane as it is much more stable. However, for this reaction to occur, the bond formed between the two carbons must contain two electrons with opposite spins due to Pauli's principle, which states that if the radicals were initially in a triplet state, they would not form ethane due to their spins being conserved. But radicals can be converted from triplet to singlet state through the application of an external magnetic field as it could influence the direction of spin. This epitomises that magnetic fields could actually affect the yield of chemical products – by applying a field or changing its direction, we could, for example, favour ethane as the product of the reaction.

A radical pair and at least one hyperfine interaction are necessary for radical pair mechanism. The mechanism proposes that when a molecule absorbs a photon, an electron is excited and donated to the acceptor molecule, forming a radical pair in a singlet and highly non-equilibrium state. Hyperfine interactions then occur within the radical pair, which breaks the symmetry of its initial parallel spins, causing them to oscillate between singlet and triplet states (Adams et al, 2018). Such oscillations are sensitive to external magnetic fields, which would then determine the yields of chemical products formed. This theory has been proven through a molecule called carotenoid-porphyrin-fullerene (CPF) triad (Rodgers & Hore, 2009), where porphyrin acts as the photon acceptor and is followed by two rapid electron transfers, producing the radical pair in its singlet state. Essentially, when porphyrin absorbs a green photon, an electron reaches an excited state and is transferred to the fullerene, producing a radical pair. This is followed by an electron being transferred from the carotenoid to replace the original electron in the porphyrin, hence overall resulting in a radical pair: one radical on the carotenoid and another on the fullerene. Spectroscopic experiments have shown that the lifetime of the radical pair produced can be affected by extremely weak magnetic fields of 50 μT , which is approximately the same as the magnitude of the geomagnetic field.

In a biological context, the radical pair mechanism is manifested through magnetoreception, a type of sensing that allows animals to detect the Earth's magnetic field. A protein called cryptochrome has been identified as a photoreceptor that forms radical pairs when blue photons are absorbed. They are naturally occurring, and are found in many living organisms,

including the eyes of migratory birds. Cryptochromes contain a molecule called flavin adenine dinucleotide (FAD) with chains of tryptophan amino acids. Similar to the CPF molecule, when blue light is absorbed by the FAD molecule, an electron is transferred from the tryptophan to the flavin, which causes the tryptophan to attract another electron from the second tryptophan, resulting in a radical pair in singlet state (Hore & Mouristen, 2022). The radical pair then interconverts between singlet and triplet state, which can be influenced by the geomagnetic field depending on its direction. The radicals can then return to its original position or undergo a conformational change to form a different protein depending on their state, and the new protein could cause the release of neurotransmitters to the birds' brain.

Therefore, this essay conclusively demonstrates that provided a highly non-equilibrium state, a seemingly insignificant chemical interaction with external magnetic fields can play an essential role in influencing reaction kinetics and can be essential to explaining some of the most fascinating biological phenomenon.

References

- [1] Adams, B., Sinayskiy, I., & Petruccione, F. (2018). An open quantum system approach to the radical pair mechanism. *Nature*, 8(1). <https://doi.org/10.1038/s41598-018-34007-4>
- [2] Hore, P. (2019, May 24). Peter Hore on Radical pair mechanism of magnetoreception [Video], FENS. <https://www.youtube.com/watch?v=FytxLiHlah4>
- [3] Hore, P. J., & Mouritsen, H. (2022, April 1). How migrating birds use quantum effects to navigate. *Scientific American*. <https://www.scientificamerican.com/article/how-migrating-birds-use-quantum-effects-to-navigate/>
- [4] Milo R., et al (2009). kT in units of kJ/mol - Generic - BNID 107841. (n.d.) <https://bionumbers.hms.harvard.edu/bionumber.aspx?id=107841>
- [5] Rodgers, C. T., & Hore, P. J. (2009). Chemical magnetoreception in birds: The radical pair mechanism. *Proceedings of the National Academy of Sciences*, 106(2), 353–360. <https://doi.org/10.1073/pnas.0711968106>
- [6] Walling, C. T. (2024, August 10). Radical | Reactions, Properties & Uses. *Encyclopedia Britannica*. <https://www.britannica.com/science/radical-chemistry/Magnetic-properties-of-free-radicals>
- [7] Wong, J. (2020, December 25). Application of Reaction Kinetics in Everyday Life. *My Chem Cafe*. <http://mychemcafe.com.sg/2020/12/application-of-reaction-kinetics-in-everyday-life/>



07

Medicine

How does the body respond to haemorrhage?

WAI WONG

A haemorrhage occurs when vascular tissue is damaged, causing acute blood loss. Any and all types of bleeding can be thus defined as a haemorrhage. The response of the human body in event of a haemorrhage is complex and often differs according to the severity and location of the haemorrhage, as well as the aid rendered to the afflicted body. This essay will explore the short term (almost immediate) responses of the body, where the target is to halt acute blood loss and stabilise systems affected by the decrease in volemia, along with the long-term response, where the objective would be the replacement of blood and repair of tissue damage. I will be delving into the mechanisms of the cardiovascular system, the nervous system, and the RAAS system, as well as their interactions.

Before discussing the response to a haemorrhage, an outline of its effects is needed. A haemorrhage would first cause a decrease in volemia as less blood returns to the heart per unit time. This then reduces CO because of the decreased SV (See see Equation 1). The reduced CO brings a tendency to decrease regarding arterial blood pressure, calculated by multiplying CO with TPR (See see Equation 2). Diminished arterial blood pressure negatively affects tissue perfusion as the driving force for blood through capillaries is reduced. The inadequate perfusion, if allowed to persist, may cause hypovolemic shock through extensive hypoxia at a cellular level[3][4][7]. The last stage of this process would be multi-organ failure due to a lack of nutrients and oxygen for normal bodily function, which results in a cascade of problems.

Equation 1: $(CO=HR \times SV)$

Equation 2: $(MAP=CO \times TP)$

The body would attempt to prevent such an outcome from happening. To respond to the haemorrhage, the body must first detect the abnormal reduction in volemia. The detected variable would not be volemia; instead, arterial baroreceptors, located in the aortic arch and the carotid sinus, flag the abnormal reduction in arterial blood pressure and send information to the medulla oblongata[1]. The sympathetic nervous system is then further activated, and various responses begin. Tachycardia occurs as a result of sympathetic signalling and increased adrenaline production. The increase in heart rate compensates for the decrease in stroke volume to attempt to produce the same CO as pre-haemorrhage levels. Vasoconstriction in non-essential organs occurs to increase TPR, thus directing the remaining blood to essential organs.

The adrenal medullae situated at the apex of the kidneys respond by releasing adrenaline, which complements vasoconstriction and tachycardia already present in the cardiovascular response. In addition to redirection of residual blood to essential areas through increasing TPR, the CNS reduces activity in non-essential systems to minimise their metabolic requirements, thus requisitioning more blood for vital systems. Residual blood flow direction would also naturally favour the vital organs such as the brain and heart as they have a constant high requirement for oxygen and nutrients due to their higher per kilogram metabolic rate. The volume of tissue fluid within capillary beds decreases by flowing back into capillaries to help maintain blood volume. This is beneficial for the maintenance of adequate pressure and perfusion in essential areas and complements the aforementioned mechanisms.

The RAAS is activated by a combination of three distinct variables; reduced renal perfusion, stimulation by the sympathetic nervous system, and decreased sodium concentration at the macula densa within the kidney[2]. The system begins with the release of the renin enzyme from the juxtaglomerular cells located in the kidneys. The enzyme then catalyses the conversion of angiotensinogen, an inert precursor protein manufactured in the liver, into angiotensin I in the bloodstream. The angiotensin I will then be further converted to the peptide hormone angiotensin II by ACE. Angiotensin II functions as a potent vasoconstrictor alone and complements other critical roles in the RAAS as well[8]. Notably, it stimulates the release of the peptide hormone vasopressin (ADH) and the steroid hormone aldosterone. Vasopressin increases the rate of water reabsorption in the kidneys, and aldosterone promotes the reabsorption of sodium (as does angiotensin II)[6]. They both

function to increase volemia given that both water and salt must be replaced. Angiotensin II has some additional effects, such as acting on the hypothalamus to stimulate the thirst sensation since increasing fluid intake can help restore volemia. It also increases the activity in the sympathetic nervous system, which as previously mentioned, is responsible for many responses that deal with haemorrhage.

The body's long term response to a haemorrhage is very different. In the short term, the objective is to stabilise arterial pressure and combat adverse effects of the haemorrhage, whereas the long term response will focus on permanently repairing damaged structures and replenishing volemia. The two main effects of a haemorrhage can be categorised as fluid loss and structural damage. Fluid loss is remedied by erythropoiesis and restoration of volemia through erythropoietin release by the kidneys and increased production of plasma proteins as well as stimulation of thirst to encourage fluid intake respectively. Tissue damage repair will involve haemostasis and inflammation, which fights infection and removes debris. Proliferation(restoring function) and remodelling(reducing scar tissue) follow as the actual repair processes for creating new tissue[5].

However, the aforementioned responses are not perfect counters to a haemorrhage. It comes down to whether the rate of blood loss is high enough to overwhelm the short term responses. If this is the case, MAP decreases due to the insufficient blood. The reduced MAP has a direct negative impact on perfusion of the heart. Muscle tissue with inadequate perfusion can no longer contract or extend with its normal intensity. Thus, the under-perfused cardiac muscles' contractility is reduced. Reduction in contractility further decreases heart rate and stroke volume. This in turn further reduces perfusion, setting up a positive feedback loop where reduced perfusion exacerbates itself through the action of its direct effects. The persistence of this loop ultimately causes the cessation of normal bodily function and death.

In summary, the body responds to a haemorrhage through cascade mechanisms to prevent further losses, maintain arterial blood pressure and repair tissue damage, some of which is triggered by the sympathetic nervous system, which causes various cardiovascular responses as well as promoting certain endocrine secretions. The wound area is repaired and restored to original functionality in the long term. Understanding the physiology involved in this process would allow for optimal medical intervention to benefit the patient.

References

- [1] Armstrong, M., Moore, R.A. and Kerndt, C.C. (2023). Physiology, Baroreceptors. [online] National Library of Medicine. Available at: <https://www.ncbi.nlm.nih.gov/books/NBK538172/> [Accessed 2 Aug. 2024].
- [2] Fountain, J.H., Lappin, S.L. and Kaur, J. (2023). Physiology, renin angiotensin system. [online] National Library of Medicine. Available at: <https://www.ncbi.nlm.nih.gov/books/NBK470410/>.
- [3] Hooper, N. and Armstrong, T. (2022). Hemorrhagic Shock. [online] National Library of Medicine. Available at: <https://www.ncbi.nlm.nih.gov/books/NBK470382/>.
- [4] Koya, H. and Paul, M. (2023). Shock. [online] Nih.gov. Available at: <https://www.ncbi.nlm.nih.gov/books/NBK531492/>.
- [5] Schultz, G.S., Chin, G.A., Moldawer, L. and Diegelmann, R.F. (2024). Principles of Wound Healing. [online] Nih.gov. Available at: <https://www.ncbi.nlm.nih.gov/books/NBK534261/#:~:text=Angiogenesis> [Accessed 7 Aug. 2024].
- [6] Scott, J.H. and Dunn, R.J. (2023). Physiology, Aldosterone. [online] PubMed. Available at: <https://www.ncbi.nlm.nih.gov/books/NBK470339/>.
- [7] Standl, T., Annecke, T., Cascorbi, I., Heller, A.R., Sabashnikov, A. and Teske, W. (2018). The nomenclature, definition and distinction of types of shock. *Deutsches Aerzteblatt Online*, [online] 115(45), pp.757–768. doi:<https://doi.org/10.3238/>

arztebl.2018.0757.

[8] Morris, D.L., Sanghavi, D. and Kahwaji, C.I. (2020). Angiotensin II. [online] PubMed. Available at: <https://www.ncbi.nlm.nih.gov/books/NBK499912/>.

How Can the Different Gasses in the Anesthetic Circuit be Measured? Can all Methods be Used for all Gasses?

JAMAL AL LOGMAN

1. Abstract

The monitoring of gas in an anesthetic circuit is essential to a patient's safety during an anesthetic intervention. Through the use of equipment such as paramagnetic oxygen analysers, and IR spectroscopy analysers, doctors can receive detailed measurements to then take the necessary responses for patient care. This review will explore the different techniques and principles used in hospitals to take accurate measurement of the gasses flowing through an anesthetic circuit.

2. Introduction

Monitoring the different gasses in the anesthetic circuit is of great importance. This is reflected by the emphasis of both the Association of Anesthetists of Great Britain and Ireland (AAGBI); and the Royal College of Anesthetists National Audit Project 4 (NAP 4) has put on monitoring gas concentrations during anesthetic interventions, [1]. This essay will discuss the different methods anesthetists use to measure the concentration of gasses in an anesthetic circuit and explain the principles and reasoning behind the techniques.

3. Mainstream vs Sidestream Monitoring Systems

Contemporary monitoring systems are split into two primary systems: mainstream and sidestream monitoring systems. A mainstream monitoring system is incorporated directly within the breathing system, resulting in fast and accurate measurements of gas concentrations in the circuit. However, rather than being used in surgical theater, mainstream monitoring systems are commonly found in pre-hospital care or emergency departments. This is a direct result of their limitations as they can only measure one gas at a time while adding weight to the breathing system. The alternative is the sidestream monitoring system, which can sample multiple gasses simultaneously at a constant rate via a tube that transports the sampled gas to the sampling site. The extraction point for gas, accessed via a sampling port, is generally situated within an airway adapter. This placement is strategically chosen to facilitate the capture of gas specifically during the exhalation phase [2]. Sidestream systems serve as the primary choice for use within an operating room. Nonetheless, it is important to note that due to the nature of gasses traveling within a tubular system, results offered by a sidestream monitoring system are delayed and are susceptible to gas condensation; however, in practice, this is not a significant concern [3].

4. Measurement of Oxygen

Depending on the gas being measured, a different technique will prove to be most optimal. Notably, oxygen (O_2), unlike other gasses, contains valence electrons that have yet to be paired. This characteristic makes oxygen a paramagnetic gas, thus being attracted to magnetic fields [4]. Paramagnetic oxygen analyzers leverage the inherent paramagnetic properties of oxygen to achieve precise measurements, and their use has been established in various applications. [5]

Modern paramagnetic oxygen analyzers consist of two chambers: a sampling chamber and a reference chamber with a pressure transducer in between. The sampling chamber is filled with oxygen from the breathing circuit via a sampling tube, while the reference chamber is filled with a predetermined amount of oxygen as illustrated in Figure 1. Oxygen molecules within the two chambers are attracted and agitated due to a generated magnetic field, turning on and off rapidly. The induced attraction and agitation of oxygen molecules results in pressure changes on either side of the pressure transducer. The pressure difference across the transducer is proportional to the partial pressure difference of the two oxygen samples (sample and reference). The measured pressure difference can then be transformed into electrical signals that are directly

proportional to the concentration of oxygen, displayed as a percentage on a screen [6].

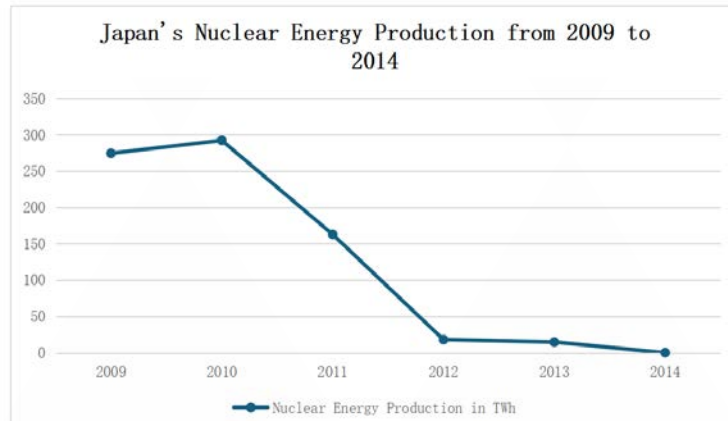


Figure 1: A schematic displaying the different chambers and position of the pressure transducer within a paramagnetic oxygen analyzer [7].

5. Measurement of Carbon Dioxide, Nitrous Oxide, and Volatile Agents

In contrast, other gasses used in an anesthetic circuit, including carbon dioxide (CO₂), nitrous oxide (N₂O), and volatile agents, are considered to exhibit diamagnetism traits. Unlike oxygen molecules, they lack unpaired electrons in the valence shell, rendering a paramagnetic analyzer useless [8]. Alternatively, techniques like IR absorption spectroscopy, Raman scattering, and Mass spectrometry have been found to be effective. While Raman scattering and Mass spectrometry deliver more accurate and rapid results compared to IR absorption spectroscopy, their high cost and the small, portable nature of IR analyzers make IR absorption spectroscopy the most attractive option for frequent use in a surgical theater [1].

The principal basis behind the IR absorption technique relies on the fact that a molecule containing two or more different atoms will absorb IR at specific frequencies. Utilizing this principle, the Beer-Lambert law, which states that light attenuation through a medium (in this case, the absorption of IR light by the gas) is proportional to the concentration of the light absorbers (gas) within the substance (mixture), can be used to measure the concentration of multiple gasses within an anesthetic circuit, as long as they meet the prerequisite conditions [9]. In an anesthetic circuit, an IR source is positioned behind a variable filter allowing for the transmission of a variety of IR frequencies. To take measurements, sample gas is introduced into a sampling chamber parallel to a reference chamber providing a baseline for concentration calculation. The design of these chambers is optimized to enhance absorbance by leveraging the increased path length, hence improving result accuracy.

IR emitted from the source passes through the filter, determining the IR wavelength. The sample gas absorbs a part of the emitted IR, while the remainder is detected by a photodetector at the end. Through processing, the concentration and partial pressure of the measured gas can be simultaneously displayed on a screen for the anesthetist to read [1]. Nonetheless, specific molecules such as CO₂, N₂O, water vapor, and alcohol may interfere with the absorption peaks of volatile agents due to overlapping absorption wavelengths. To overcome this issue, modern machines analyze multiple absorption peaks of the volatile agents, allowing for accurate measurements [6].

6. Conclusion

By taking advantage of the physical and chemical properties of gasses running through an anesthetic circuit, the precise technique needed to take exact measurements of gas concentration can be employed. Overcoming the limitations of each technique by proposing new ambitious solutions allows for the safe monitoring of a patient during an anesthetic intervention to be achieved.

7. References

- [1]: Duncan, A., & Pratt, O. W. (2021). Measurement of gas concentrations. *Anaesthesia & Intensive Care Medicine*, 22(3), 190-193.
- [2]: Pierry, A. T. (2009). U.S. Patent No. 7,556,039. Washington, DC: U.S. Patent and Trademark Office.
- [3]: BRADBROOK, C. (2013). Advanced patient monitoring during anesthesia: Part Two.
- [4]: Paramagnetic method. HiQ. (n.d.). https://hiq.linde-gas.com/en/analytical_methods/other_methods/paramagnetic_method.html, accessed on 11th of August, 2024
- [5] Manning, Andrew C., Ralph F. Keeling, and Jeffrey P. Severinghaus. "Precise atmospheric oxygen measurements with a paramagnetic oxygen analyzer." *Global Biogeochemical Cycles* 13.4 (1999): 1107-1115.
- [6]: Garg, R., & Gupta, R. C. (2013). Analysis of oxygen, anesthesia agent and flows in anesthesia machine. *Indian Journal of Anaesthesia*, 57(5), 481-488.
- [7]: 2015B13 describes the fuel cell and the paramagnetic oxygen analyser. discuss their use in anesthetic practice. *Anaesthesia Primary Exam*. (n.d.). https://ketaminenightmares.com/pex/saqs/other/equipment/2015B13_fuel_cell_and_paramagnetic_oxygen_analyser.htm, accessed on 11th of August, 2024
- [8]: Iwasaka, M. (2009, March). Effects of gradient magnetic fields on CO₂ sublimation in dry ice. In *Journal of Physics: Conference Series* (Vol. 156, No. 1, p. 012029). IOP Publishing.
- [9]: Narayan, R. (2018). *Encyclopedia of biomedical engineering*. Elsevier.



08

Computer Science

Overview of JPEG-1 and JPEG 2000 compression

ERIC ZIMING LU

1. Abstract

In this report, I will cover the mechanics behind JPEG-1 and JPEG 2000 compression and provide a brief comparison between the two algorithms at the end.

2. Introduction

During the 1970s and 1980s, computers struggled to load images. This is due to several reasons:

- Limited computation speed
- Limited storage

To mitigate these problems, experts in computer science have gathered and invented JPEG compression in 1986. As technology continues to thrive within consecutive years, computers became increasingly powerful. CPUs and GPUs were faster, RAM and cache size were increased, and HDDs and SSDs had more storage capacity. That means, the problems with loading speed and limited storage were diminished, and customers' expectations in the electronics sector will rise sharply. To keep pace with the times, the Joint Photographic Experts Group invented JPEG 2000 to replace JPEG-1. Not only JPEG 2000 compresses an image with higher compression ratio and quality, it also includes additional features and parameters that can be tuned to the user's needs. Unfortunately, JPEG 2000 adoption was extremely slow due to the popularity of JPEG-1, and hardware and software were specifically optimized for JPEG compression. The commercial sector did not bother to make their hardware and software compatible with JPEG 2000 given that JPEG-1 already does reasonably good job at compressing images and that the cost of switching to JPEG 2000 is huge. In this report, we'll cover the mechanics behind JPEG-1 and JPEG 2000 compression and provide a brief comparison between the two algorithms.

3. JPEG compression

JPEG compression involves the following steps shown in fig 1.

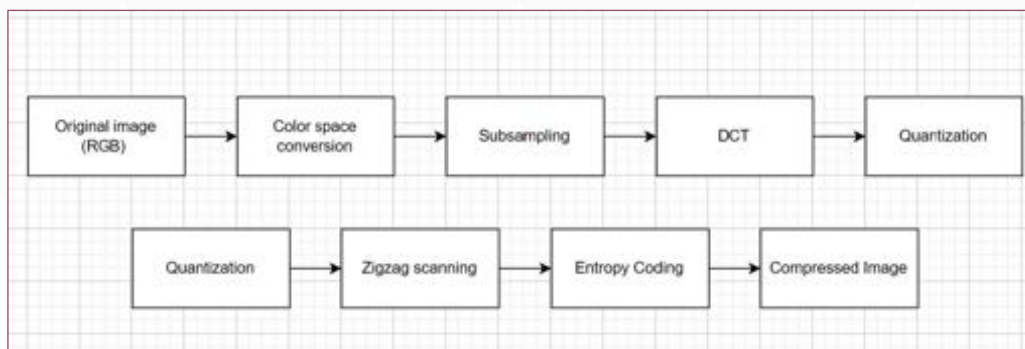


Fig 1: Overview of JPEG compression

3.1 Color-space conversion

To backup some knowledge required for understanding why we need to perform a color-space conversion: the human eye consists of rod and cone cells. Rod cells are responsible for scotopic vision (i.e. receives light), whilst cone cells are responsible for photopic vision (i.e. receives color). As humans, we have 20 times more rod cells than cone cells [1], allowing more luminance information to be processed by our brain. Because of this, we are very sensitive to changes in brightness than changes in color. This characteristic is something that we take advantage of when performing JPEG compression: we

can subtly remove some chrominance information without affecting the luminance information.

To remove chrominance information without our eyes noticing, the first thing to consider is that we need to isolate chrominance and luminance information. This is exactly what color-space conversion does.

Traditionally, an image consists of three channels: RGB. This means that each pixel within the image are represented by a vector of three values: (R,G,B), each ranging from 0 to 255. There are also other ways of representing an image with three channels, and the one that does the job of isolating luminance and chrominance is YCbCr. In this representation:

- Y indicates the brightness of a pixel
- Cb indicates the blueness of a pixel
- Cr indicates the redness of a pixel

The formula for transferring from RGB color space to YCbCr color space is given by

$$\begin{aligned}
 Y &= 0.299R + 0.587G + 0.114B \\
 Cb &= -0.168736R - 0.331264G + 0.5B + 128 \\
 Cr &= 0.5R - 0.418688G - 0.081312B + 128
 \end{aligned}$$

Fig 2. Formula for RGB -> YCbCr color space transform

The values of Y, Cb, and Cr are quantized to fit into the 8-bit (0-255) range, which results in round-off errors, thereby making the process irreversible.

3.2 Subsampling

Typically, we divide an image into blocks with 4x2 pixels during this process. To reduce chrominance information, we can average the values of nearby pixels depending on the J:a:b ratio:

- J refers to the number of pixels horizontally present in the pixel block
- a refers to the number of merged pixels in first column
- b refers to the number of changes in color value with reference to the first column of pixels.

To visualize this, take a look at fig 3.

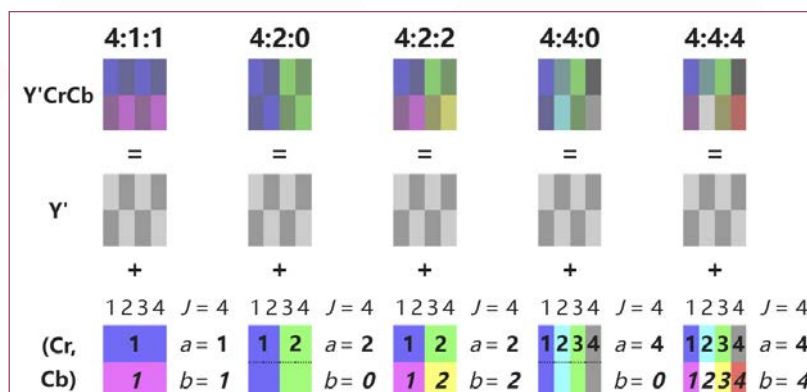


Fig 3. Color space conversion

By averaging the adjacent colors, we reduce the spatial resolution of chrominance information and thereby shrinking the file size.

3.3 Discrete Cosine Transform

We have reduced the image's file size by subsampling, but this won't result in a high compression ratio. We need to further the image by playing some visual tricks.

In 1.1, we talked about how human eyes are less sensitive to chrominance than luminance information. Another useful

characteristic of the human eye that we haven't talked about is that we are less sensitive to high spatial frequencies than low spatial frequencies. Here, we are referring to the change in color value of neighboring pixels. To illustrate what high and low frequencies might look like, take a look at fig 4 below.



Fig 4. This image includes huge amounts of high frequency data that could be potentially removed

In fig 4, the yellow boxed region consists of smooth, gradient color changes, which are low frequency information. In contrast, the red boxed region consists of complex shapes and edges, which indicates high frequency in that area. Our eyes might not notice if we delicately add some brown to the snow on the tree branches, proving that removing high frequency information makes little difference in our eyes. To benefit from this property, we need to think of a way of isolating high frequency and low frequency information.

If we look at a row of pixel, and transform it into the frequency domain, the graph of its frequency variation would look extremely complicated. Fortunately, we have mathematical tools at our disposal to effectively break these frequencies down into simpler waves, and DCT is one such tool.

3.3.1 1-dimensional DCT

The mathematical formula for 1D DCT is given below:

$$X[k] = \alpha(k) \sum_{n=0}^{N-1} x[n] \cos \left[\frac{\pi}{N} \left(n + \frac{1}{2} \right) k \right]$$

We have:

- $\alpha(k)$ being the normalization factor, defined as:

$$\alpha(k) = \begin{cases} \sqrt{\frac{1}{N}} & \text{if } k = 0 \\ \sqrt{\frac{2}{N}} & \text{if } k \neq 0 \end{cases}$$

- k being the index of the 8 basis functions of the transform
- $x[n]$ being the input signal
- N being the length of the signal (i.e. number of pixels in a row)
- $\cos \left[\frac{\pi}{N} \left(n + \frac{1}{2} \right) k \right]$ being the basis function, which can also be represented by the notation

$$\phi_k(x) = \cos \left(\frac{\pi(2x+1)k}{2N} \right)$$

Note that if you have uncovered discrete cosine transform in other areas of computer science, you might immediately notice that the normalization factor $\alpha(k)$ isn't included in traditional DCT. In JPEG compression, every step we take should be orthonormal such that we can perform IDCT correctly (i.e. reconstruct the image).

Looking at the formula, it is trivial to see that the number of different basis functions depends on N . The choice of N should be chosen carefully, as it can affect the computational complexity and image quality (we'll talk about this in section 1.1.3). What the formula is essentially doing is breaking down the signal $x[n]$ into N basis functions $\phi_k(x)$, where k ranges from 0 to $N-1$. As k increase, we see that the frequency of $\phi_k(x)$ would increase as well. We are able to construct the original frequency

$x[n]$ with relative quantities of each of these basis functions.

3.3.2 2-dimensional DCT

Clearly, an image is made up of rows as well as columns of pixels. We cannot ignore the columns, so we will have to perform DCT in the vertical direction as well, which extends 1D DCT to 2D DCT. The formula for 2D DCT isn't science rocket:

$$X[u, v] = \alpha(u)\alpha(v) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} x[x, y] \cos \left[\frac{\pi(2x+1)u}{2N} \right] \cos \left[\frac{\pi(2y+1)v}{2N} \right]$$

The formula employs another dimension and multiplies it by our previous formula. Notice that we have N at the denominator of both of the basis functions. This means that we apply 2D DCT only to square regions which brings up the concept of block splitting.

3.3.3 Block splitting

As said earlier, the dimensions of the blocks to be processed by DCT should be carefully chosen. This is a tradeoff between computational complexity and energy compaction. For example, choosing 4×4 as the region for each block would result in low energy compaction due to less frequency information being captured, therefore leading to more noises and decreased image quality. Conversely, if we choose 128×128 as the size of our blocks, we will consume a lot of computational resources. Remember that computer CPUs had little cache and memory in the 1990s when JPEG was invented. Processing such a big chunk of data at a time would have been computationally infeasible or time consuming. As a result, the Joint Photographic Experts Group have agreed upon standardizing the block size as 8×8 pixels for balancing the two factors.

3.3.4 Padding

We've decided to split an image into 8×8 pixel blocks. However, not all images have dimensions divisible by 8. The trick to get around with this is to pad extra pixels into the image. These pixels are typically assigned with values of $YCbCr = (0,0,0)$ (though other variations are possible), and are not differentiated from other pixels during the rest of the JPEG compression. These padded pixels are truncated during the decompression process, specifically after IDCT for displaying.

3.4 Quantization

Performing DCT upon an 8×8 block would return us an 8×8 matrix. Each number in the matrix indicates how much of its corresponding basis function is required to build up the frequency present in the original 8×8 pixel block.

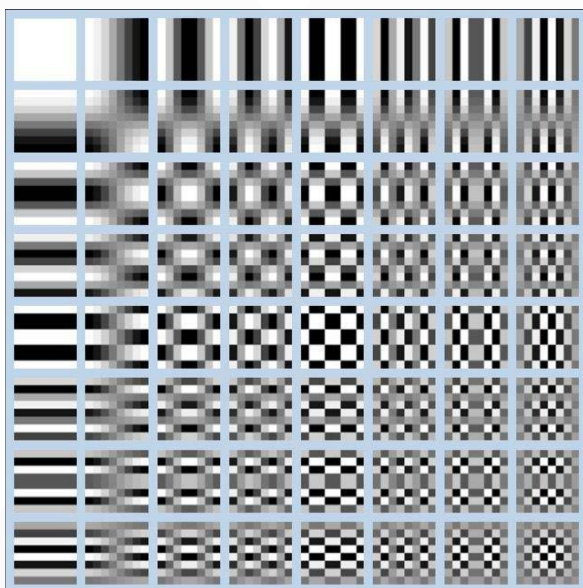


Fig 5. Each block is a visual representation of its basis function: The whiter an area is, the closer it is to the crest; the blacker an area is, the closer it is to the trough.

$$\begin{pmatrix} 340.2 & 23.4 & -12.1 & 6.5 & 2.3 & -1.7 & 3.9 & 1.2 \\ 13.8 & -5.2 & 7.1 & -4.3 & -2.6 & 1.9 & 0.4 & 2.8 \\ -9.7 & 3.2 & -1.3 & 0.8 & -2.4 & 1.6 & 0.3 & -1.1 \\ 5.4 & -2.9 & 1.2 & -3.6 & -1.4 & 0.7 & 1.0 & 2.3 \\ 3.1 & -1.5 & 0.9 & -2.1 & 1.5 & 0.2 & 0.1 & 1.6 \\ 1.8 & 0.6 & -1.0 & 1.4 & 0.3 & -1.2 & 1.7 & 0.5 \\ 0.9 & 1.1 & -1.4 & 0.7 & 1.9 & 0.8 & -1.3 & 0.2 \\ 1.3 & 0.4 & 0.7 & 1.1 & 0.9 & -1.5 & 0.6 & 1.1 \end{pmatrix}$$

Fig 6. An example of the coefficient matrix after DCT

The coefficients at the top left (representative of low frequency signals) of the matrix tends to be greater compared to the bottom right coefficients (representative of high frequency signals) of the matrix. Before abandoning these high frequency signals right away, we can further distinguish between high frequencies and low frequencies. This process is known as quantization.

two stages:

- Building the binary Huffman tree
- Encoding the data

To build a binary Huffman tree, we analyze the frequency of the occurrence of each number from the set of tuples we receive after RLE and arrange them into a priority queue sorted from the least to most frequent occurring numbers. Since the tree is binary, each node including the root would only consist of two child nodes. We can follow the below steps to obtain a binary Huffman tree:

- Extract two numbers from the queue and assign them as leaf nodes
- Sum up their corresponding frequencies. This will be the frequency of their parent node
- Insert the parent node back to the priority queue and repeat until only one node remain
- The remaining node will be the root of the tree

Now that we have built the tree, we can start assigning binary variable length codes to each node. A symbol will be assigned a shorter code the more frequently it occurs. We replace the output from RLE with Huffman codes to produce a bit stream. As a result, we have reduced the length of the bitstream, and data can be stored more effectively because we are not sending extra wasted bits for representing a symbol.

4. JPEG 2000

JPEG 2000 follows the same architecture of JPEG-1 with slight variations at each stage. The most notable change is that it uses discrete wavelet transform instead of discrete cosine transform.

4.1 Color-space conversion

JP2 is designed to be much more flexible than JPEG-1 as mentioned in the introduction. JP2 still supports the traditional RGB → YCbCr conversion while introducing the new RGB → YCgCo conversion as an alternative choice. Here, Cg indicates the greenness of an image, and Co indicates the orangeness of an image. To transform from RGB to YCgCo, we use the formula below:

$$\begin{aligned}
 Y &= \frac{R + G + B}{4} \\
 Cg &= \frac{R - B}{2} \\
 Co &= \frac{2G - R - B}{4}
 \end{aligned}$$

Notice that the formula is very clean. Division by 4 will at most introduce 2 decimal places, which can be easily represented without round-offs, allowing the conversion to be practically reversible.

4.2 Chroma subsampling

The process of subsampling chrominance information for both CbCr and CgCo layers are the same. The only difference compared to JPEG-1 at this stage is that chroma subsampling is skipped if we are on lossless mode because the averaging of pixels is irreversible.

4.3 Discrete wavelet transform

In JPEG-1, we split an image into 8x8 pixel blocks. However, in JP2, we are more flexible with the sizes of these blocks.

4.3.1 Tiling

Before performing a transform, we need to split the image into smaller pixel blocks just like JPEG-1. In JP2, these blocks are called tiles. We are free to parameterize the width and height of the tiles – we can even treat the whole image as one single tile. These tiles are uniform in size. By default, they have the size of 4096x4096 due to the increase in CPU cache size over past years which allowed more efficient processing of larger blocks of image data at a time.

4.3.2 sign shifting

After tiling, every pixel in a tile is shifted up a common value. The common value would typically be the half of the maximum possible value of the pixel range. This is to ensure the efficiency of calculating the Discrete wavelet transform.

4.3.2.1D Discrete wavelet transform

When we apply 1D DWT to a signal, we are essentially filtering the signal through a high pass and low pass filter. The high pass filter removes low frequencies which enhances the details (e.g. sharpens the edges) of the image, while the low pass filter removes high frequencies which blurs the image. This approach successfully isolates high and low frequency regions of a tile. Mathematically, 1D DWT can be expressed by:

$$\text{Approximation Coefficients: } A[n] = \sum_k x[k] \cdot h[2n - k]$$

$$\text{Detail Coefficients: } D[n] = \sum_k x[k] \cdot g[2n - k]$$

Where:

- $x[k]$ is the input signal
- $h[n]$ is the low pass filter
- $g[n]$ is the high pass filter
- $A[n]$ is the output coefficients of the low pass filter (so called the approximation coefficients)
- $D[n]$ is the output coefficients of the high pass filter (so called the detail coefficients) You may recognize that the formula for 1D DWT looks very similar to the formula for discrete convolution:

$$(f * g)[n] = \sum_{m=-\infty}^{\infty} f[m]g[n - m]$$

In fact, 1D DWT literally convolve the input signal $x[k]$ with a low or high pass filter. The 2 within the $g[2n-k]$ and $h[2n-k]$ allows down sampling with a factor of 2 to take place. As a result, we obtain half of the approximation coefficients and half of the detail coefficients, which in total adds to the correct number of signals that fits into a row of pixel in a tile. The actual formula for the high and low pass filters may vary depending on the type of wavelet transform we are using. Here, I will post the formula for H and L filters of lossless 5/3 discrete wavelet transform:

$$H(z) = -\frac{1}{8}(z^2 + z^{-2}) + \frac{1}{4}(z^1 + z^{-1}) + \frac{3}{4}$$

$$L(z) = -\frac{1}{2}(z^1 + z^{-1}) + 1$$

4.3.3 2D DWT

Similar to what we did in JPEG-1 during DCT, we extend DWT to 2 dimensions. We achieve this by performing DWT on the rows within a tile, then the columns. To help classifying high and low frequency regions, we can label the signals that passed through the low pass filter as L, and the signals that passed through the high pass filter as H. Together, we should obtain four quadrants: LL, LH, HL, and HH as shown in figure 10.

4.3.4 Recursive DWT

Subbands are the coefficients that are output after DWT. To further isolate low frequency and high frequency information, we should not consider the LH, HL, HH subbands because they already contain high frequency data.



fig 10. LL, LH, HL, and HH subbands

Instead, we can perform DWT locally within the LL subband. This process can be repeated multiple times depending on the desired level of decomposition.

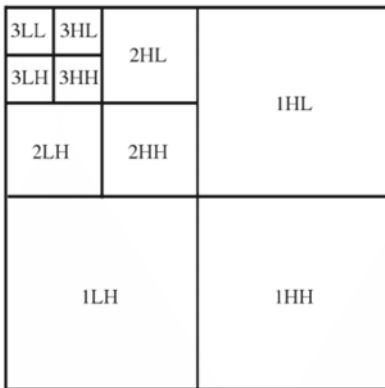


Fig 11. Recursive DWT of a within a tile

4.4 Quantization

Quantization in JPEG-1 is applied by dividing a block of coefficients to a quantization matrix. In jp2, we apply quantization locally to each subband at a time. Since we have varying sizes of subbands, we cannot predefine a quantization matrix. Instead, we have quantization parameters that are calculated from each individual subbands, and we use them to determine our quantization step size (the strength of quantization on a local subband). Additionally, JP2 introduces deadzone quantization. A deadzone is a subband with large amounts of coefficients with values near 0. Coefficients whose absolute values fall within deadzones are directly quantized to zero. The formula for calculating a quantized coefficient is shown below:

$$q_b = \text{sign}(y_b) \left\lfloor \frac{|y_b|}{\Delta_b} \right\rfloor$$

Where:

- q_b is the quantized coefficient
- y_b is a wavelet coefficient in subband b
- Δ_b is the quantization step

Obviously, we also have the lossless route, which would mean that no quantization happens such that we don't lose any data.

4.5 Entropy encoding

Entropy encoding is very different between JPEG-1 and JP2. JP2 implements more efficient encoding techniques, and the user is free to choose from which encoding techniques to use, allowing higher flexibility.

4.5.1 Codeblock decomposition

Subbands might be quite big in their dimensions at the current stage. We can break them into 64x64 or 32x32 blocks of pixels as these codeblock sizes have been proven to work well with different encoding algorithms. These are just sizes that generally strikes the best balance in trade-offs between encoding efficiency and computational complexity. Ultimately, the decision of the dimensions of codeblocks are decided and parameterized by the user. Nowadays our CPU have large caches that are capable of handling 64x64 code-blocks efficiently, and according to recent studies, we should always use 64x64 codeblocks for better image quality and compression time [2].

4.5.2 Encoding

Unlike JPEG-1, JP2 involves much advanced encoding mechanisms and are split into tier1 and 2, and we will only briefly cover the process of what typically happens during each stage. The compression algorithms can be changed or tailored to the user's specific needs.

Typically:

- We first apply bit-plane encoding to each quantized coefficient within a codeblock.
- We may then apply Embedded Block Coding with Optimal Truncation (EBCOT) which encodes the bit-plane encoded data in a way that supports multiple quality levels and resolutions. EBCOT is mandatory and is considered to be tier 1 encoding.
- At tier 2 encoding, the bit streams are rearranged and grouped in a way that improves spatial organization.
- After tire 2 encoding, we can start packaging the file into JP2 format.

5. Comparison between JPEG and JPEG 2000 in each stage

Stages	JPEG	JPEG 2000
Color space transformation	Only allows RGB -> YCbCr transformation	Allows both RGB -> YCbCr and RGB -> YCgCo transformation
Subsampling	Mandatory	Nonmandatory for lossless route
Block splitting/tiling	Image is split into 8x8 pixel blocks	Image is split into tiles with dimensions defined by the user.
DCT/DWT	Decomposes signal into cosine waves. This may be only applied once to each block	Decomposes signal into wavelets. This can be done recursively to LL subbands
Quantization	Uses standardized matrixes	Calculated in each subband taking lots of parameters into consideration
Entropy encoding	Simple and mandatory. RLE + Huffman	Complicated but flexible. The user decides what type of algorithm to use. Only ECBOT is mandatory.
File format	.jpg/jpeg	.jp2

6. Reference

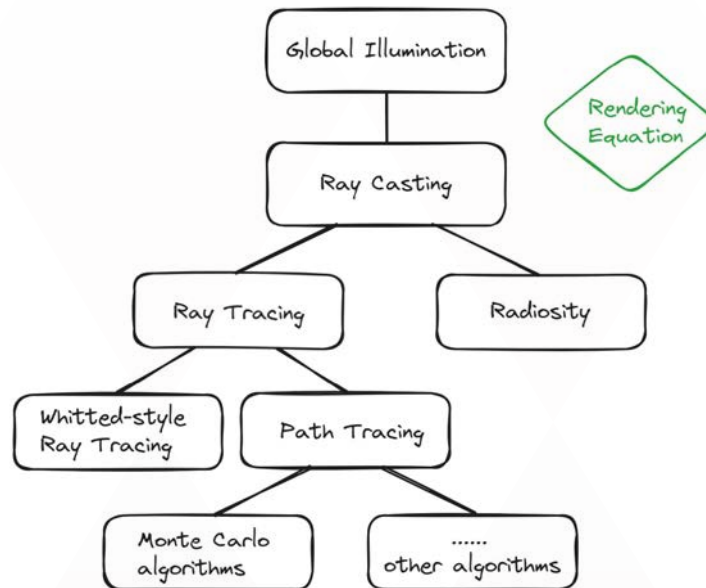
- [1] Khan Academy, [Online]. Available: <https://www.khanacademy.org/science/health-and-medicine/nervous-system-and-sensory-infor/sight-vision/v/photoreceptors-rods-cones> [Accessed: Aug. 09, 2024].
- [2] D. Barina and O. Klima, "JPEG 2000: Guide for Digital Libraries," Brno University of Technology, Faculty of Information Technology, Centre of Excellence IT4Innovations, Brno, Czech Republic, June 5, 2020. [Online]. Available: [(PDF) JPEG 2000: guide for digital libraries (researchgate.net)]. [Accessed: Aug. 09, 2024].

Comparison between Ray Tracing and Radiosity

CHANG LIU

1. Introduction

Computer graphics is a large topic, mainly focused on researching how to use computers to simulate 3D scenes. Since the concept was proposed, computer graphics scientists have been constantly striving to make the generated models more realistic. In real-life simulations, one of the most important thing is lighting and shadowing. Therefore, many models for simulating light have appeared, known as Global Illumination Models, which are mainly implemented with Ray Casting.



Pic 1-1

Most illumination models are based on Bidirectional Reflectance Distribution Function (BRDF) and Rendering Equation, which will both be discussed later in this essay.

The two well-known illumination models are Ray Tracing and Radiosity, which are the two models that will be compared later in this essay, among which Ray Tracing also has many different types, such as Whitted-style Ray Tracing and Path Tracing.

2. Ray Tracing

2.1 Definition

Ray tracing, a method for calculating the path of waves or particles through a system, which is used for 3D image generation, implements light simulation by recursion.

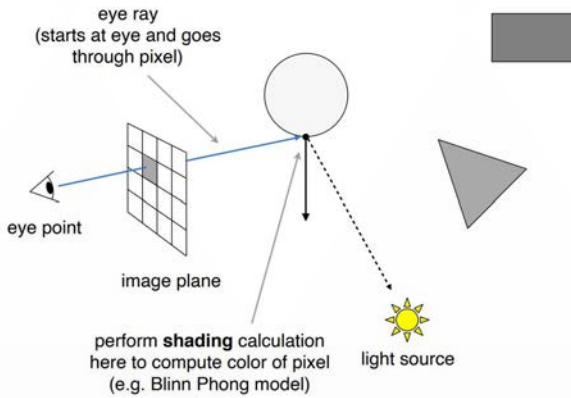
2.2 Whitted-style Ray Tracing

2.3 Algorithm

In Whitted-style Ray Tracing, all objects will be split in to 3 types, reflective, both reflective and refractive, diffuse and glossy. Due to the reversibility of optical path, if there is a light path from the light source to the viewer's eye, calculating a reversed ray from the viewer's eye to the light source will be equivalent.

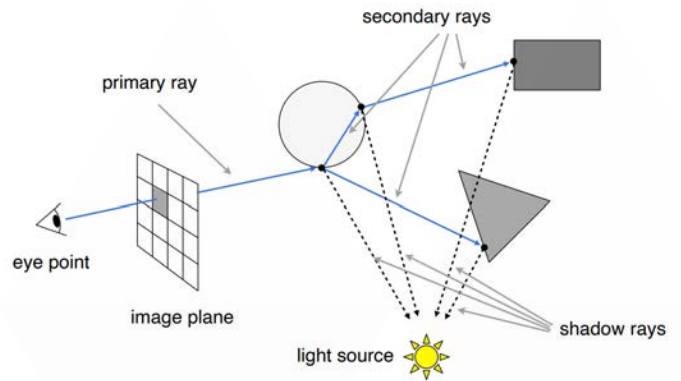
Therefore, instead of simulating the real light path and wasting calculation that is unnecessary from certain viewing point,

Ray Tracing calculates the reversed ray. We draw rays, starting from the eye point, crossing the centre of every pixel on the screen, and calculate the colour of the pixel.



Pic 2-2-1 From GAMES101, Lingqi Yan, UC Santa Barbara

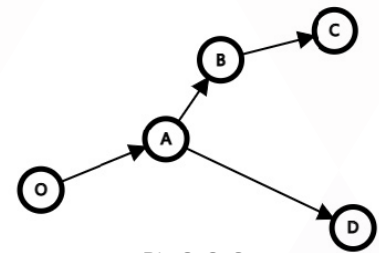
If there are no object that is both reflective and refractive, like the example in pic 2-2-1, the ray starts from the eye point, (get reflected by reflective surfaces several times), and ends at either a light source or a diffuse and glossy object, the colour of the pixel will be the colour of the point where the ray ends (in pic 2-2-1's case, point A).



Pic 2-2-2 From GAMES101, Lingqi Yan, UC Santa Barbara

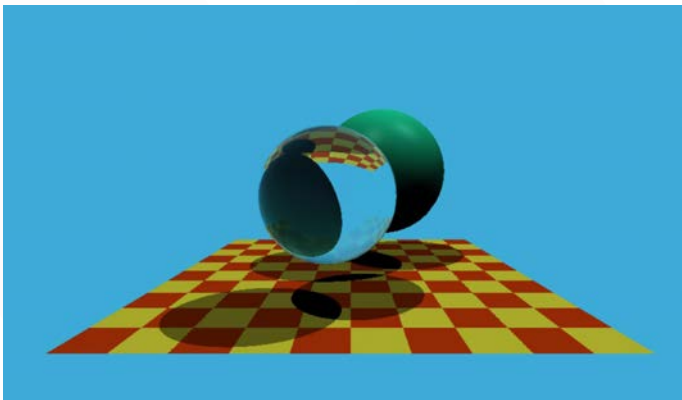
When there are objects that are both reflective and refractive, like the example in pic 2-2-2, the ray branches when it meets such surfaces (point A), then the two rays meet their ending surface separately (point C and point D). The final colour of this pixel will be the sum of the colours of all the ending points.

Therefore, the whole scene can be simplified as a directed graph (pic 2-2-3) and calculated by recursive algorithms.



Pic 2-2-3

2.4 Implementation



This is my implementation of Whitted-style Ray Tracing for a scene with two light sources and three objects:

- a green diffuse reflective and glossy sphere
- a transparent glass sphere
- a board with grids that is divided into two triangles

This implementation of the algorithm was carried out by the author,

- using C++ in Linux
- pure software programmed using CPU instead of GPU

2.5 Radiometry

Symbol	Name	Unit	Unit symbol
ϕ	Radiant flux	Watt; lumen	W; lm
I	Radiant intensity	Watt per solid radian; candela	W/sr; cd
L	Radiance	Watt per solid radian per square meter; nit	W/(sr·m ²)
E	Irradiance	Watt per square meter	W/m ²

The problem now comes to how to calculate the colours of the ending points of rays, which are from diffuse reflective surfaces.

Every rough object has its texture and colour mapping, so to calculate its real colour in the scene, we need to calculate the irradiance at its position E_p according to the radiant flux of the light source(s) Φ and distance between the object and the light source(s) r .

$$E_p = \frac{\Phi}{4\pi r^2}$$

2.6 Rendering Equation

However, Whitted-style Ray Tracing algorithm is physically incorrect, hence it will generate images that does not look real. The defect of this algorithm is that, in the real world, there is not only the direct illumination from the light source(s), but also the indirect illumination from diffuse reflection from other rough objects. The real situation can be described as the rendering equation below:

$$L_o(p, \omega_o) = L_e(p, \omega_o) + \int_{\omega_i} L_i(p, \omega_i) f_r(p, \omega_i, \omega_o) (n \cdot \omega_i) d\omega_i$$

Where

- $L_o(p, \omega_o)$ stands for total output radiance from position p to exit vector ω_o .
- $L_e(p, \omega_o)$ stands for emission radiance (the light emitted by the object itself) from position p to exit vector ω_o .
- $L_i(p, \omega_i)$ stands for total input (incident) radiance from incident vector ω_i to position p .
- $f_r(p, \omega_i, \omega_o)$ stands for the Bidirectional Reflectance Distribution Function (BRDF), which represents how much light is reflected into the outgoing direction ω_o from the incoming direction ω_i .
- n stands for the normalized normal vector, ω_i stands for the normalized incident vector, hence $(n \cdot \omega_i)$ stands for the cosine of the incident angle.

Since the input radiance at any angle must also be the total output radiance from the direction $-\omega_i$, the rendering equation can be optimized with one less variable:

$$L_o(p, \omega_o) = L_e(p, \omega_o) + \int_{\omega_i} L_o(p', -\omega_i) f_r(p, \omega_i, \omega_o) (n \cdot \omega_i) d\omega_i$$

Where p' stands for the position that has an output radiance in the direction $-\omega_i$

2.7 Path Tracing

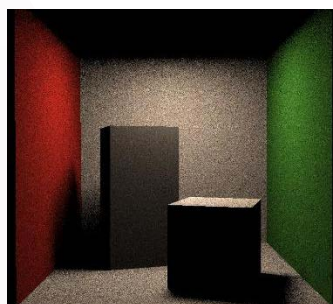
Path Tracing Algorithm is an upgraded version that combines Whitted-style Ray Tracing Algorithm with the Rendering Equation. On the basis of Whitted-style Ray Tracing, in Path Tracing, the ray does not end at the first rough surface, it goes on spreading to almost every direction due to diffuse reflection from other rough surfaces, hence the colour of the pixel is calculated using the Rendering Equation.

It is worth noting that L_o appears in both sides of the Rendering Equation, so in order to calculate all the L_o in the scene, we have to solve a system of linear equations using matrices. However, if we calculate all the possible directions that the ray can branch to, there will be millions of equations, which will slow down the algorithm to an unacceptable extent. Therefore, it is an approximation of the Rendering Equation, using the Monte Carlo Method. Instead of traversing all the diffusing rays spreading out, the Monte Carlo Method chooses a few of them according to their possibility to be seen, and returns the weighted average.

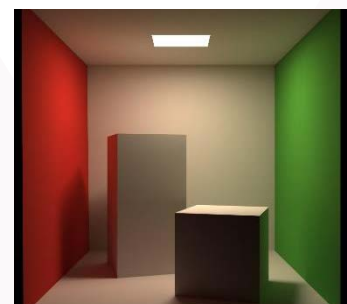
2.8 The Cornell Box



Pic 2-6-1 Real Photo



Pic 2-6-2 Whitted-style Ray Tracing



Pic 2-6-3 Path Tracing (Monte Carlo)

The Cornell Box is a famous scene for Global Illumination.

- Pic 2-6-1 is the real photo taken (in this model the red cuboid has a specular reflective surface).
- Pic 2-6-2 is the simulation done by Whitted-style Ray Tracing (this scene is slightly different with the one in Pic 2-6-1, it is not that the defects of this algorithm that caused this problem).
- Pic 2-6-3 is the simulation done by Path Tracing (this scene is the same as the one in Pic 2-6-2).

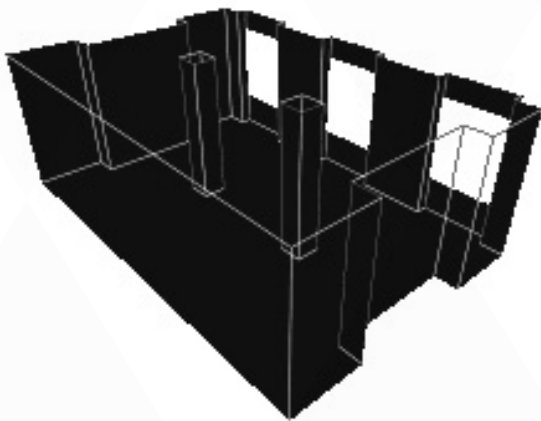
3. Radiosity

3.1 Definition

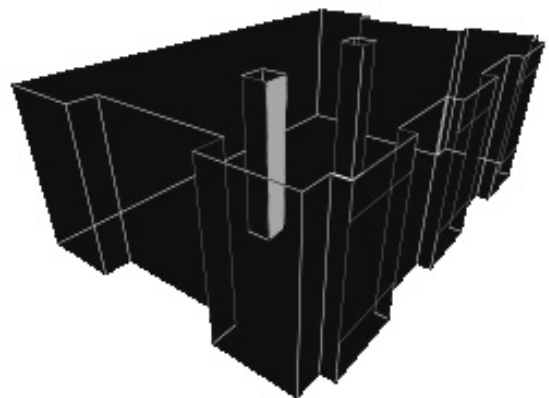
Radiosity is a rendering algorithm which gives a realistic rendering of shadows and diffuse light. It is an application of the finite element method to solving the rendering equation for scenes with surfaces that reflect light diffusely.

3.2 Algorithm

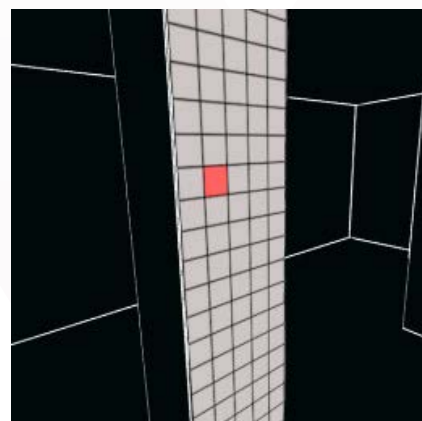
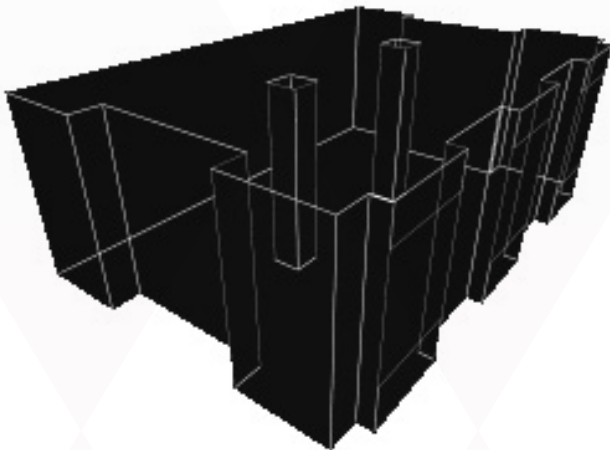
The surfaces of every object in the whole scene are divided up into many small patches (usually triangles since any polygon can be represented by a set of triangles, or pixels since it can be easily represented using coordinates), which are calculated separately. Due to the limited quantity of patches, Radiosity is a finite element method.



Pic 3-2-1



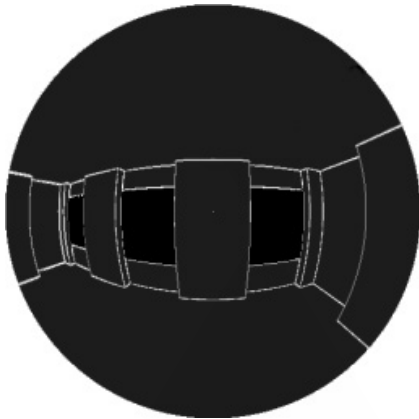
Pic 3-2-2



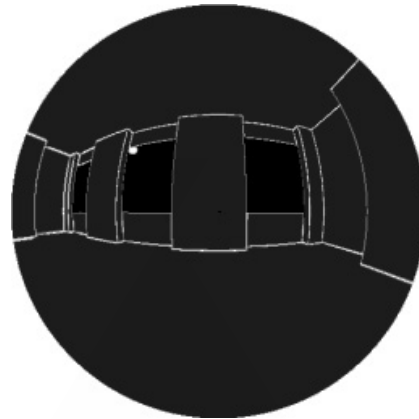
Pic 3-2-3

From Hugo-Elias-Radiosity

For example, this scene (pic 3-2-1) is a room with three windows, two pillars and some alcoves, and there is a light source outside the windows. Locally displayed, one of the surfaces of a pillar (the highlighted one in pic 3-2-2) is divided up into pixels (pic 3-2-3).



Pic 3-2-4

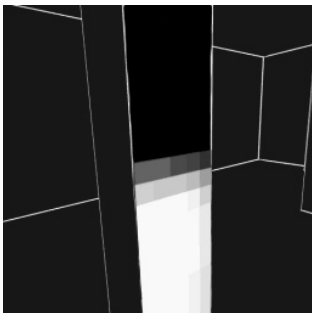


Pic 3-2-5

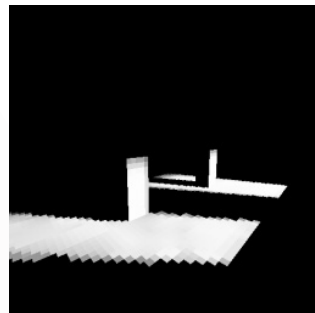
From Hugo-Elias-Radiosity

Viewing from each patch, by adding together all the light it sees, we can calculate the total amount of light from the scene reaching the patch.

Pic 3-2-4 is the vision from the red pixel on the pillar (pic 3-2-3), and Pic 3-2-5 is the vision from a lower pixel on the same pillar. In the vision of the red pixel now, the world is completely dark, while the lower pixel can see a part of the light source through the window. Therefore, the colour of the red pixel is black, while the lower pixel is grey or white, which is a little brighter than the red pixel (pic 3-2-6).



Pic 3-2-6



Pic 3-2-7

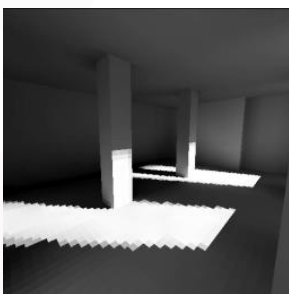
Entire Room Lit: 1st Pass



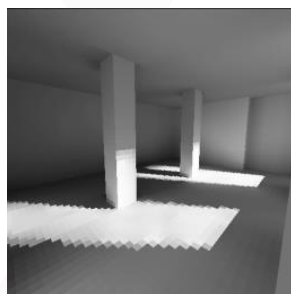
Pic 3-2-8

View from the red pixel after 1st Pass

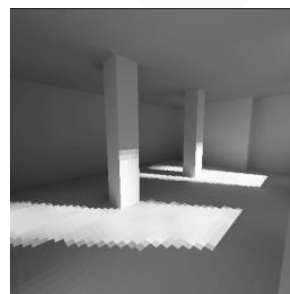
Having repeated this process for all the patches, the room is lit with direct light from the light source (pic 3-2-7). Pic 3-2-8 is the vision from the red pixel after the first pass, now its world is brighter due to the diffuse reflection of the floor and the wall, and now the light source becomes a set that includes the original light source(s) and the bright parts of the floor and the wall.



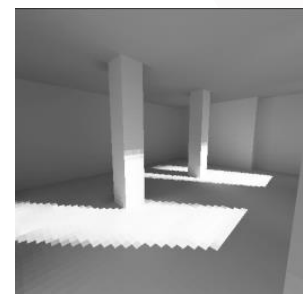
Pic 3-2-9

 Entire Room Lit: 2nd Pass


Pic 3-2-10

 Entire Room Lit: 3rd Pass


Pic 3-2-11

 Entire Room Lit: 4th Pass


Pic 3-2-12

 Entire Room Lit: 16th Pass

From Hugo-Elias-Radiosity

By repeating the above operations, every time the additional light in the room becomes a new part of light source due to their diffuse reflection. With more and brighter light sources, the room becomes brighter and more realistic every time as simulation continues.

4. Comparison between Ray Tracing (Whitted-style) and Radiosity

4.1 Similarities and Differences

The two global illumination algorithms have many differences

		Ray Tracing		Radiosity
		Whitted-style Ray Tracing	Path Tracing	
Similarities	Purpose	Global Illumination		
Differences	Specular reflection	Solved		Not solved
	Diffuse reflection	Partially solved	Approximately solved	Solved
	Refraction	Solved		Not solved
	Shadows	Sharp edges, looks contrived		Soft edges, looks realistic
	Algorithm	Simulates the reversed ray from the camera to the light source(s)		Simulates the light from the light source(s) to all surfaces of all objects
	Vision	Only from the current camera vision		Can be observed from any direction

- The first difference is rendering effect. Whitted-style Ray Tracing can show specular reflection, refraction and direct diffuse reflection that bounces only once. Path Tracing can show specular reflection, refraction and approximate diffuse reflection that bounces only once. Radiosity can only show Diffuse reflection, but much more precise than Path Tracing.
- The second difference is the starting point. Ray tracing follows all rays from the eye of the viewer back to the light sources. Radiosity simulates the diffuse propagation of light starting at the light sources.
- The third difference is the vision. Since Ray tracing follows the rays from the eye to calculate the specular reflections, the result generated depends on the position of the eye. So Ray tracing is not usually used to do large amount of work such as animation films, because every frame needs to be calculated completely separately. While Radiosity simulates the diffuse reflections of every patch that are independent from the position of the eye, so we only have to run the whole algorithm once and can then easily calculate the image from any view using rasterization (an algorithm that can project a 3D scene onto a 2D screen).

4.2 Combination

From the comparison above, we can see that Ray Tracing and Radiosity both have their own advantages and disadvantages. So the effect will be better if the two are combined together. Pic 4-2-1 is generated by Whitted-style Ray Tracing, the image shows specular reflections of the floor, but all the shadows have sharp edges, and the ceiling and the upper part of the wall is completely dark.

Pic 4-2-2 is generated by Radiosity, the image shows no specular reflection, but the shadows have realistic soft edges, and the ceiling and the upper part of the wall is lit.

Pic 4-2-3 is generated by the combined algorithm, the image shows the floor which has both specular reflection and diffuse reflection, the shadows have realistic soft edges, and there are nice indirect illumination.



Pic 4-2-1
Whitted-style Ray Tracing



Pic 4-2-2
Radiosity From Radiosity - Ray Tracing



Pic 4-2-3
Combination

5. Reference

- [1]. 全局光照模型 (Global Illumination Model) 概览、绘制方程 (Rendering Equation) from Zhihu
- [2]. Ray Tracing from Wikipeda
- [3]. Radiosity from Wikipeda
- [4]. GAMES101, Lingqi Yan, UC Santa Barbara from Bilibili
- [5]. Hugo-Elias-Radiosity from Github
- [6]. 光的能量与颜色——辐射度量学, 光度学, 色度学 from Zhihu
- [7]. Radiosity - Ray Tracing



FitzEd

EDUCATIONAL PROGRAMMES
FITZWILLIAM COLLEGE, CAMBRIDGE



FitzEd
EDUCATIONAL PROGRAMMES
FITZWILLIAM COLLEGE, CAMBRIDGE



60

Palaeobiology: Evolution and Behaviour

Mass-speciation and extinction events in Felidae

WANZHANG

1. Introduction

Felidae, a family originating approximately 30 million years ago and persisting to this day, has been profoundly influenced by several factors. Mass speciation events—rapid formations of numerous new species within short geological periods—and mass extinction events—widespread and rapid die-offs of a significant proportion of species globally—have played pivotal roles in felid evolution. Climate is widely considered a major driving factor behind these events.

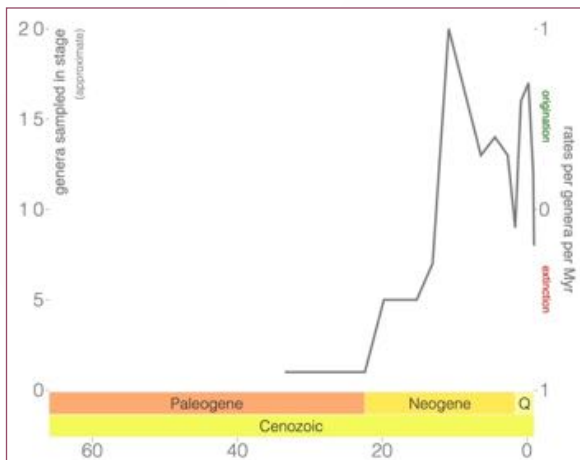


Figure 1.1 The general trend of the development of the Felidae family from the late Paleogene period till now

In this essay, general pattern of the development of Felidae will be summarized, and the potential relationship between climate and mass speciation and extinction events of Felidae, as well as its relationship with other clades, will be examined and discussed. It is hypothesized that there is an obvious relationship between the climate and the development of the Felidae.

It is impossible to clearly describe the development of Felidae in a few sentences, so a general chart (fig 1.1) is presented as a brief summary. Through the chart, five main turning points are clearly shown, and they are respectively in: (1) the late Paleogene period (about 22 MA), (2) the middle Neogene period (about 10 MA), (3) the late Neogene period (about 6 MA and 4 MA), (4) the early Quaternary period (about 2 MA), and (5) the late Quaternary period (about 0.5 MA).

2. Method

In this essay, The Paleobiology Database is used for creating a map of the Felidae habitat locations during each time period. Paleocharts is also used for plotting the relationships between the change in environmental factors and the change in Felidae and other clades.

3. Results

Climate aspect, temperature

In the chart (fig 3.1), there is a weak negative relationship between the temperature and the Felidae. The most obvious time period was between 15 MA and 9.5 MA, which was in the Neogene period. Although the expansion of the Felidae started from the late Paleogene period, it was not until the early Neogene period that the number had begun to rise rapidly which was also the time that the world temperature started to fall.

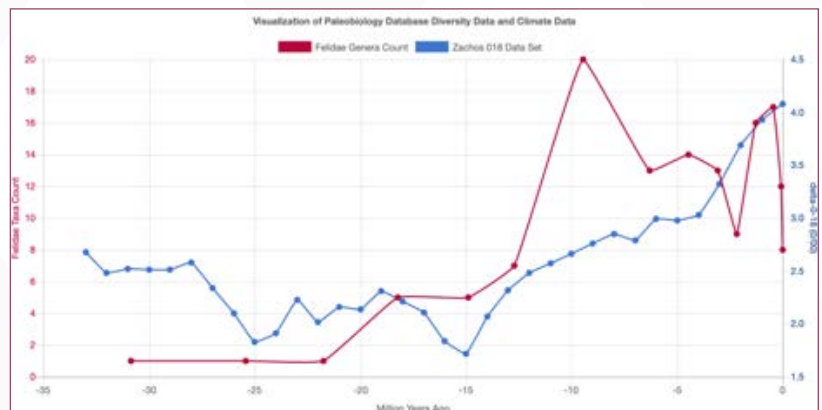


Figure 3.1 The general relationship of the temperature and the Felidae family from the Rupelian period (33.9mya) till now

The spread out of the Felidae can also show the weak negative relationship between the temperature and the Felidae family. During the Oligocene period, the Felidae's habitats were in North America, Asia, Europe, but in the Pliocene period, Felidae's habitats could also be found in Africa and South America (fig 3.2) The graph in fig 3.1 also shows that the general trend of the temperature since the Oligocene period to the Pliocene period is a decrease. Therefore, the cool temperature is beneficial for the mass speciation of the Felidae.



Figure 3.2 The Felidae's distribution in the Oligocene period (left) and the Pliocene period (right)

The fall of temperature may be due to a combination of both short-term factors, such as changes in atmospheric greenhouse gas concentrations, Antarctic ice form, and long term factors, including tectonic changes. During this period, the atmospheric carbon dioxide levels were lowering, which trapped less heat in the Earth's atmosphere, driving the overall cooling trend. A increase in ice sheet coverage could also allow more solar radiation to be reflected to space. Moreover, the tectonic changes, such as the closing of the Tethys sea between Africa and Eurasia, altered the global ocean circulation patterns. [1, 2, 3]

The fall in global and regional temperature created new habitats and ecological niches that allowed Felidae to expand their range and colonize new areas. The distribution of Felidae's prey species is also influenced by the rising temperature, creating new foraging opportunities for the Felidae.

Climate aspect, sea level

From the chart (fig 3.3), a weak relationship between the sea level and the Felidae is shown, and the most obvious time period was between 14 MA and 10.75 MA, which is nearly the same as the temperature fall.

The fall of the sea level enabled the Felidae to reach more resources and even new continents, speeding up the process of the mass-speciation of the Felidae family.

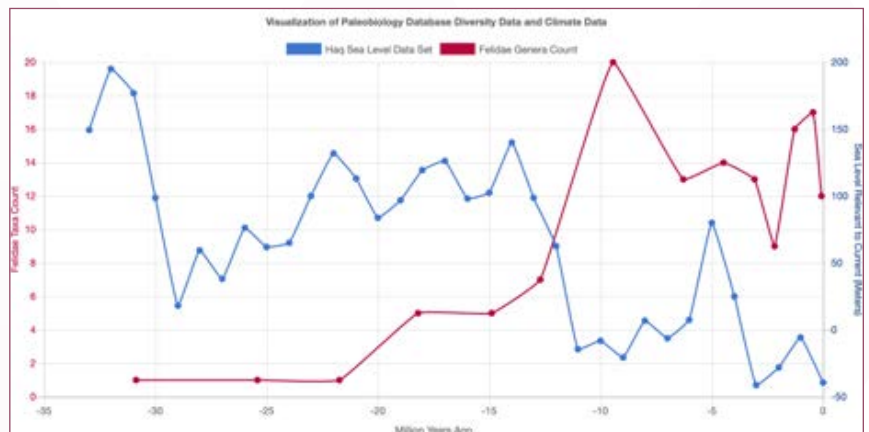


Figure 3.3 The general relationship of the sea level and the Felidae family from the Rupelian period (33.9mya) till now

Geographical aspect, tectonic movements

From fig 3.2, the movements of the plates during the time between the Oligocene period and the Pliocene period can also be shown. During these years, the North and the South America continents moved closer together, and so did the Asia, Europe and Africa continents. This change made it easier for the spread of the Felidae as it was easier for them to reach new continents, and new habitat and resources could also be provided. Therefore, tectonic movements can also be a reason for the mass speciation of the Felidae family. [4]

A single Felidae is also closely related to other clades, including their competitor and predator: the Canidae and the Ursidae,

as well as their prey: the Leporidae, the Cervidae, and the Bovidae. The two charts below (fig 3.4 and fig 3.5) show the changes in the number of these different clades.

However, these two graphs clearly show that the trend of the different clades' changes, include the Felidae family, is generally the same and happens at the same time, and there is little causality. Therefore, the threat from other clades cannot be classified as one of the changing

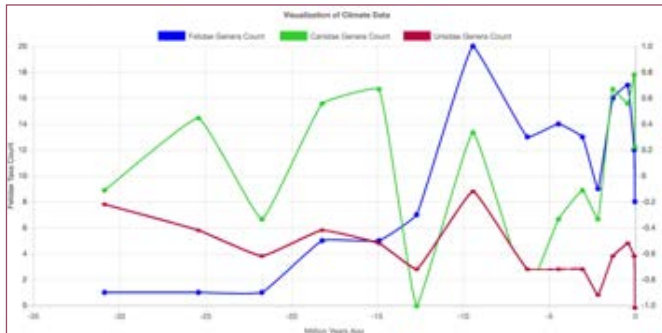


Figure 3.4 (left) The changes in the number of Felidae and its competitor and predator clades

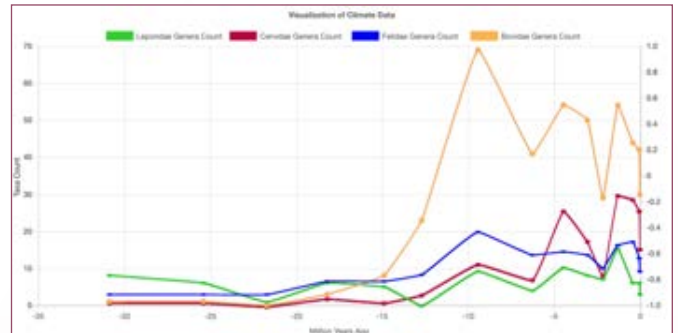


Figure 3.5 (right) The changes in the number of Felidae and its prey clades

4. Discussion

All the results above shows that climate changes, changes in temperature and sea level especially, as well as tectonic movements can all be the factor of the development of the Felidae family. However, the relationships between all three factors and the mass-speciation of the Felidaees are really limited: only a few upward turning points are explained and no downwards turning points can be explained.

The limited relationship may be due to a wide-spread of the Felidae family, so that they are less likely to be deeply affected by a single climate event in a single place. Its strong ability to adapt enable the Felidaees to exploit different food sources, navigate different terrains, and occupy different ecological niches, which may also be the reason why it won't be affected by climate change [5]. There can also be a possibility that the speciation and extinction is affected by hominin behaviors, such as hunting or deforestation. However, the result can also be affected by the preservation bias due to a lack of proper data, so further investigations are required.

5. References

- [1] Mudelsee, M., T. Bickert, C. H. Lear, and G. Lohmann (2014), Cenozoic climate changes: A review based on time series analysis of marine benthic $\delta^{18}O$ records, *Rev. Geophys.*, 52, 333–374, doi:10.1002/2013RG000440.
- [2] Anagnostou E, John EH, Edgar KM, Foster GL, Ridgwell A et al. 2016. Changing atmospheric CO2 concentration was the primary driver of early Cenozoic climate. *Nature* 533:7603380–84
- [3] Shiling Yang, Yongda Wang, Xiaofang Huang, Minmin Sun, Jingtai Han, Xu Wang, Zuoling Chen, Shihao Zhang, Wenying Jiang, Zihua Tang, Zhaoyan Gu, Shangfa Xiong, Zhongli Ding, Pliocene CO2 rise due to sea-level fall as a mechanism for the delayed ice age, *Global and Planetary Change*, 2024, 104431, ISSN 0921-8181
- [4] Janis, C. M. (1993). Tertiary Mammal Evolution in the Context of Changing Climates, Vegetation, and Tectonic Events. *Annual Review of Ecology and Systematics*, 24, 467–500.
- [5] Luo, ZX. Transformation and diversification in early mammal evolution. *Nature* 450, 1011–1019 (2007). <https://doi.org/10.1038/nature06277>

The user variability in muscle creation of the shoulder musculature of Coelophysis

MEIMEI XIE

1. Introduction

This study aims to primitively evaluate the reliability of musculoskeletal models by building the shoulder musculature of Coelophysis (a genus of coelophysid theropod dinosaur that lived approximately 215 to 208.5 million years ago during the Late Triassic period from the middle to late Norian age in what is now the southwestern United States) three times.

Creating musculoskeletal model is one of the most important ways to analyze fossil movement. Musculoskeletal modelling involves the creation of a model that attaches muscles and/or other soft tissues for simulation and hypothesis testing of locomotor function, behavior and performance (Bishop et al., 2020). However, as the number of studies using musculoskeletal model increases, the reliability of the model in question can become debatable and thus their reliability in providing replicable outcomes should be fully tested (Demuth et al., 2023). Though there are some stratagems (Hicks et al., 2015) that can be used to verify the validation of musculoskeletal models, the complexity of the 3D morphology and 3D environments still caused some models biologically inaccurate.

Here, the shoulder musculature of Coelophysis was used as a case study to determine whether the inaccuracy of musculoskeletal model is considerable when estimating fossil's moment arm. The muscles were modelled simply as straight lines of action. This hypothetical muscle that connected the scapula to the humerus was created by the same user three times. By recreating the same muscle three times, we can deduce whether slight differences in modeling can cause significant discrepancies of moment arms.

2. Methods

This study used the OpenSim Creator (provide URL/DOI here) to generate the hypothetical shoulder musculature of Coelophysis. A skeletal 3D model of Coelophysis was inserted to the program. An overall view of the dinosaur's skeleton is shown in figure 1.

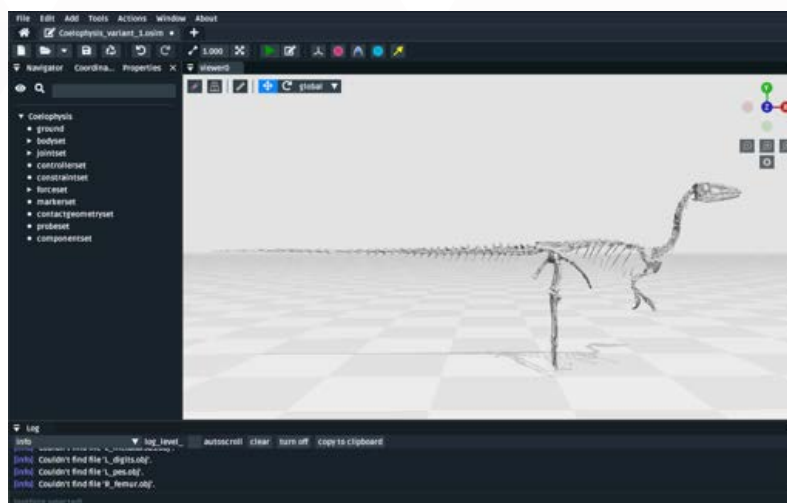


Figure 1. An overview of the Coelophysis skeletal structure in OpemSim Creator

Click “Add” - “Millard2012EquilibriumMuscle” on the right corner to create a new muscle. A window shown in figure 2 where we can set up the name, the force, the path points and any other properties of the muscle will pop up. Here we only need to edit the name and choose 2 path points “body” and “R-arm” (right arm) of the muscle. The new muscle is shown on the

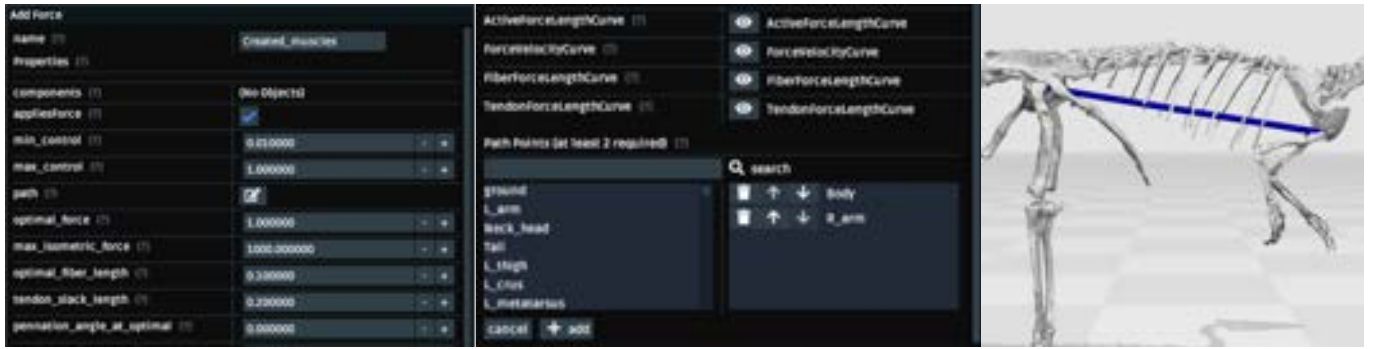


Figure 2. the window of editing new muscle and the created muscle (blue)

When we click on the created muscle, the muscle on the viewer will be highlighted in orange, and the muscle in the navigator will turn yellow (figure 3). Click on the path points to change the attaching points of the muscle (shown in the red boxes in figure 3).

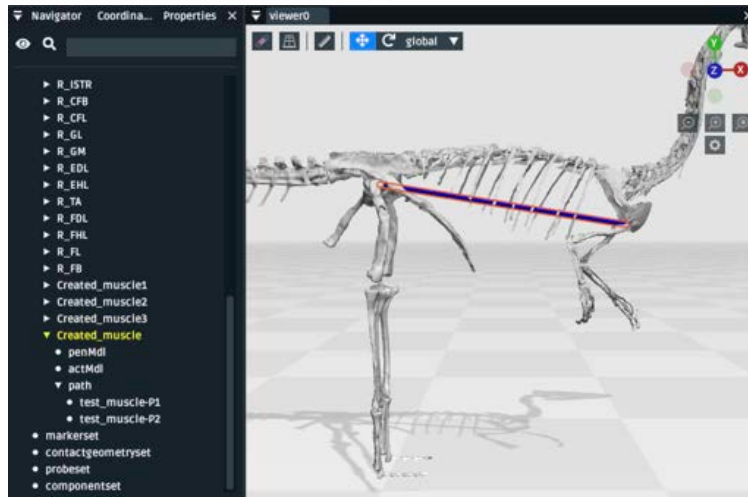


Figure 3. primary position of the new muscle

As our aim is to create a shoulder muscle, we have to attach the two ends on scapula and humeri. However, due to the limitations of the program, we have to create each via points of the muscle from shoulder to arm, namely that we have to put one point on the shoulder and the other point on the turning point of the shoulder to ensure the muscle will not pass through the bone when doing movements. To achieve this, we can first change the socket of the second point from body to right arm by right clicking "test_muscle-P2", and click "sockets" - "Action change". Choose "Body" on the popped-up window.

The starting point on the shoulder and the first turning point rotating with the shoulder is shown in figure 4.

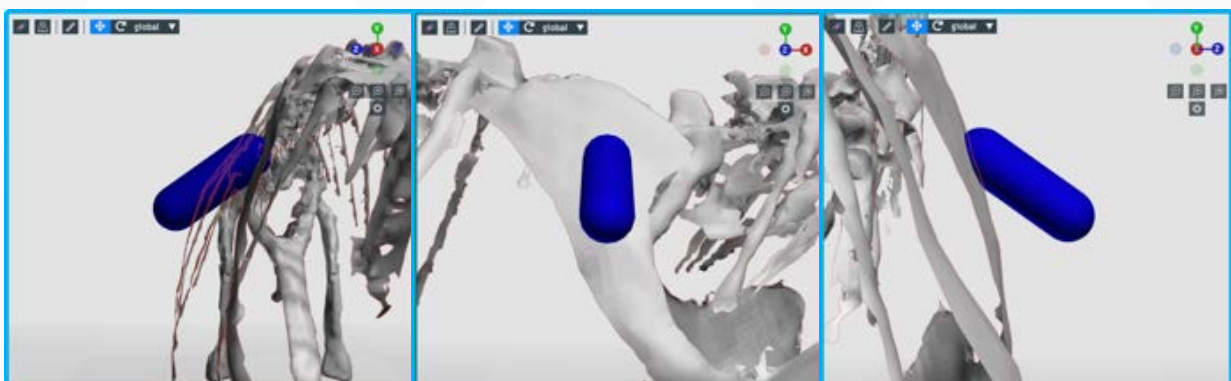


Figure 4. the first two point of the muscle

To create the second turning point rotating with the arm, right click “Created muscle”, and click “Add path point”. Click “R-arm” on the popped-up window. Adjust the position of that point same as before. Create the final point on the humeri using the same method previously.

In figure 5, a completed shoulder musculature module is displayed



Figure 5. the completed shoulder muscle structure

Hide the muscle and repeat the above procedures two times. Ultimately, three same muscles are generated shown in figure 6.

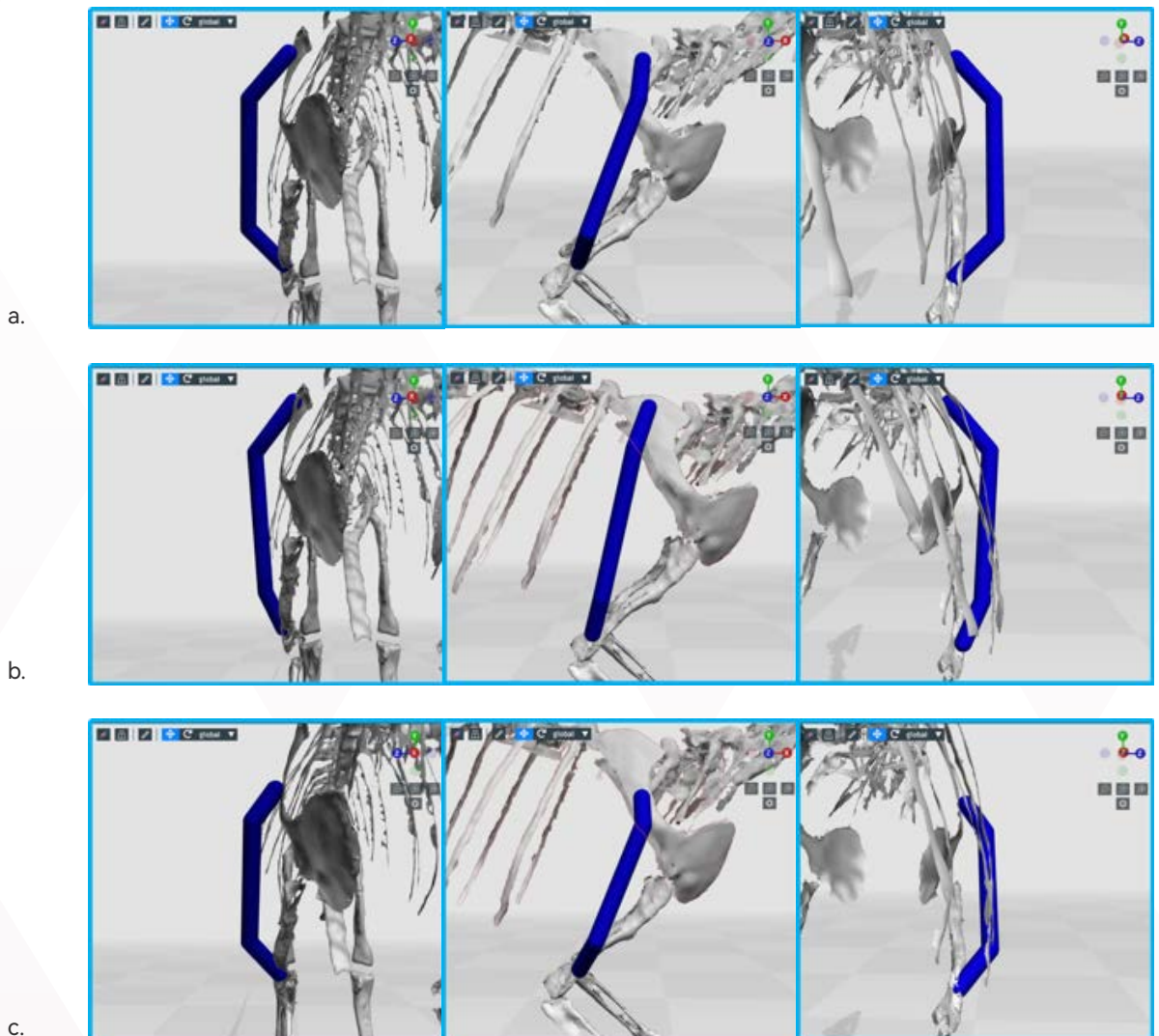


Figure 6. the three muscles: (a) muscle 1, (b) muscle 2, (c) muscle 3

After that, plot the graph of how the moment arm changes as the angle of the joint changes. Moment arm is the perpendicular distance between the point where the muscle force is applied and the joint. A greater moment arm means it is easier for the body to move in that angle. Thus, scientists often use moment arm to find how the extinct animal move their body and their locomotion. To see the differences of the moment arms, we export the extension value–moment arm graph and combined them into one chart shown in figure 7.

3. Results and discussion

In general, the moment arm trends are roughly the same. However, we have to be more critical toward the accuracy of the moment arms.

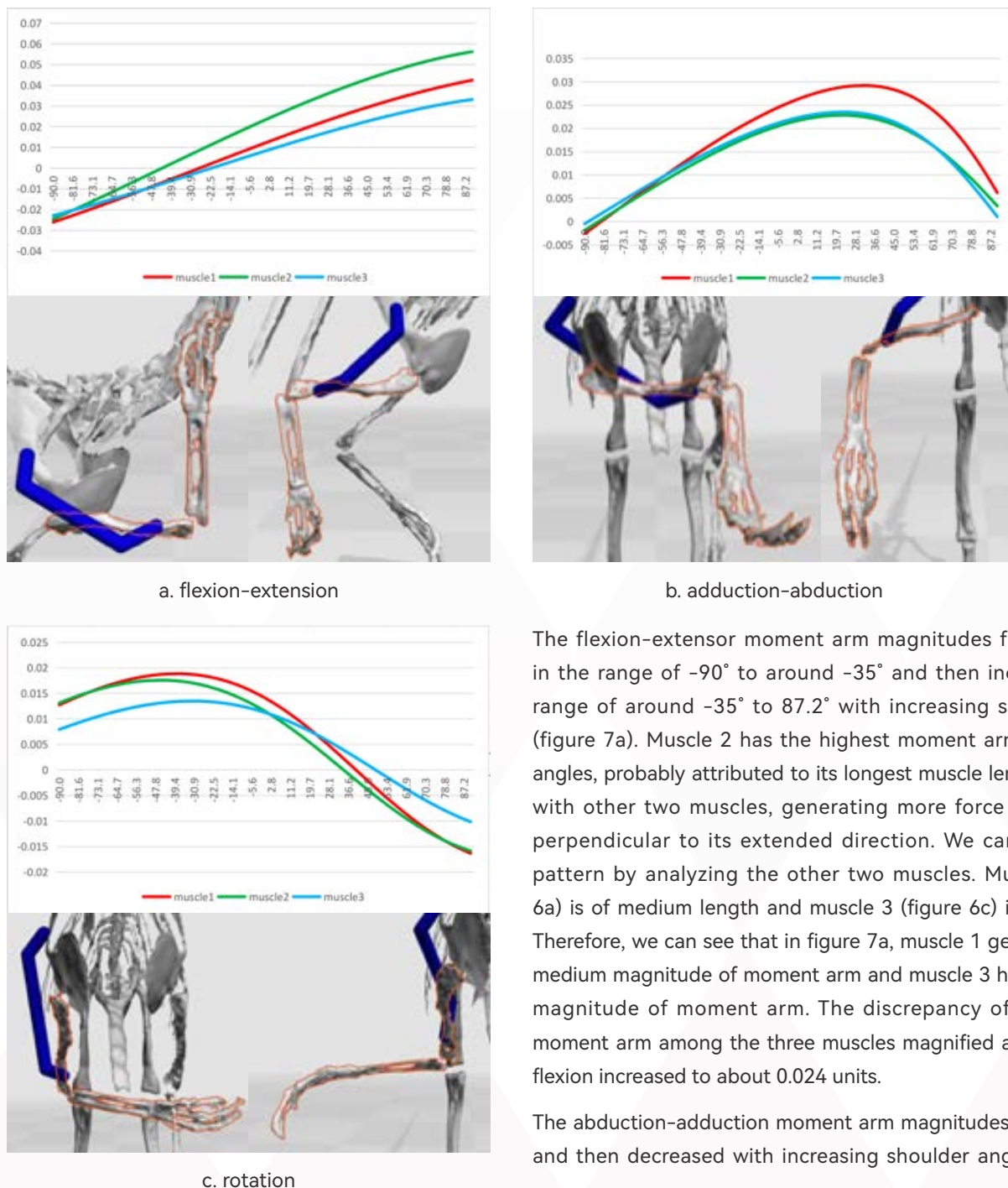


Figure 7. the (a) flexion-extension/(b) adduction-abduction/(c) rotation value (units: degrees)-moment arm graph

The flexion-extensor moment arm magnitudes first decreased in the range of -90° to around -35° and then increased in the range of around -35° to 87.2° with increasing shoulder angle (figure 7a). Muscle 2 has the highest moment arm in almost all angles, probably attributed to its longest muscle length compared with other two muscles, generating more force when moving perpendicular to its extended direction. We can identify this pattern by analyzing the other two muscles. Muscle 1 (figure 6a) is of medium length and muscle 3 (figure 6c) is the shortest. Therefore, we can see that in figure 7a, muscle 1 generally has the medium magnitude of moment arm and muscle 3 has the smallest magnitude of moment arm. The discrepancy of the extensor moment arm among the three muscles magnified as the shoulder flexion increased to about 0.024 units.

The abduction-adduction moment arm magnitudes first increased and then decreased with increasing shoulder angle (figure 7b).

While muscle 2 and muscle 3's trend almost coincided together, muscle 1 fluctuated in a wider range. To be specific, muscle 2 and 3 reached the maximum moment arm at about 0.024 units, yet muscle 1 reached the maximum moment arm at about 0.029 units. This is within the user variability in muscle creation. Living species also display differences in moment arms for the same muscle, correlated to slight anatomical differences in the build of their skeleton and muscle strength.

The rotation moment arm is more complicated (figure 7c). In general, all three muscles' moment arm magnitude first increased, then decreased, and finally increased as the shoulder rotating angle increases. Muscles 1, 2 and 3 reached its peak value at the angle of 35°, 45° and 29° respectively. Compared to muscle 1 and 2, muscle 3's moment arm magnitude varied in a smaller range, and the moment arms did not follow the same pattern. Instead, one of the muscles has an entirely different pattern, indicating that the moment arms peaked at different joint rotations. For example, Muscles 1 and 2 peaked during moments of internal rotation, whilst Moment 3 instead peaked during moment of external rotation. Ultimately, such differences can correlate to different functional capabilities of the muscle/model, and any significant differences in muscle function can entirely change the inferences of locomotory behaviour of a fossil individual. Thus it is imperative to acknowledge in all published studies, the degree of user-variability in muscle creation. Interestingly, the trend of the three lines formed a centrosymmetric graph.

Factors causing the discrepancies can be any feature of the muscles. The most obvious factors we can think of is the locations of the attaching points of the muscles. Their differences can directly cause the change in muscle length and the angle of the muscle line. These factors can affect the magnitude of moment arm, changing the force the muscle needed to drag adjacent bones to different angles. For example, a longer moment arm leads to a smaller force than doing the same action a shorter moment arm. Another factor causing the differences is the number of turning points. In figure 8, the number of turning points on muscle 3 is changed from 0 to 4. Although there are some differences, the trends are generally the same.

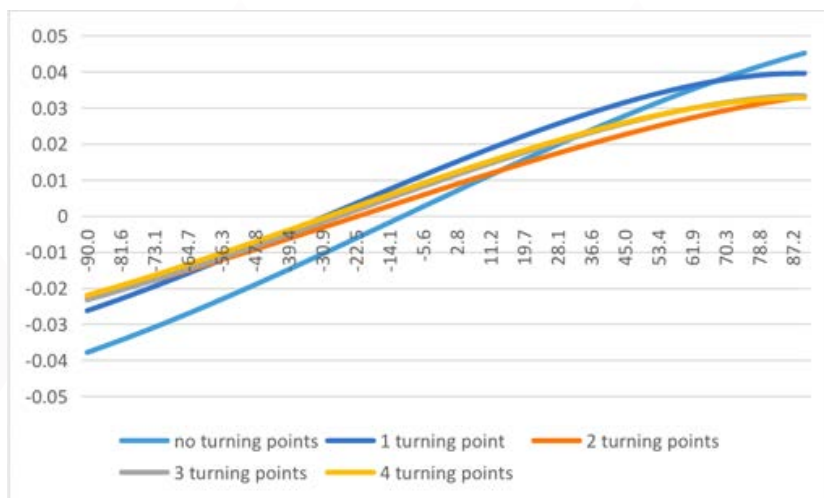


Figure 8. the differences between the moment arms for different numbers of turning points in muscle 3 when doing extension movements.

4. Conclusion

In most cases, the musculoskeletal module is generally the same. Our case study showed that flexion-extension and adduction-abduction movements module is quite reliable. Even changing the number of turning points will not affect the overall trend of the moment arm magnitude, too. Nevertheless, there are some conditions where the module is quite debatable such as the rotation moment arm. Consequently, we can refute the hypothesis that musculoskeletal models of extinct species are unreliable, and instead acknowledge that in this study, the muscle modelling was replicably created with a few exceptions. Future studies should repeat this methodology to ensure that the full user-variability of muscle modelling is captured, thus ensuring that erroneous conclusions are not reached (Hicks et al., 2015).

5. References

- [1] Bishop, P. J., Cuff, A. R., & Hutchinson, J. R. (n.d.). How to build a dinosaur: Musculoskeletal modeling and simulation of locomotor biomechanics in extinct animals. <https://doi.org/10.5061/dryad.73n5tb2v9>
- [2] Demuth, O. E., Herbst, E., Polet, D. T., Wiseman, A. L. A., & Hutchinson, J. R. (2023). Modern three-dimensional digital methods for studying locomotor biomechanics in tetrapods. In *Journal of Experimental Biology* (Vol. 226). Company of Biologists Ltd. <https://doi.org/10.1242/jeb.245132>
- [3] Hicks, J. L., Uchida, T. K., Seth, A., Rajagopal, A., & Delp, S. L. (2015). Is My Model Good Enough? Best Practices for Verification and Validation of Musculoskeletal Models and Simulations of Movement. In *Journal of Biomechanical Engineering* (Vol. 137, Issue 2). American Society of Mechanical Engineers (ASME). <https://doi.org/10.1115/1.4029304>



FitzEd

EDUCATIONAL PROGRAMMES
FITZWILLIAM COLLEGE, CAMBRIDGE

FitzEd
EDUCATIONAL PROGRAMMES
FITZWILLIAM COLLEGE, CAMBRIDGE



10

Philosophy in Cambridge: Past and Present

Fitzwilliam College Summer School Journal · 2024 Summer

Is Language Perfect? Could There Ever Be a Perfect Language?

YIFEI LI

1.

I argue that languages are perfect to themselves, and only perfect to themselves. It is impossible to have a specific language that can be judged as perfect. First, I will explain how languages have its self-consistency by analysing the mechanism of language, that is, language is the description of fact. Second, I will clarify why in this paper perfectness implies omnipotence including depicting Truth, and how it failed due to inversion. In conclusion, I reject the possibility of a perfect language through analysing the limit and externality born in language.

Wittgenstein defined language as picturing reality in *Tractatus Logico-Philosophicus* by saying “The picture can represent any reality whose form it has.” (Proposition 2.1) that its purpose is to describe the events in reality accurately. Then he made an edition of this definition in *Philosophical investigations*, that language is a form of life that closely associates with cultural and social practices. This use gives the symbols meaning and makes every language always perfect to itself because it serves the practices of a specific system. For any concept being not referred by a symbol in a language, the concept is not demanded by the culture who uses the concept. For example, in a village located in a place that never had climate variation, there would not be words for seasons, because there is no need to distinguish a certain period of time to another. When an English speaker fails to translate a specific German word to an English word, the self-consistency of each language is still unshaken. Indeed, the failure occurs precisely because that German word did not exist in English.

Nevertheless, the translation failure above indicates a certain inconsistency between different languages. The inconsistency is triggered by differences between every culture. These differences may originate from different geographic structures, types of food, invented tools, social system, etc. Even for artistic expressions, like painting and music, there is similar inconsistency. The emotions and thoughts contained within them could be only understood by those sensitive to the concepts. Imagine playing music to a group of amoebae, creatures that can't have any concept of emotions, their response would be no more than them moving away to avoid being harmed by high volume soundwaves.

If all languages are perfect to themselves, why would people desire a perfect language that overrides all other languages? Some philosophers desire a perfect language that is capable to explain the truth behind reality, in this case, perfectness mean the language can avoid any kind of distortion of the truth. Thus, they imagine such omnipotence being involved in “super languages” such as mathematics, logic, etc. For instance, Leibniz imagines a perfect language, *Characteristica Universalis*, that could represent all human knowledge and make reasoning as mathematics.

The assumption of a perfect language is plausible in circumstances that a transcendent truth exists, hidden beneath appearances. There has been a solid system supporting this. Plato states that everything in this world is a shadow of its ideal form in the world of ideas by saying “We must think of the Forms as the true realities of which the things we see and touch are but shadows.” (Plato, *Phaedo*). For Plato, there are Forms hidden behind presentation, our language is also a shadow guided by absolute meanings of reality. In this case, our language is to describe our inner experience of these implicit truth and forms. Therefore, there must be a perfect language to describe truth.

Nevertheless, none of this demonstrates why a perfect language can exist; or why any given language is not perfectly consistent with itself. The Truth that has been discussed refers to a kind of ontology, but language itself is a limitation imposed upon ontology. In language, rules such as definitions and grammar were set, to regulate the ineffable experience into something intersubjectively understandable and approvable, and to relate such experiences to symbolic appearances. The meaning precedes the symbol, the signifier precedes the signified. As Heidegger said “language is the house of being. In its home man dwells. Those who think and those who create with words are the guardians of this home.” (*Being*



And Time,1927) People use language to refer to experience and approach reality, but not invent language by describing experiences of reality. Since language is symbols born from culture, language could only shape and cover up the truth with cultural characteristics.

However, where the failure lies is not with language itself, it is with the expectations of these philosophers, and with their assumption of a language of truthTheir goal is to use language to describe something beyond language, but truth is only within language. I can never have any language beyond it, because it is not within the scope of my imagination. My reality is my language, because language is the only way that reality is presented to me.

In conclusion, all languages are forms of culture and life engaging with the world. They are and can only be perfect to themselves, failing in translating with each other is a proof of its self-consistency. A super language that interprets Truth cannot exist, because Truth in this case is imagined to be above the language of imagination, which is fundamentally impossible. However, this failure does not shake the self-consistency of languages, it just denied the assumption of an omnipotent language.

5. References

- [1] Wittgenstein, L., 2023. Tractatus logico-philosophicus. Accessed 8th August 2024: <http://public-library.uk/pdfs/9/292.pdf>
- [2] Martin Heidegger,1962. Being And Time. Accessed 8th august 2024:<https://archive.org/details/beingtime0000heid>
- [3] Plato,1951. Phaedo. Accessed 8th August 2024:<https://archive.org/details/phaedopl00plat>

Should impose restrictions on speech? If so, when and how? If not, why not?

LIWEI LIN

In this thesis, I assume that free speech is equivalent to free communication, and speech will be centred on as a moral issue rather than a constitutional one. To demonstrate more conclusively, I will first analyze the possible viewpoints of the opponents by concentrating on democracy theory. Subsequently, I will critique the internal constraints of this theory including the neglect of unequal discourse power and the inherent inconsistency of any appeal to democracy as a justification. I conclude that freedom of speech needs to be restricted for those with higher power and be encouraged for those who are overlooked over time.

In seeking to justify or restrict free speech, there are a plurality of interests. However, democracy theory is the only theoretical viewpoint appraised in this thesis. Put simply, democracy theory emphasizes the interests of democratic citizens. For instance, Rawls (2020) insists that citizens should be treated as free and equal moral agents by possessing the moral right to decide what to express or choose to listen to. Scholars, such as Whitten (2021), believe that free speech establishes a relatively equal relationship between citizens, where it promotes democracy. Likewise, Heinze (2016) and Barendt (2005) hold that freedom of speech is a significant component of democracy because restrictions on free speech hinder participation in a democracy.

Regarding democracy theory, I will spend all the remaining focus on elucidating the internal constraints of free speech in order to argue that the theory used to justify free speech itself justifies imposing some restrictions. This involves two parts: i) inequality of discourse power among different groups; ii) inconsistency of the term democracy.

Firstly, I firmly deny Rawls' claim that free speech can lead to equality among citizens. As Foucault asserts, "Discourse is power" (1971). Discourse is guided by the people in power and utilized as a means for mastering their power. Imagine a president and a student expressing the same opinion on the same topic; the weight of their words differ in society due to the preexisting distinction between their social status and power. Having power as "president" makes speech a tool for further exercising that power, forming a dominant discourse that inculcates from top to bottom.

Audiences who lack discourse power can easily be manipulated by ideology from superiors, which affect their self-identities and behaviors. An example mentioned by Bettcher and Shrage (2009) is that, when a woman is raped, the rapist often has a mysterious saying that "her mouth says no, but her eyes say 'yes'." The significance of this narrative—that of no means yes—is as evidence that women often have very little or no socially-recognized ethical authority to express consent. I maintain that this power imbalance is the outcome of the subordinated femininity which makes women's discourse power weaker than that of men. Foucault (1971) once wrote that: "There is not one but many silences, and they are an integral part of the strategies that underlie and permeate discourses". Thus, unrestricted speech will constantly expand the hegemony and subsidiary characteristics between the subject of speech and the object of listening, going against the goal of democracy and equality.

Secondly, if citizens are encouraged to truly practice democracy, then they should be free to discuss any issue, including whether trust should be suspended about their "democratic speech rights" (Howard, 2024). Greene and Simpson (2017) point out that "Why should we insist on a conception of democracy that contains a self-destruct mechanism? Merely stipulating that democracy requires this is not enough". Therefore, simply stating that freedom of speech constitutes democracy is too arbitrary as in fact democracy contradicts this theory.

Bibliography

[1] Barendt, E. (2005). Freedom of speech. OUP Oxford.

[2] Bettcher, T. M., & Shrage, L. (2009). Trans identities and first-person authority. *You've changed: sex reassignment and personal identity*, 1, 98-120.

[3] Foucault, M. (1971). Orders of discourse. *Social science information*, 10(2), 7-30.

[4] Greene, A. R., & Simpson, R. M. (2017). *Tolerating hate in the name of democracy*.

[5] Heinze, E. (2016). *Hate speech and democratic citizenship*. Oxford University Press.

[6] Rawls, J. (2020). Political liberalism. In *The New Social Theory Reader* (pp. 123-128). Routledge.

[7] Rawls, J. (2017). A theory of justice. In *Applied ethics* (pp. 21-29). Routledge.

[8] Whitten, S. (2021). *A republican theory of free speech: critical civility*. Springer Nature.

Resources from the Internet

[1] Howard, Jeffrey W., "Freedom of Speech", *The Stanford Encyclopedia of Philosophy* (Spring 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <<https://plato.stanford.edu/archives/spr2024/entries/freedom-speech/>>.



11

Nuclear Engineering

Assessing Public Opinion on Nuclear, Benefits and The Future of Nuclear

MALIK ALAMRI

1. Introduction

A negative stigma has surrounded nuclear power for many years, which has slowed down the nuclear industry,[1] but public reception to nuclear has been improving with time as the economic and environmental benefits become clearer and information becomes more accessible to the public.[2] People are more accepting of nuclear energy as they learn more about it[3]. Generation IV nuclear reactors are reactor designs proposed by the Gen IV International Forum (GIF), which will be manufactured in the future and have improved safety, economics and sustainability compared to other energy sources[4]. Gen IV nuclear reactors consist of 6 designs chosen by the GIF[5]. The GIF consists of 14 active countries and focuses more on sharing R&D rather than the manufacture of actual nuclear reactors[6]. The GIF currently focuses on 4 aspects related to gen IV nuclear reactors, which are: Sustainability, Safety and Reliability, Economics and Proliferation Resistance and Physical Protection.[7]

2. Purpose of Essay

This essay investigates how negative public opinion and distrust of media and government outlets affects the production of nuclear power and proposes certain methods which should be taken to show the benefits nuclear reactors and how Gen IV nuclear reactors will continue to improve and build upon these aspects while being economically viable.

3. Problems Preventing Growth of Nuclear Energy

Nuclear energy is one of the safest, cleanest and efficient energy sources available[8],but the negative public opinion stunts the growth of nuclear worldwide[9].The negativity stems around the public's unawareness on the benefits of nuclear power and backlash towards media and government outlets[10].Chernobyl, Fukushima and 3 Mile Island have altered the public opinion on nuclear and have produced drastic decreases in nuclear energy after their occurrences[11].

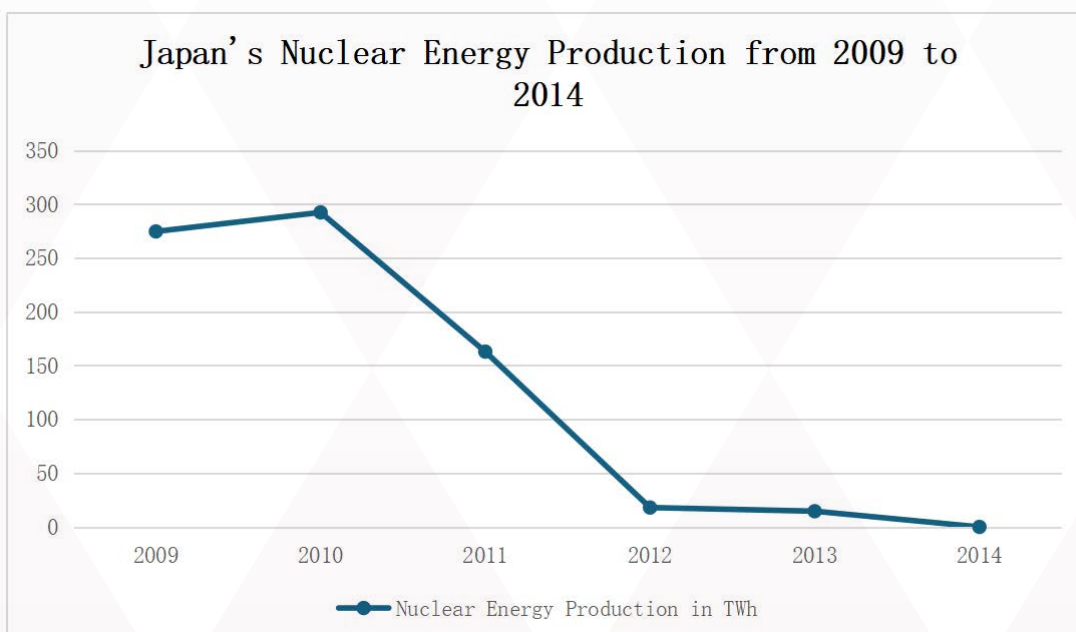


Figure 1: Japanese Nuclear Power Production in TWh Before and After the Fukushima Meltdown

Japan’s production of nuclear energy significantly decreased after the Fukushima meltdown in 2011, with Japan suspending all of their remaining nuclear power plants in 2013 and relying on natural gas imports[12]. After Fukushima, the Japanese prime minister at the time, Naoto Kan said, “Japan should aim for a society that does not depend on nuclear energy,” Mr. Kan said, “We should reduce our dependence in a planned and gradual way, and in the future, we should aim to get by with no nuclear energy.”[13] Japanese citizens protested the use of nuclear energy for years after the Fukushima meltdown, which implies that the meltdown had long-lasting effects on Japanese opinions on nuclear power[14]

4. Safety and Environmental Benefits to Nuclear Power

Even considering the catastrophes which have happened in nuclear reactors, it remains one of the safest energy sources. (xv) Gen IV nuclear reactors aim to maintain this level of safety and improve upon it by having both active and “passive” safety systems which use laws of physics such as gravity, and this would render a reactor meltdown and other accidents to be physically impossible (xvi)

Nuclear power is also one of the most environmentally friendly energy sources in the world. (xx) Producing far less CO2 emissions than coal, gas and oil. (xxi) Fears of radioactive material entering the environment also become unwarranted as nuclear reactors have multiple systems, redundant backup systems and large containment structures to absolutely ensure that no radioactive or hazardous waste is released in areas which can affect life. (xxii) All decommissioned radioactive fuel is also kept in very secure storage facilities away from human life. (xxiii)

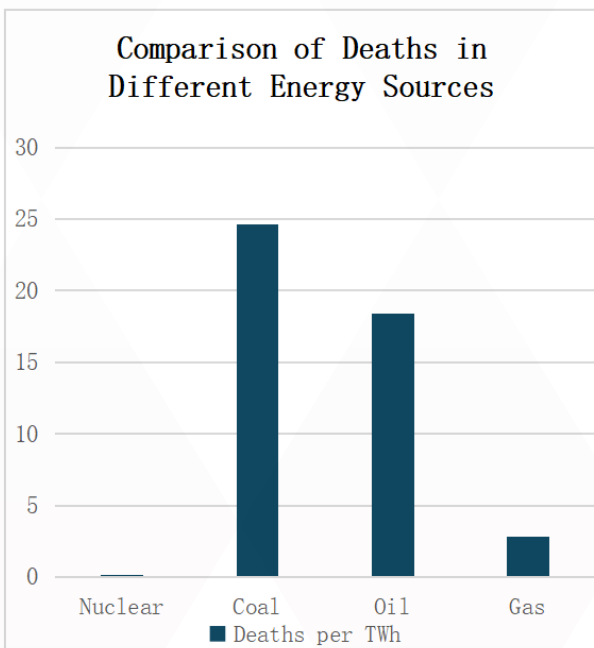


Figure 3: Comparison of deaths per TWh In Coal, Oil, Gas and Nuclear

Nuclear is by far the safest energy source compared to coal, oil and gas, with nuclear energy having 0.03 deaths (/TWh) compared to coal which has 24.62 deaths (/TWh). (xvii) Using this chart, nuclear energy is approximately 821x safer than coal energy. Chernobyl and Fukushima are the only major catastrophes to occur in over 18,500 cumulative years in which nuclear reactors have been operating. (xviii) The design of a nuclear power plant is made to be as safe as possible, having multiple methods and backup systems of shutting down a reactor in case of a meltdown. (xix)

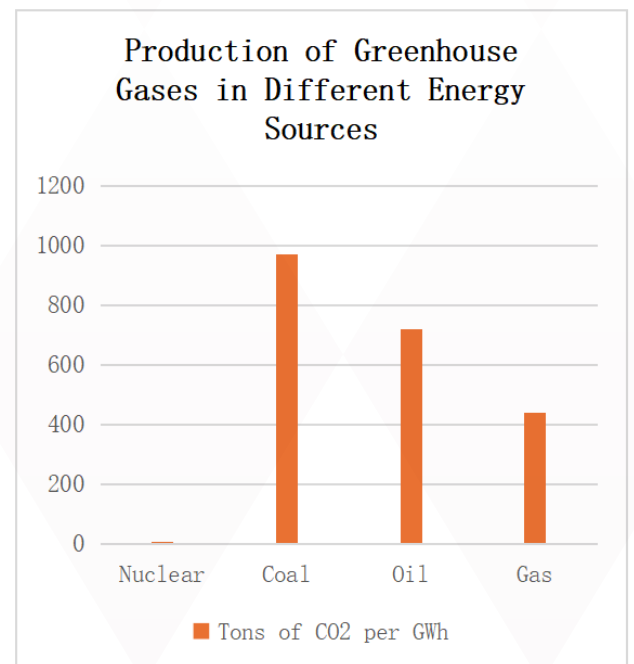


Figure 2: Comparison in the production of greenhouse gases in tons per GWh in Coal, Oil, Gas and Nuclear

The GIF aims to manage and minimise their nuclear waste produced, (xxiv) and they will produce sustainable energy that meets clean air objectives, so the low rate at which CO2 is produced by nuclear will remain the same or be potentially even lower. (xxv)

5. Gen IV International Forum Goals

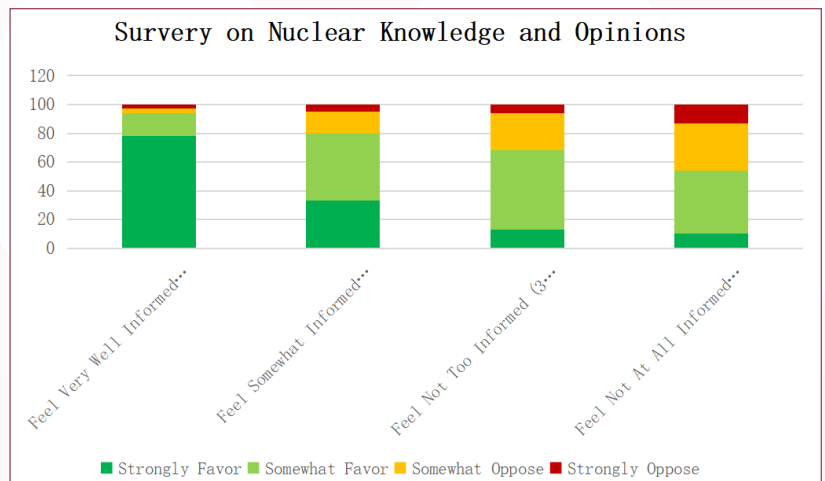
Sustainability-1	Generation IV nuclear energy systems will provide sustainable energy generation that meets clean air objectives and provides long-term availability of systems and effective fuel <u>utilisation</u> for worldwide energy production.
Sustainability-2	Generation IV nuclear energy systems will <u>minimise</u> and manage their nuclear waste and notably reduce the long-term stewardship burden, thereby improving protection for the public health and the environment.
Economics-1	Generation IV nuclear energy systems will have a clear life-cycle cost advantage over other energy sources.
Economics-2	Generation IV nuclear energy systems will have a level of financial risk comparable to other energy projects.
Safety and Reliability-1	Generation IV nuclear energy systems operations will excel in safety and reliability.
Safety and Reliability-2	Generation IV nuclear energy systems will have a very low likelihood and degree of reactor core damage.
Safety and Reliability-3	Generation IV nuclear energy systems will eliminate the need for offsite emergency response.
Proliferation Resistance and Physical Protection	Generation IV nuclear energy systems will increase the assurance that they are very unattractive and the least desirable route for diversion or theft of weapons-usable materials and provide increased physical protection against acts of terrorism.

The goals set by the GIF show that measures are being taken in order to make Gen IV nuclear reactors as safe, economical and sustainable as possible while maintaining and improving upon the efficiency and effectiveness of previous nuclear reactors (xxvi). Gen IV reactors are designed to be safer, more efficient, and more sustainable than previous generations (xxvii). Gen IV reactors use state-of-the-art materials, designs, and cooling systems, and are intended to be more economical, scalable and adaptable. (xxviii).

6. Proposed Method of Gaining Public Support

Research and studies show that as people think they know more about nuclear energy and the overall benefits, they tend to be more welcoming of it, but people who think they are uninformed on nuclear tend to produce the opposite reaction, having negative feelings on it. (xxix)

Figure 4: American opinions on nuclear power after being asked how knowledgeable they are on the subject (xxx)



People who feel uninformed on nuclear energy much more likely strongly oppose it than people who think they feel very informed (xxxix). So, to gain overall support from the general public, a proposed idea would be to inform people on nuclear energy, how it works, and the benefits of it. The outlet from which this message is spread needs to be considered carefully as surveys also reveal that some outlets are much less likely to be trusted as sources of academic material (xxxix) The GIF has support from numerous major countries, so an effective method of gaining the public's trust would be to utilize nuclear related government branches and have a campaign spreading the effectiveness and benefits of nuclear energy and gen IV nuclear reactors specifically to ensure no backlash when they start being manufactured.

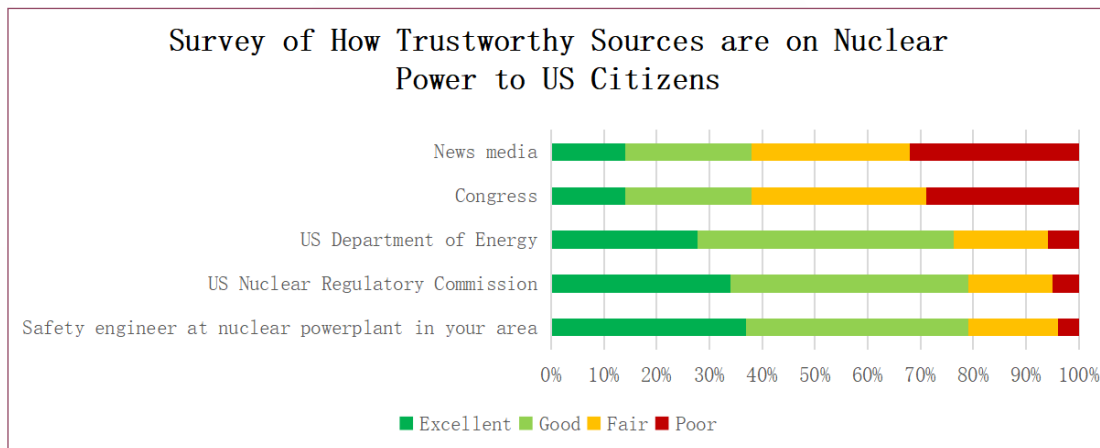


Figure 5: American opinions on how trustworthy specific sources are for nuclear power information (xxxix)

7. Conclusion

In conclusion, even taking into consideration the accidents that have happened with nuclear power, it still remains one of the best energy sources. One of the biggest factors preventing more growth is the negative public reception, but this paper explores a certain method which can be used by the GIF to gain public support of nuclear and Gen IV nuclear reactors.

8. References

- [I] Dr Hans Blix, The post-Chernobyl outlook for nuclear power, IAEA Bulletin, Autumn 1986, (p.9)
- [II] Ann S. Bisconti, PhD, 2023 National Nuclear Energy Public Opinion Survey: Public Support for Nuclear Energy Stays at Record Level For Third Year in a Row, Bisconti Research Inc. (<https://www.bisconti.com/blog/public-opinion-2023>), April-May 2023, Cited on 8/8/2024
- [III] Ann S. Bisconti, May 2021 National Public Opinion Survey: Support for Nuclear Energy Groups with Climate Change Concerns, Bisconti Research Inc. (<https://www.bisconti.com/blog/climate-change-concerns>), June 15, 2021, Cited on 8/8/2024
- [IV] Hannah Ritchie and Pablo Rosado, Nuclear Energy, Our World in Data, (<https://ourworldindata.org/nuclear-energy>), July 2020, Cited on 8/8/2024
- [V] Gen IV International Forum, Technology, Generation IV Systems, (https://www.gen-4.org/gif/jcms/c_59461/generation-iv-systems), No date, Cited on 8/8/2024
- [VI] Gen IV International Forum, Generation IV Goals, (https://www.gen-4.org/gif/jcms/c_9502/generation-iv-goals), No date, Cited on 8/8/2024
- [VII] Gen IV International Forum, Generation IV Goals, (https://www.gen-4.org/gif/jcms/c_9502/generation-iv-goals), No

date, Cited on 8/8/2024

[VIII] Hannah Ritchie and Pablo Rosado, Nuclear Energy, Our World in Data, (<https://ourworldindata.org/nuclear-energy>), July 2020, Cited on 8/8/2024

[IX] Dr Hans Blix, The post-Chernobyl outlook for nuclear power, IAEA Bulletin, Autumn 1986, (p.9)

[X] Editorial Board, Germany is closing its last nuclear plants. What a mistake., Washington Post, (<https://www.washingtonpost.com/opinions/2022/01/01/germany-is-closing-its-last-nuclear-plants-what-disaster/>), Jan 1, 2022, Cited on 8/8/2024

[XI] Hannah Ritchie and Pablo Rosado, Nuclear Energy, Our World in Data, (<https://ourworldindata.org/nuclear-energy>), July 2020, Cited on 8/8/2024

[XII] BBC News, Japan halts last nuclear reactor at Ohi, (<https://www.bbc.co.uk/news/world-asia-24099022>), 15 September 2013

[XIII] Hiroko Tabuchi, Japan Premier Wants Shift Away From Nuclear Power, The New York Times, (https://www.nytimes.com/2011/07/14/world/asia/14japan.html?_r=1&hp), July 13, 2011, Cited on 8/8/2024

[XIV] Alexander Brown, The Anti-nuclear Movement and Street Politics in Japan after Fukushima, ASAA, (<https://asaa.asn.au/anti-nuclear-movement-street-politics-japan-fukushima/>), June 25, 2018, Cited on 8/8/2024

[XV] Hannah Ritchie and Pablo Rosado, Nuclear Energy, Our World in Data, (<https://ourworldindata.org/nuclear-energy>), July 2020, Cited on 8/8/2024

[XVI] The LFR provisional System Steering Committee, SAFETY DESIGN CRITERIA FOR GENERATION IV LEAD-COOLED FAST REACTOR SYSTEM, Gen IV International Forum, March 2021, (p.20)

[XVII] Hannah Ritchie and Pablo Rosado, Nuclear Energy, Our World in Data, (<https://ourworldindata.org/nuclear-energy>), July 2020, Cited on 8/8/2024

[XVIII] World Nuclear Association, Safety of Nuclear Power Reactors, (<https://world-nuclear.org/information-library/safety-and-security/safety-of-plants/safety-of-nuclear-power-reactors>), Updated March, 2022, Cited on 8/8/2024

[XIX] Gen IV International Forum, Benefits and Challenges, Question 8., (https://www.gen-4.org/gif/jcms/c_40368/benefits-and-challenges#c_43122), No date, Cited on 8/8/2024

[XX] Hannah Ritchie and Pablo Rosado, Nuclear Energy, Our World in Data, (<https://ourworldindata.org/nuclear-energy>), July 2020, Cited on 8/8/2024

[XXI] Hannah Ritchie and Pablo Rosado, Nuclear Energy, Our World in Data, (<https://ourworldindata.org/nuclear-energy>), July 2020, Cited on 8/8/2024

[XXII] U.S.NRC, High level waste, (<https://www.nrc.gov/waste/high-level-waste.html>), Updated March, 2020, Cited on 8/8/2024

[XXIII] U.S.NRC, High level waste, (<https://www.nrc.gov/waste/high-level-waste.html>), Updated March, 2020, Cited on 8/8/2024

[XXIV] Gen IV International Forum, Generation IV Goals, (https://www.gen-4.org/gif/jcms/c_9502/generation-iv-goals), No

date, Cited on 8/8/2024

[XXV] Gen IV International Forum, Generation IV Goals, (https://www.gen-4.org/gif/jcms/c_9502/generation-iv-goals), No date, Cited on 8/8/2024

[XXVI] Gen IV International Forum, Generation IV Goals, (https://www.gen-4.org/gif/jcms/c_9502/generation-iv-goals), No date, Cited on 8/8/2024

[XXVII] Gen IV International Forum, Generation IV Goals, (https://www.gen-4.org/gif/jcms/c_9502/generation-iv-goals), No date, Cited on 8/8/2024

[XXVIII] Robert Rapier, Fourth Generation Nuclear Reactors Take A Big Step Forward, Forbes, (<https://www.forbes.com/sites/rrapier/2023/04/24/fourth-generation-nuclear-reactors-take-a-big-step-forward/>), April 24, 2023, Cited on 8/8/2024

[XXIX] Ann S. Bisconti, May 2021 National Public Opinion Survey: Support for Nuclear Energy Groups with Climate Change Concerns, Bisconti Research Inc. (<https://www.bisconti.com/blog/climate-change-concerns>), June 15, 2021, Cited on 8/8/2024

[XXX] Ann S. Bisconti, May 2021 National Public Opinion Survey: Support for Nuclear Energy Groups with Climate Change Concerns, Bisconti Research Inc. (<https://www.bisconti.com/blog/climate-change-concerns>), June 15, 2021, Cited on 8/8/2024

[XXXI] Ann S. Bisconti, May 2021 National Public Opinion Survey: Support for Nuclear Energy Groups with Climate Change Concerns, Bisconti Research Inc. (<https://www.bisconti.com/blog/climate-change-concerns>), June 15, 2021, Cited on 8/8/2024

[XXXII] Ann S. Bisconti, May 2021 National Public Opinion Survey: Support for Nuclear Energy Groups with Climate Change Concerns, Bisconti Research Inc. (<https://www.bisconti.com/blog/climate-change-concerns>), June 15, 2021, Cited on 8/8/2024

[XXXIII] Ann S. Bisconti, May 2021 National Public Opinion Survey: Support for Nuclear Energy Groups with Climate Change Concerns, Bisconti Research Inc. (<https://www.bisconti.com/blog/climate-change-concerns>), June 15, 2021, Cited on 8/8/2024

Nuclear Reactors for Medical Isotopes: An Overview

FATIMAH AHMED ALABDULLAH

1. Introduction

Radioisotopes are atoms with an unstable balance of neutrons and protons or excess energy in their nucleus. Nuclear reactors produce these radioactive materials by generating high levels of neutrons through fission and activation[1]. Radioisotopes are used in so many fields including medicine, for the diagnosis and treatment of a wide range of different health diseases[2]. Each year, global demand for these therapies rises by up to 5%, showing they are key components of radiopharmaceuticals[2]. Therefore, this essay gives an overview of radioisotope production using nuclear reactors, most used medical isotopes, and their applications.

2. Nuclear reactors for isotopes production

2.1 Types of reactors

Most reactor-based isotopes are generated using research reactors. Unlike power reactors which are used for electricity generation, one of the main roles of nuclear research reactors is to produce radioisotopes for medical applications since long ago. Research reactors are small, specialized nuclear reactors that are mainly used for research and isotope production rather than electricity generation. They generate neutron beams, often with lower power outputs than power reactors, and require highly enriched uranium. Their design prioritizes efficient neutron generation, often requiring specific cooling, moderation, and reflecting systems[4,5,6].

2.2 Production method

Radioisotopes are created in reactors by exposing target materials to high neutron fluxes. In light-water reactors, targets are stored in capsules and lowered into the reactor core for irradiation, then transferred to labs for processing. Heavy-water reactors require more complicated assemblies with many target capsules. The quality and specific activity of the generated radioisotopes is determined by the target material and irradiation settings[7].

2.3 Operating research reactors

As of June 2021, the Research Reactor Database of IAEA (The International Atomic Agency Emergency) represented 223 operating research reactors around the world, divided as shown in the table:[6]

Table 2.3: 2021 operational research reactors around the world. (Data from[6])

Country	Operational research reactors	Country	Operational research reactors	Country	Operational research reactors	Country	Operational research reactors
Russia	52	Canada	5	Kazakhstan	4	Indonesia	3
USA	50	Germany	5	Belarus	3	Japan	3
China	16	Italy	5	Belgium	3	Ukraine	3
India	7	Brazil	4	Czech Republic	3	Others	45
Argentina	5	Iran	4	France	3	Total	223

2.4 The most used research reactors for isotope production

The high flux reactor (HFR) located in Petten, the Netherlands, is the most commonly used reactor for the production of medical isotopes. A 45 MW light water-cooled and moderated research reactor designed similarly to the Oak Ridge Research (ORR) reactor. It's a multipurpose reactor with 17 available core positions for irradiation experiments. It's operated under contract by the Netherlands Energy Research Foundation and funded by a programme between Netherlands and Germany[8].

2.5 challenges

Approximately 50% of the operating research reactors around the world have reached 50 years of operating and 20% of them have reached 60 years based on the IAEA Research Reactor Database. Therefore, there is an urgent need to establish an aging management and modification plan to enhance the effectiveness and maintain safety[6].

3. Half-life of radioisotopes

The half-life of radioactive material is a time period for the radioactive decay process, the faster the rate of decay, the shorter the half-life. And It's different for each radioisotope[9].

4. Common medical isotopes and their applications

According to the World Nuclear Association, there are about 40 active radioisotopes generated by reactors and used for medical purposes. They can be used to diagnose and / or treat various diseases including cancer, heart diseases, thyroid illness, and even brain diseases such as Alzheimer's[2].

Table 4.0: Some common radioisotopes in medicine. (Data from[2])

Radioisotope	Application	Half-life
Technetium-99m (The most common radioisotope for diagnostic)	Image the skeleton and heart muscle. Brain, thyroid, lungs (perfusion and ventilation). Liver, spleen, kidney (structure and filtration rate). Gall bladder, bone marrow, salivary and lacrimal glands, heart blood pool, infection, and numerous specialized medical studies.	6h
Bismuth-213	Targeted alpha therapy (TAT), especially cancers.	46min
Caesium-137	Low-intensity sterilization of blood and in brachytherapy.	30yr
Holmium-166	Diagnosis and treatment of liver tumors.	26h

Rhenium-186	Pain relief in bone cancer. Beta emitter with weak gamma for imaging.	3.8d
Selenium-75	Study the production of digestive enzymes.	120d

5. Conclusions

Medical isotopes are promising tools in diagnosing and treating various diseases. Yet, challenges exist and safety proportions must be followed. Moreover, new plans need to be prepared for any unexpected event. There is a need for new reactors and / or modernizing projects for existing reactors. If this was done with maintaining the safety, more radioisotopes would exist which means more patients would be helped.

6. References

- [1] What are radioisotopes?. ANSTO. (n.d.). <https://www.ansto.gov.au/education/nuclear-facts/what-are-radioisotopes#:~:text=Radioisotopes%20in%20medicine,to%20make%20an%20accurate%20diagnosis>.
- [2] Radioisotopes in medicine. World Nuclear Association. (n.d.-a). <https://world-nuclear.org/information-library/non-power-nuclear-applications/radioisotopes-research/radioisotopes-in-medicine>
- [3] Health and Nuclear Medicine. National Nuclear Laboratory. (2024, February 12).<https://www.nnl.co.uk/focus-areas/health-and-nuclear-medicine/>
- [4] IAEA. (2016, July 15). Radioisotope production in research reactors. IAEA.<https://www.iaea.org/topics/radioisotope-production-in-research-reactors#:~:text=Radioisotope%20production%20in%20reactors%20is,by%20bombardment%20with%20thermal%20neutrons>.
- [5] International Atomic Energy Agency (IAEA). (2023, December). Research Reactors. <https://www.iaea.org/sites/default/files/researchreactors.pdf>
- [6] Research reactors. World Nuclear Association. (n.d.-a). <https://world-nuclear.org/information-library/non-power-nuclear-applications/radioisotopes-research/research-reactors>
- [7] Production methods. NIDC: National Isotope Development Center. (n.d.).<https://www.isotopes.gov/production-methods>
- [8] Ahlf, J. (1989, January 1). High flux reactor petten present and future programme.inis.iaea.org. https://inis.iaea.org/search/search.aspx?orig_q=RN%3A21088004
- [9] Libretexts. (2023, April 3). 8.3: Half-life of Radioisotopes. Chemistry LibreTexts. [https://chem.libretexts.org/Bookshelves/Introductory_Chemistry/Introduction_to_General_Chemistry_\(Malik\)/08%3A_Nuclear_chemistry/8.03%3A_Half-life_of_radioisotopes#:~:text=The%20half%20life%20\(t1,the%20shorter%20the%20half%20life](https://chem.libretexts.org/Bookshelves/Introductory_Chemistry/Introduction_to_General_Chemistry_(Malik)/08%3A_Nuclear_chemistry/8.03%3A_Half-life_of_radioisotopes#:~:text=The%20half%20life%20(t1,the%20shorter%20the%20half%20life)



 **FitzEd**
EDUCATIONAL PROGRAMMES
FITZWILLIAM COLLEGE, CAMBRIDGE

Fitzwilliam College
Cambridge
CB3 0DG
+44 (0)1223 332000

<https://www.fitz.cam.ac.uk/educational-programme>
Fitzwilliam College Online Summer School

In Mainland China the Programme is offered in partnership with ASEEDER China for outstanding high school students.

 **阿思丹 ASEEDER**

<https://www.seedasdan.com/en/fitz/>