Term 2 Math Methods Notes

Physsoc Council 2024

June 2, 2024

1 Fourier Series

A good portion of the content in this section is based on the "Principles of Mathematical Analysis" by Walter Rudin, along with material from MATH 321 and 420 at UBC.

L^p spaces and orthogonality:

To begin a discussion on Fourier Series, it is imperative that the perspective one looks at functions with is broadened slightly. In particular, we will consider a special class of functions, namely:

Definition 1.1. *We denote by $L^p[a,b]$:*

$$L^p[a,b] = \left\{ f : [a,b] \to \mathbb{C} \;\middle|\; \|f\|_p^p = \int_a^b |f|^p \, dx < \infty \right\}$$

In particular, we will focus on either $L^1[a,b]$, the set of all integrable functions, and $L^2[a,b]$, the set of all square integrable functions $f : [a,b] \to \mathbb{C}$.

Above, we make use of the following notation:

Notation 1.2. $\|f\|_p = (\int_a^b |f|^p \, dx)^{p^{-1}}$

It can be seen that $L^p[a,b]$ is in fact a vector space over $\mathbb{C}$:

Proposition 1.1. *$L^p[a.b]$ is a vector space over $\mathbb{C}$, where the addition $(f+g)(x) = f(x) + g(x)$ denotes the pointwise addition of functions f, g.*

Proof. It suffices to show that $L^p[a,b]$ is closed under scalar multiplication, and the pointwise addition operation:

- For $c \in \mathbb{C}$, $f \in L^p[a.b]$, $|cf| = |c||f|$, and thus it follows that $\int_a^b |cf|^p = |c^p| \int_a^b |f|^p \, dx = |c^p| \|f\|_p^p < \infty$, which implies $cf \in L^p[a,b]$.
- For $f, g \in L^p[a.b]$, via the triangle inequality:

  $$|f+g|^p \leq (|f| + |g|)^p = (|f| + |g|)^p \leq (2 \max\{|f|, |g|\}^p) = 2^p \max\{|f|, |g|^p\} \leq 2^p (|f|^p + |g|^p)$$

  Integrating both sides yields $\int_a^b |f+g|^p \, dx = \|f+g\|_p^p \leq 2^p (\|f\|_p^p + \|g\|_p^p) < \infty$, which implies $f + g \in L^p[a.b]$

This completes the proof. □

As it turns out, we have shown that these particular sets of functions can be thought of as vector spaces. Of all values of p, $p = 2$ is in particular special. As it turns out, we can do even more for $p = 2$ specifically (more discussion on this can be found later in the section). Recall the definition of the inner product:

Definition 1.3. *Let V be a vector space over $\mathbb{C}$ (or a subset of $\mathbb{C}$) and let $\vec{u}, \vec{v} \in V$, $k \in \mathbb{C}$. An inner product is a mapping*

$$\langle \vec{u}, \vec{v} \rangle : V \times V \to \mathbb{C}$$

with the following properties:

1. $\langle \vec{u}, \vec{u} \rangle \geq 0$ *with equality only when* $\vec{u} = 0$
2. $\langle \vec{u}, \vec{v} \rangle = \langle \vec{v}, \vec{u} \rangle^*$
3. $k\langle \vec{u}, \vec{v} \rangle + \langle \vec{u}, \vec{w} \rangle = \langle \vec{u}, k\vec{v} + \vec{w} \rangle$

We can then introduce an inner product on $L^2[a, b]$ via:

Proposition 1.2. *We define the inner product on $L^2[a, b]$ for functions $f, g \in L^2[a, b]$ via* [1]*:*

$$\langle f, g \rangle = r \int_a^b f^* g \, dx$$

where f^* refers to the complex conjugate (function) of f, and $r \in \mathbb{R}$, $r > 0$ (Usually $r = 1$ or $(b - a)^{-1}$)

Proof. It suffices to show that $\langle ., . \rangle$ satisfies the axioms of an inner product outlined in definition 1.3:

- The first part of the first axiom is easily verified, as $\langle f, f \rangle = \|f\|_2^2 \geq 0$, via the monotonicity property of the integral. If $\langle f, f \rangle = 0$, then for our intents and purposes, $f = 0$.[2]
- Let $f = f_r + if_i$, and so on. Then:

$$\begin{aligned}
\langle f, g \rangle^* &= \left(\int_a^b f^* g \, dx \right)^* \\
&= \left(\int_a^b f_r g_r + f_i g_i \, dx + i \int_a^b f_r g_i - f_i g_r \, dx \right)^* \\
&= \left(\int_a^b f_r g_r + f_i g_i \, dx \right) - i \left(\int_a^b f_r g_i - f_i g_r \, dx \right) \\
&= \int_a^b f_r g_r - i f_r g_i + i f_i g_r + f_i g_i \, dx \\
&= \int_a^b (g_r - i g_i)(f_r + i f_i) \, dx = \int_a^b g^* f \, dx = \langle g, f \rangle
\end{aligned}$$

[1] You may see the complex conjugate on the second argument instead of the first online. This is mostly a matter of convention - physics generally sticks with linearity in the second argument, while math the first.

[2] This is actually not true - consider $f(x) = 1$ if $x = 0$ and $f(x) = 0$ otherwise. It turns out L^p spaces are actually equivalence classes of functions, where functions equivalent almost everywhere (with respect to a measure) are considered identical. Essentially, these pathological examples are lumped together in the equivalence class of the 0 function. tl;dr we can assume for $f \geq 0$, $\int f \, dx = 0 \implies f = 0$.

- Let $k \in \mathbb{C}$, and $h \in L^2[a,b]$. Then:

$$\begin{aligned} k\langle f,g\rangle + \langle f,h\rangle &= k\int_a^b f^* g \; dx + \int_a^b f^* h \; dx \\ &= \int_a^b f^*(kg+h) \; dx = \langle f, kg+h\rangle \end{aligned}$$

Having shown that all axioms are satisfied, this verifies $\langle .,. \rangle$ is an inner product, as required. □

Having an inner product, we can then talk about what it means for functions to be orthogonal:

Definition 1.4. *Two functions $f, g \in L^2[a,b]$ are said to be orthogonal iff $\langle f, g\rangle = 0$*

As an example of two orthogonal functions, we can see that $\sin x$ and $\cos x$ are orthogonal on $[-\pi, \pi]$. In fact, on symmetric intervals $[-a, a]$, any product of an odd and even function would be orthogonal, and this lets us come up with a plethora of examples of function orthogonality.

Now that we have seen that orthogonality is not difficult to find in functions, one may wonder how this is applied, and why our additional machinery to understand functions as vectors is even useful, apart from being a rather nice analogy. It is a common theme to try and express a function in terms of better understood ones, the reader has likely already done so several times when applying Taylor's theorem! As it turns out, $L^2[a,b]$, and orthogonality (especially of trigonometric functions such as sine and cosine), turn out to be extremely useful in approximating functions using trigonometric ones (Fourier series), as opposed to monomials like in Taylor's theorem.

Additional content on L^p spaces:

In this optional section, we go through a few preliminary inequalities that often come up with L^p spaces [3], as well as some discussion on why L^2 is special.

Proposition 1.3. *Young's inequality: if $a, b, p, q \in \mathbb{R}$, $a, b \geq 0$, $p, q > 1$, with $p^{-1} + q^{-1} = 1$, then:*

$$ab \leq p^{-1}a^p + q^{-1}b^q$$

Proof. If either $a = 0$, or $b = 0$, then the claim is clearly true. Assume then that $a, b \neq 0$. Then consider the function $h(x) = \log x$. As $h'(x) = x^{-1}$, and $h''(x) = -x^{-2} < 0$ for all $x > 0$, h is concave for $x > 0$. Thus, from the definition of concavity:

$$\begin{aligned} \log\left(p^{-1}a^p + q^{-1}b^q\right) &= h(p^{-1}a^p + q^{-1}b^q) \\ &\geq p^{-1}h(a^p) + q^{-1}h(b^q) \\ &= p^{-1}\log a^p + q^{-1}\log b^q = \log a + \log b = \log ab \end{aligned}$$

exponentiating both sides completes the proof. □

Proposition 1.4. *Holder's inequality: Let p, q be as above. Then for $f \in L^p[a.b]$, $g \in L^q[a,b]$:*

$$\|fg\|_1 \leq \|f\|_p \|g\|_q$$

[3] This content is not entirely necessary for the discussions below, but proves that L^p spaces form a metric space!

Proof. If $\|f\|_p = 0$ (or $\|g\|_q = 0$), then $f = 0$ ($g = 0$), which means $fg = 0$, so the inequality holds in this case. We then assume $\|f\|_p, \|g\|_q \geq 0$, and by virtue of being elements of their L^p spaces, $\|f\|_p, \|g\|_q < \infty$. Then, we define $F = f\|f\|_p^{-1}$, $G = g\|G\|_p^{-1}$, so that $\|F\|_p = \|G\|_q = 1$. Then via Young's inequality, we note that:

$$|F||G| = |fg| \leq p^{-1}|F|^p + q^{-1}|G|^q$$

Then integrating both sides yields:

$$\|FG\|_1 \leq p^{-1}\|F\|_p^p + q^{-1}\|G\|_q^q = p^{-1} + q^{-1} = 1$$

Since $\|FG\|_1 = \|f\|_p^{-1}\|g\|_q^{-1}\|fg\|_1$ by construction, multiplying both sides of the equality above by $\|f\|_p\|g\|_q$ completes the proof. □

Proposition 1.5. *Minkowski Inequality: For a given $p > 1$, if $f, g \in L^p[a, b]$, then:*

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p$$

Proof. To do so, note that, via the triangle inequality:

$$|f + g|^p \leq |f + g||f + g|^{p-1} \leq |f||f + g|^{p-1} + |g||f + g|^{p-1}$$

We then note that $p^{-1} + q^{-1} = 1 \implies q = p(p-1)^{-1}$. Then, we observe that:

$$\int (|f + g|^{p-1})^q \, dx = \int |f + g|^p \, dx = \|f + g\|_p^p < \infty$$

as $f + g \in L^p[a, b]$, which implies $|f + g|^{p-1} \in L^q[a, b]$, and that $\left\||f + g|^{p-1}\right\|_q^q = \|f + g\|_p^p$. Then since $|f|, |g| \in L^p[a, b]$, Integrating the inequality above, applying Holder's inequality, and using $q^{-1} = p^{-1}(p - 1)$:

$$\begin{aligned}
\|f + g\|_p^p = &\leq \int_a^b |f||f + g|^{p-1} \, dx + \int_a^b |g||f + g|^{p-1} \, dx \\
&= \left\|f|f + g|^{p-1}\right\|_1 + \left\|g|f + g|^{p-1}\right\|_1 \\
&\leq \|f\|_p \left\||f + g|^{p-1}\right\|_q + \|g\|_p \left\||f + g|^{p-1}\right\|_q \\
&= (\|f\|_p + \|g\|_p)\|f + g\|_p^{pq^{-1}} = (\|f\|_p + \|g\|_p)\|f + g\|_p^{p-1}
\end{aligned}$$

If $\|f + g\|_p = 0$, then the inequality is clearly true. Assuming then that $\|f + g\|_p > 0$, dividing both sides of the inequality by $\|f + g\|_p^{p-1}$ completes the proof. □

The Holder and Minkowski inequalities are frequently used in L^p theory, and are quite important. As an example, those who read the notes on Linear Algebra may remember the discussion on metric spaces (if not, see the section on manifolds!). As it turns out, the Minkowski inequality is a sort of triangle inequality for the function $\|.\|_p$. Since it is also positive definite and symmetric by definition (remembering to use equivalence classes of functions), we get that each L^p space is in fact a metric space! This is partly also why L^2 is special, as the metric induced by the norm, $\|f - g\|_2$ in this case, also completes L^2 (referring to Cauchy completeness), forming a Hilbert space. Another reason why L^2 is special is because it is the only L^p space for which one can administer an inner product.

Introducing Fourier series:

To begin our journey towards Fourier series, it is important to first go through a few[4] definitions. We also usually will be using the inner product with $r = 2\pi$, and $a, b = -\pi, \pi$, though we may use the general inner product with r and a, b. Firstly:

Definition 1.5. *For $x \in \mathbb{R}$, a trigonometric polynomial of degree N is defined via:*

$$T(x) = a_0 + \sum_{n=1}^{N} (a_n \cos x + b_n \sin x) = \sum_{n=-N}^{N} c_n e^{inx}$$

We observe that all functions $\cos x$, $\sin x$, and e^{inx} are periodic, with period 2π. As $e^{inx} = \frac{\mathrm{d}}{\mathrm{d}x}\left((in)^{-1}e^{inx}\right)$, it follows that:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{imx}\, dx = \langle 1, e^{imx}\rangle = \langle e^{i0x}, e^{imx}\rangle = \begin{cases} 1 & m = 0 \\ 0 & m \neq 0 \end{cases}$$

Motivated by this, we take the product[5] $e^{-imx}e^{inx}$, from which we obtain:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i(n-m)x}\, dx = \langle e^{imx}, e^{inx}\rangle = \begin{cases} 1 & m = n \\ 0 & m \neq n \end{cases}$$

Since $T(x)$ is a trigonometric polynomial, we may take advantage of the integral being linear to obtain:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-imx}T(x)\, dx = \langle e^{imx}, T\rangle = \begin{cases} c_m & m \in [-N, N] \cap \mathbb{Z} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

which gives us quite a nice expression and analogy for the coefficients c_m. The coefficients represent projections of T along the vector e^{imx}, much like the coefficients a, b of some vector $\vec{x} = a\vec{x}_1 + b\vec{x}_2$ in $\mathbb{R}^3$ represent the projections of $\vec{x}$ along $\vec{x}_1, \vec{x}_2$ respectively. In general, the collection of functions $\{e^{inx} : n \in \mathbb{Z}\}$ form an orthonormal family[6] with respect to the administered inner product. There are various other examples of orthonormal families. For the inner product we are currently using, an equivalent orthonormal family is:

$$\phi_{j,n} = \begin{cases} 1 & j = n = 0 \\ \sqrt{2}\cos nx & j = 0, n \neq 0 \\ \sqrt{2}\sin nx & j = 1 \end{cases}$$

The $\sqrt{2}$ factor is a consequence of our use of $\frac{1}{2\pi}$ in the denominator. In particular, we will use $\{\phi_n\}$ to refer to an arbitrary orthonormal family (wherein we use the general inner product), $\{\phi_{j,n}\}$ to the trigonometric functions above, and the functions $\{e^{inx}\}$ themselves in the remaining case, unless further specificity is needed. With the above, we then define a trigonometric series as follows:

Definition 1.6. *The standard trigonometric series for a function f on $[-\pi, \pi]$ is defined via:*

$$f(x) \sim F(x) = \sum_{-\infty}^{\infty} c_n e^{inx}$$

If f is integrable on $[a, b]$, then the right hand side is called the standard Fourier series for f, and the standard Fourier coefficients c_n are given by eq. (1)*.*

[4] Just a few, we're almost there!

[5] This is often called Fourier's trick, as it is a rather neat way to find Fourier coefficients!

[6] One may wonder if this is a basis, in which case - stay tuned :)

In general, we do not need to restrict to $\{e^{inx}\}$, and can consider any orthonormal family as follows:

Definition 1.7. *For an integrable function f, its Fourier series (with respect to a family $\{\phi_n\}$) is defined via:*

$$f(x) \sim \sum_n c_n \phi_n$$

where c_n is defined to the n^{th} Fourier coefficient with respect to ϕ_n. The use of the notation $\sim$ is to indicate that we have not yet understood if the series converges to f in any way, and is used to refer to the fact that the Fourier coefficients in the series are obtained from f.

After introducing Fourier series, one is tempted to ask - when do these series converge to the function, and what meaningful information about the function can these series provide? These are all great questions to ask at this stage, and we will answer them in the next section, restricting occasionally to $f \in L^2[-\pi, \pi]$ to the orthonormal family $\{e^{inx}\}$.

Convergence of Fourier series:

With Fourier series defined, we focus on a particularly interesting property of the partial sums of a Fourier series:

Theorem 1.8. *For a function $f \sim \sum_n c_n \phi_n$, let $f_N = \sum_{n=1}^N c_n \phi_n$ refer to the N^{th} partial sum of the Fourier series of f. Then given any N^{th} degree trigonometric polynomial $g_N = \sum_{n=1}^N d_n \phi_n$, the following holds:*

$$\int_a^b |f - f_N|^2 \, dx \leq \int_a^b |f - g_N|^2 \, dx$$

In particular, this implies that:

$$\langle f - f_N, f - f_N \rangle \leq \langle f - g_N, f - g_N \rangle$$

with equality above holding iff $c_n = d_n$ for each n.

For intuition on this result, it is nice to refer to the inner product, and picture f, f_N and g_N as arrows in a $2D$ vector space, with f_N, g_N lying along the x-axis, as in fig . Then $\langle f - f_N, f - f_N \rangle$ gives us a notion of 'distance' between f and f_N, and similarly with g_N. The content behind this is essentially that this difference is minimized by a partial sum of the orthonormal family ϕ_n (vector along the x-axis), when the coefficients are those of f_N, corresponding to minimizing the height of the triangle formed. The $\frac{\pi}{2}$ angle corresponds to this, and also to the coefficients of f_N being projections of f with respect to the chosen orthonormal family of functions. With how this is structured, this oddly has a vibe akin to that of the Pythagorean theorem. In fact, let us supplement this with the following corollary:

Corollary 1.8.1. *Bessel's inequality: Substituting $c_n = d_n$ in theorem 1.8 and taking the limit as $n \to \infty$ yields:*

$$\langle f_N, f_N \rangle = \frac{1}{2\pi} \sum_{n=1}^{\infty} |c_n|^2 \leq \langle f, f \rangle = \frac{1}{2\pi} \int_a^b |f|^2 \, dx$$

In words, Bessel's inequality states that the sum of squares of every component along an orthonormal family $\{\phi_n\}$ can at most be the length of f. The result, especially in the equality case, furthers the connections we noted above with the Pythagorean theorem. One may naturally then ask - what

conditions are necessary and/or sufficient for equality to hold above? The answer in general is unfortunately that it depends quite heavily on our orthonormal basis. As it turns out, our standard orthonormal family $\{e^{inx}\}$ is rather special, and we can quite easily state when equality holds (and why we dived into L^p (and L^2 specifically) spaces)!

With this in mind, we will restrict our attention to 2π periodic functions f [7] with the standard orthonormal family $\{e^{inx}\}$ (or equivalently $\phi_{j,n}$), and develop some more machinery for [8] Fourier series, leading to equality in the corollary above. Let us take our first step with the following definition:

Definition 1.9. *We define the Dirichlet kernel via:*

$$D_N(x) = \sum_{-N}^{N} e^{inx}$$

D_N in fact has a more useful form, as we will show in the corollary below:

Corollary 1.9.1. $D_N(x) = \frac{\sin\left(N+\frac{1}{2}\right)x}{\sin\frac{x}{2}}$

Proof. Since the sum is finite, we may regroup terms. This leads to the following:

$$D_N(x) = \sum_{-N}^{N} e^{inx} = 1 + \sum_{n=1}^{N} (e^{inx} + e^{-inx}) = 1 + 2\sum_{n=1}^{N} \cos nx$$

Multiplying and dividing the second term by $\sin\frac{x}{2}$ yields [9]:

$$D_N(x) = 1 + \frac{2}{\sin\frac{x}{2}} \sum_{n=1}^{N} \left(\cos nx \sin\frac{x}{2}\right)$$

Using the trigonometric identity $\cos a \sin b = \frac{\sin(a+b) - \sin(a-b)}{2}$. with $a = nx$, $b = \frac{x}{2}$ yields, via a telescoping sum:

$$\begin{aligned} D_N(x) &= 1 + \frac{2}{\sin\frac{x}{2}} \sum_{n=1}^{N} \frac{\sin\left(n+\frac{1}{2}\right)x - \sin\left(n+\frac{1}{2}\right)x}{2} \\ &= 1 + \frac{\sin\left(N+\frac{1}{2}\right)x - \sin\frac{x}{2}x}{\sin\frac{x}{2}} \\ &= \frac{\sin\left(N+\frac{1}{2}\right)x}{\sin\frac{x}{2}} \end{aligned}$$

One should also verify that this holds for when $x = 2k\pi$, $k \in \mathbb{Z}$, as the derivation performed above was invalid for these values of x. This is done via:

$$\lim_{x\to 2k\pi} \frac{\sin\left(N+\frac{1}{2}\right)x}{\sin\frac{x}{2}} = \lim_{x\to 2k\pi} \frac{\left(N+\frac{1}{2}\right)\cos\left(N+\frac{1}{2}\right)x}{\frac{1}{2}\cos\frac{x}{2}} = 2N+1 = D_N(2k\pi)$$

as required. □

[7] If you are worried about losing too much generality here, it's a valid mathematical concern, but it usually isn't a problem in Physics, since either the functions or the region of interest are bounded.

[8] We'll also be dropping the standard prefix for convenience, but it's a good distinction to have in general.

[9] If you're worried about when this is 0 - good catch! It's addressed a little further on.

We introduce the Dirichlet kernel because it provides us with a particularly nice way to write the standard Fourier Series F:

Corollary 1.9.2. $f_N(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x-u)D_N(u)\ du$

Proof. A good portion of the proof follows from the below of chain of equalities:

$$\begin{aligned} f_N(x) = \sum_{-N}^{N} c_n e^{inx} = \sum_{-N}^{N} \langle e^{inx}, f\rangle e^{inx} &= \frac{1}{2\pi}\sum_{-N}^{N}\left(\int_{-\pi}^{\pi} e^{-int} f(t)\ dt\right)e^{inx} \\ &= \int_{-\pi}^{\pi}\left(\sum_{-N}^{N} e^{in(x-t)}\right) f(t)\ dt \\ &= \int_{-\pi}^{\pi} D_N(x-t)f(t)\ dt \end{aligned}$$

We then perform a change of variables, given by $u = x - t$, which transforms f_N into the following:

$$f_N(x) = -\int_{x+\pi}^{x-\pi} D_N(u)f(x-u)\ du = \int_{x-\pi}^{x+\pi} D_N(u)f(x-u)\ du = \int_{-\pi}^{\pi} D_N(u)f(x-u)\ du$$

where in the last equality we use the fact that f, D_N, and hence their product are 2π periodic. □

The Dirichlet kernel can thus characterize the N^{th} partial sum of f, f_N. The following theorem elaborates on this:

Theorem 1.10. *If f is Lipshitz continuous at x, that is, there exists constants $\delta > 0$, M finite with:*

$$|f(x) - f(t)| \leq M|x - t|$$

for all $t \in (x-\delta, x+\delta)$, then $f_n \to f$ at x, that is, the Fourier series $F = f$, our function itself at x

The Dirichlet Kernel may initially seem unrelated, but it is actually rather useful in the proof of the above theorem - to prove the conditions under which Fourier series converge pointwise. You might be wondering about how strict the Lipshitz condition is. In fact, it can easily be shown that a sufficient condition for f being Lipshitz is it being $C^1[-\pi, \pi]$, or continuously differentiable (Hint: Mean Value Theorem!). Since we usually deal with (at least piecewise) continuously differentiable functions in Physics, for our intents and purposes, Fourier series do indeed converge pointwise to f, as we would desire them to.

Having characterized pointwise convergence, we end off by stating 2 extremely fundamental results. Firstly, we will finally address when Bessel's inequality is an equality in the context of the $\{e^{inx}\}$ basis, via Parseval's theorem.

Theorem 1.11. *Parseval's theorem - If f, g are 2π periodic and Riemann integrable, then:*

$$\lim_{N\to\infty}\langle f - f_N, f - f_N\rangle = \lim_{N\to\infty}\frac{1}{2\pi}\int_{-\pi}^{\pi}|f - f_N|^2\ dx = 0$$

$$\langle g, f\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} g^* f\ dx = \sum_{-\infty}^{\infty} d_n^* c_n$$

In particular, if $f = g$:

$$\|f\|_2^2 = \langle f, f\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} |f|^2 \, dx = \sum_{-\infty}^{\infty} |c_n|^2$$

and if we restrict to square integrable functions (It can be shown $L^q[a,b] \subset L^p[a,b]$ when $q > p$, due to $[a,b]$ having finite measure), then the first equality implies

$$\lim_{N\to\infty} d(f, f_N) = 0$$

Since L^2 is complete, we also have the following:

Theorem 1.12. *The family e^{inx} form a complete orthonormal basis for $L^2[a,b]$. That is, for any $f \in L^2[a.b]$, $f = \sum_{-\infty}^{\infty} c_n e^{inx}$*

These are incredibly useful results, and are invoked almost all the time in solving PDEs (and thus in Quantum Mechanics as well)! There of course is a lot more to Fourier series than just the above, but the above provides a rather steady introduction, with some additional context via L^p spaces.

2 Quantum Entanglement

2.1 Tensors as Vector Spaces

One of the most basic things we do in mathematics is constructing new spaces out of old ones. If you open any introductory graduate textbook (scarily that is a thing) in mathematics, chances are it will open by constructing the integers from the naturals, the rationals from the integers, and the reals from the rationals. From the real numbers, we build $\mathbb{R}^n$ and $\mathbb{C}$, which turn out to be two sides of the same coin. If these constructions were completely arbitrary, they wouldn't be particularly interesting. But such a construction isn't useful; if we had no way to add or multiply what use would they be? What we want is to preserve the properties of the objects we're building off of, to keep their useful features and extend them to further applications.

In particular, we're interested here in quantum mechanics. In it, the "state space" of a particle is a vector space, in particular a Hilbert space. We'd ideally like a way, like we can in classical mechanics, to consider a system of two particles as one vector space or as a combination of the single vector spaces used for each particle. To do this, we use a construction called the *tensor product*.

Definition 2.1. *Let V, W be two vector spaces over $\mathbb{C}$. The tensor product of these spaces, denoted $V \otimes W$, is a vector space over $\mathbb{C}$ consisting of sums and scalar multiples of pairs $\vec{v} \otimes \vec{w}$, where $\vec{v} \in V, \vec{w} \in W$, which satisfies the following axioms.*

1. *For all $a \in \mathbb{C}$, $a(\vec{v} \otimes \vec{w}) = (a\vec{v}) \otimes (\vec{w}) = \vec{v} \otimes (a\vec{w})$*
2. *$\vec{v}_1 \otimes \vec{w} + \vec{v}_2 \otimes \vec{w} = (\vec{v}_1 + \vec{v}_2) \otimes \vec{w}$*
3. *$\vec{v} \otimes \vec{w}_1 + \vec{v} \otimes \vec{w}_2 = \vec{v} \otimes (\vec{w}_1 + \vec{w}_2)$*

An important thing to note here : this construction is **not** just doing element-wise operations on $V \times W$[10]. For example, we have in general for tensor products that

[10] Such a construction does exist, and is called the direct sum

$$\vec{v}\otimes\vec{w}+\vec{v}'\otimes\vec{w}'\neq(\vec{v}+\vec{v}')\otimes(\vec{w}+\vec{w}')$$
$$a(\vec{v}\otimes\vec{w})\neq(a\vec{v})\otimes(a\vec{w})$$

Both of these inequalities would in fact be equalities if we just did element-wise operations. We choose instead to use the tensor product for a very particular property it has.

Theorem 2.2. *Let V, W, Z we vector spaces over $\mathbb{C}$. $V\otimes W$ is the unique vector space such that for any bilinear map $\varphi: V\times W\to Z$ (that is a map that's linear in both arguments) there exists a unique linear map $\psi: V\otimes W\to Z$ such that $\varphi=\psi\circ\iota$, where $\iota: V\times W\to V\otimes W$ is the natural map $\iota(\vec{v},\vec{w})=\vec{v}\otimes\vec{w}$.*

The proof of this is very much in the realm of abstract algebra, and is not particularly enlightening for our purposes, so we won't go over it here. However, we can see some of its consequences, and potential why this property is so useful.

Corollary 2.2.1. *Let V, W be vector spaces over $\mathbb{C}$, and $A: V\to V, B: W\to W$ linear maps. Then there exists a unique linear map $A\otimes B: V\otimes W\to V\otimes W$ such that $(A\otimes B)(\vec{v}\otimes\vec{w})=(A\vec{v})\otimes(B\vec{w})$.*

Proof. Let $A\times B: V\times W\to V\otimes W$ be the bilinear map given by $(A\times B)(\vec{v},\vec{w})=(A\vec{v})\otimes(B\vec{w})$. By the uniqueness property of tensor spaces, there exists a unique linear map $C: V\otimes W\to Z$ such that $A\times B=C\circ\iota$. This would be the unique linear map such that

$$(A\times B)(\vec{v},\vec{w})=(C\circ\iota)(\vec{v},\vec{w})\implies(A\vec{v})\otimes(B\vec{w})=C(\vec{v}\otimes\vec{w})$$

so the map C is the linear map $A\otimes B$ we claimed existed. □

This is a great first step; it gives us a canonical way of combining a pair of linear operators in each vector space to produce a linear operator in the tensor space. All we need to apply this to quantum mechanics then is a way to extend the inner product. First though, we need to recall some facts about inner products (definition 1.3), and dual spaces from last term, along with some results on bases. The reader may recall the definition of a dual space.

Definition 2.3. *Let V be a vector space over a field $\mathbb{C}$. A linear function $f: V\to\mathbb{C}$ is called a linear functional. The set V' of all linear functionals on V is a vector space in itself, and referred to as the dual space of V.*

Dual spaces, naively, seem to give us another way of representing inner products. Indeed, the map $\langle\vec{v},\cdot\rangle:\vec{u}\mapsto\langle\vec{v},\vec{u}\rangle$ is a linear functional, for any fixed $\vec{v}$. The problem is that there's no canonical way to pair linear functionals with inner products, that is unless we're in a Hilbert space (a particularly nice inner product space for our purposes). In this case, we get the following result

Theorem 2.4. *Suppose $\mathcal{H}$ is a Hilbert space over $\mathbb{C}$. Then there exists a canonical bijective map $f:\mathcal{H}\to\mathcal{H}'$[11] with the following property, for any $a\in\mathbb{C},\vec{v},\vec{u}\in\mathcal{H}$ (this property is called anti-linearity)*

$$f(a\vec{v}+\vec{u})=a^*f(\vec{v})+f(\vec{u})$$

In particular, this map is given by

$$f(\vec{v})(\vec{u})=\langle\vec{v},\vec{u}\rangle$$

Again, the proof of this is beyond the scope of these notes. The point of this theorem is not so much the existence of the map f, that would work in any inner product space, but that it's *bijective*. This is what gives us our perfect correspondence between inner products and dual vectors[12], and

[11] This is technically not true, and the map is instead bijective between $\mathcal{H}$ and continuous dual vectors. For our purposes, this difference isn't too important, since physics tends to assume continuity anyways.

[12] Also known as the dual correspondence in QM

it's what allows us to extend our inner products in a consistent way in quantum mechanics.

I did mention earlier that we need a result on bases for inner products; we prove that now.

Theorem 2.5. *Let V, W be vector spaces over $\mathbb{C}$ with bases $\{\vec{b}_i\}_{i=1}^n, \{\vec{c}_i\}_{i=1}^m$. Then the set $\{\vec{b}_i \otimes \vec{c}_j\}_{i,j=1}^{n,m}$ is a basis for $V \otimes W$.*

Proof. We first prove that its spanning, for which it suffices to prove that any tensor of the form $\vec{v} \otimes \vec{w} \in V \otimes W$ can be expressed as a linear sum of elements in the set. Since $\{\vec{b}_i\}_{i=1}^n, \{\vec{c}_i\}_{i=1}^m$ are bases, $\exists \alpha_i, \beta_j \in \mathbb{C}$ such that $\vec{v} = \sum_{i=1}^n \alpha_i \vec{b}_i$ and $\vec{w} = \sum_{j=1}^m \beta_j \vec{c}_j$. Thus,

$$\vec{v} \otimes \vec{w} = \Big(\sum_{i=1}^n \alpha_i \vec{b}_i\Big) \otimes \Big(\sum_{j=1}^m \beta_j \vec{c}_j\Big) = \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j (\vec{b}_i \otimes \vec{c}_j)$$

as required. The proof of linear independence is more complicated and involves some heavy usage of dual bases, so we'll avoid it here. □

Finally, we can get to the point.

Theorem 2.6. *Let V, W be inner product spaces over $\mathbb{C}$. Then the map $\langle \cdot, \cdot \rangle : (V \otimes W) \times (V \otimes W) \to \mathbb{C}$ given by*

$$\langle \vec{u} \otimes \vec{w}, \vec{v} \otimes \vec{x} \rangle = \langle \vec{u}, \vec{v} \rangle \langle \vec{w}, \vec{x} \rangle$$

with linearity in the second argument and anti-linearity in the first is a well-defined inner product on $V \otimes W$.

Note that there's an abuse of notation here; we use the angled brackets for all inner products, even ones that operate on different spaces.

Proof. Existence follows from the previous theorem on the bases of tensor spaces, so we just need to show that this is an inner product. Indeed, we get

$$\langle \vec{v} \otimes \vec{w}, \vec{v} \otimes \vec{w} \rangle = \langle \vec{v}, \vec{v} \rangle \langle \vec{w}, \vec{w} \rangle$$

Which is always positive, and zero if and only if $\vec{v}$ or $\vec{w}$ is zero. But any vector tensor with the zero vector is a zero vector (to check this, look at $\vec{v} \otimes \vec{0} + \vec{v} \otimes \vec{w} = \vec{v} \otimes \vec{w}$), so the first property of inner products is satisfied. For the second, we note that

$$\langle \vec{u} \otimes \vec{w}, \vec{v} \otimes \vec{x} \rangle = \langle \vec{u}, \vec{v} \rangle \langle \vec{w}, \vec{x} \rangle = (\langle \vec{v}, \vec{u} \rangle \langle \vec{x}, \vec{w} \rangle)^* = \langle \vec{v} \otimes \vec{x}, \vec{u} \otimes \vec{w} \rangle^*$$

The third is simply part of our definition. □

There's one final property that we need, which again will not be proven.

Theorem 2.7. *The tensor product of two Hilbert spaces, with the inner product of the above theorem, is a Hilbert space.*

2.2 Composite Systems and Entanglement

Interesting note about tensor product Hilbert spaces - the number of basis vectors of the composite Hilbert space (and hence its dimension) is given by $\dim(\mathcal{H}_{\text{composite}}) = \prod_{i=1}^{n} \dim(\mathcal{H}_i)$ - this is exponential in the number of systems being composed. For example for n spin-1/2 particles (with $\dim(\mathcal{H}_i) = 2$), the dimension of the composite Hilbert space is $\dim(\mathcal{H}^n) = \prod_{i=1}^{n} 2 = 2^n$. For $n = 300$ we have $\dim(\mathcal{H}^n) \sim 10^{90}$ which already exceeds the number of atoms in the observable universe ($10^{78} - 10^{82}$). This high dimensionality is a reason[13] for why quantum systems are hard to simulate classically.
Let us give a concrete example of $n = 2$ spin-1/2 particles. An ONB for the Hilbert spaces $\mathcal{H}_A, \mathcal{H}_B$ is $\{|\uparrow\rangle, |\downarrow\rangle\}$, so the basis states of the composite Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B$ are:

$$\begin{aligned} |\uparrow\uparrow\rangle_{AB} &\coloneqq |\uparrow\rangle_A \otimes |\uparrow\rangle_B \\ |\uparrow\downarrow\rangle_{AB} &\coloneqq |\uparrow\rangle_A \otimes |\downarrow\rangle_B \\ |\downarrow\uparrow\rangle_{AB} &\coloneqq |\downarrow\rangle_A \otimes |\uparrow\rangle_B \\ |\downarrow\downarrow\rangle_{AB} &\coloneqq |\downarrow\rangle_A \otimes |\downarrow\rangle_B \end{aligned} \tag{2}$$

And so:

$$\mathcal{H}_{AB} = \text{span}(\{|\uparrow\uparrow\rangle_{AB}, |\uparrow\downarrow\rangle_{AB}, |\downarrow\uparrow\rangle_{AB}, |\downarrow\downarrow\rangle_{AB}\}) = \{\alpha|\uparrow\uparrow\rangle_{AB} + \beta|\uparrow\downarrow\rangle_{AB} + \gamma|\downarrow\uparrow\rangle_{AB} + \delta|\downarrow\downarrow\rangle_{AB} : \alpha, \beta, \gamma, \delta \in \mathbb{C}\} \tag{3}$$

A question we now ask - are all states in a composite Hilbert space able to be written as a tensor product of states of the individual subsystems (as the notation $\mathcal{H}_{AB} = \mathcal{H}_A \otimes \mathcal{H}_B$ might suggest)? The answer is a *no* - this leads to our definition of entanglement, which will play a key role in the entire discussion of this chapter:

Definition: Entanglemement

A pure quantum state $|\Psi\rangle$ in a composite Hilbert space $\mathcal{H} = \bigotimes_{i=1}^{n} \mathcal{H}_i$ is *entangled* if it cannot be written as the tensor product of states from the subsystems $\mathcal{H}_1, \ldots \mathcal{H}_n$, i.e.:

$$|\Psi\rangle \neq |\psi_1\rangle_1 \otimes |\psi_2\rangle_2 \otimes \ldots \otimes |\psi_n\rangle_n \tag{4}$$

for *any* choice of states $|\psi_i\rangle_i \in \mathcal{H}_i$.

For the case of $n = 2$ subsystems, we have bipartite entanglement defined as:

Definition: Bipartite entanglement

Let $\mathcal{H}_A, \mathcal{H}_B$ be Hilbert spaces and define the composite Hilbert space $\mathcal{H}_{AB} = \mathcal{H}_A \otimes \mathcal{H}_B$. A pure state $|\Psi\rangle_{AB} \in \mathcal{H}_{AB}$ is *entangled* if:

$$|\Psi\rangle_{AB} \neq |\psi\rangle_A \otimes |\phi\rangle_B \tag{5}$$

for any choice of local states $|\psi\rangle_A \in \mathcal{H}_A, |\phi\rangle_B \in \mathcal{H}_B$.

[13] There are some subtleties here; specifically, we require an extremely large number of parameters to describe highly entangled states (entanglement to be defined extremely shortly). Product (i.e. unentangled) states, are efficiently simulable because we may describe the subsystems individually, and therefore the whole state efficiently. The argument is actually a layer more nuanced than this, because certain types of entangled states (stabilizer states - see the Gottesman-Knill Theorem) are efficiently simulable. But this is far beyond the scope of this course.

A specific example of bipartite entanglement is given by the Bell state $|B_{11}\rangle$ (also called the singlet state - this name for it will perhaps become clearer after we begin our study of addition of angular momenta):

$$|B_{11}\rangle = \frac{|\uparrow\rangle_A \otimes |\downarrow\rangle_B - |\downarrow\rangle_A \otimes |\uparrow\rangle_B}{\sqrt{2}} \tag{6}$$

It is a useful exercise to use the definition of entanglement given above to prove that the above Bell state is indeed entangled (hint: try a proof by contradiction).

2.3 Optional: a review of quantum measurement

Mathematical object of interest: projectors

Definition: Projectors

A linear operator Π is a *projector* if it satisfies:

$$\Pi^2 = \Pi^\dagger = \Pi. \tag{7}$$

Below are examples of projectors in matrix representations:

$$\Pi_1 \cong \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \Pi_2 \cong \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{8}$$

Π_1 is a rank 1 projector, while Π_2 is rank $2 > 1$. Why we call an operator with the properties in Eq. (7) a projector might not be obvious, but the nomenclature is elucidated by the above examples. A projector projects a state into a lower-dimensional subspace of the Hilbert space. Π_1 has the property of taking a three-dimensional vector and projecting it into a 1-dimensional subspace, while Π_2 has the property of taking a three-dimensional vector and projecting it into a 2-dimensional subspace. $\mathbb{I}$ is a projector (though a trivial one), and is a projection from a space to itself. In Fig. 1 we visualize the action of Π_1, Π_2 for the case when our vector space is $\mathbb{R}^3$ (but one should keep in mind that this is for the sake of intuition, and the Hilbert spaces we use in quantum mechanics are, of course, complex).

Now let's return to the bra-ket formalism and what projectors look like in this abstract setting. First, recall that we can write an observable A in the form:

$$A = \sum_i a_i |a_i\rangle\langle a_i| \tag{9}$$

where $|a_i\rangle$ is the eigenstate of A corresponding to the eigenvalue a_i, and $\{|a_i\rangle\}_i$ is an ONB. In the non-degenerate case, each of the a_is are distinct. However, in general degenerate eigenvalues (where $a_i = a_j$ for some i, j) are possible, and the current form of the expression does not make this particularly clear. With our knowledge of projectors, let us now rewrite the above as:

$$A = \sum_a a\Pi_a \tag{10}$$

where each of the as are *distinct* eigenvalues of A, and

$$\Pi_a = \sum_{a_i = a} |a_i\rangle\langle a_i| \tag{11}$$

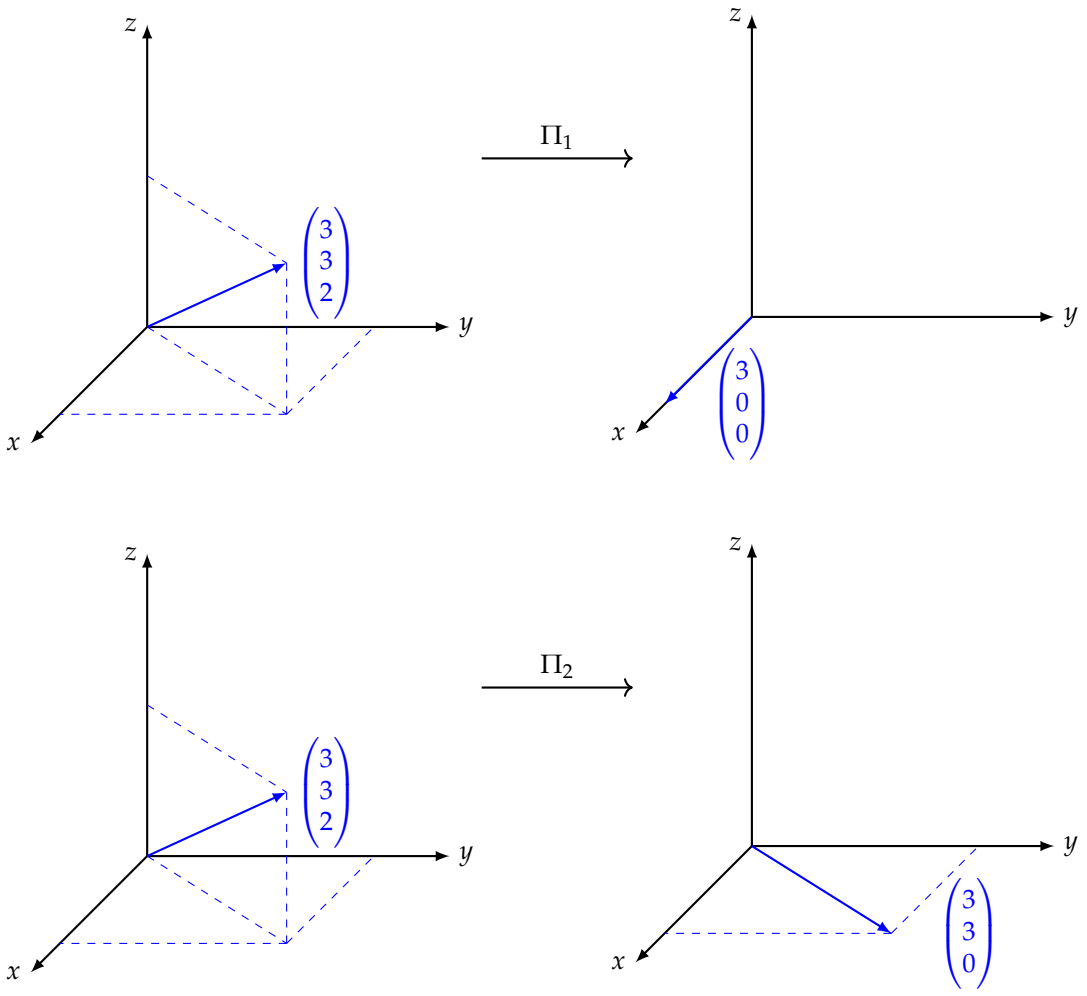

Figure 1: Visualization of the action of projectors Π_1, Π_2 (as defined in Eq. (8)) on a vector in $\mathbb{R}^3$. Π_1 can be visualized as projecting the given vector onto the one-dimensional subspace that is the x-axis subspace; preserving the x-component of the vector, and nullifying the y and z-components. Π_2 can be visualized as projecting the given vector onto the two-dimensional subspace that is the xy-plane; preserving the x and y components of the vector and nullifying the z-component.

is the *projector onto the eigenvalue-a subspace*. Let us verify that these are indeed projectors. First, we verify that they are Hermitian:

$$\Pi_a^\dagger = \left(\sum_{a_i=a} |a_i\rangle\langle a_i|\right)^\dagger = \sum_{a_i=a} (|a_i\rangle\langle a_i|)^\dagger = \sum_{a_i=a} |a_i\rangle\langle a_i| = \Pi_a. \tag{12}$$

where in the second-to-last equality we use that $(|a\rangle\langle b|)^\dagger = |b\rangle\langle a|$ (which follows immediately from the definition of the Hermitian adjoint; the proof is left to the reader!). Next, we show that they are idempotent (that is, they square to themselves):

$$\Pi_a^2 = \left(\sum_{a_i=a} |a_i\rangle\langle a_i|\right)^2 = \sum_{a_i=a}\sum_{a_j=a} |a_i\rangle \langle a_i|a_j\rangle \langle a_j| = \sum_{a_i=a}\sum_{a_j=a} |a_i\rangle\langle a_j| \,\delta_{ij} = \sum_{a_i=a} |a_i\rangle\langle a_i| = \Pi_a. \tag{13}$$

So they are indeed projectors! In this form, we have decomposed the observable A into the parts associated with each eigenvalue in a clear way. These projectors have some properties of note, described in the theorem below.

Proposition

Let $\{\Pi_a\}_a$ be the set of projectors associated to an observable A (with $\Pi_a = \sum_{a_i=a} |a_i\rangle\langle a_i|$ being the projector onto the eigenvalue-a subspace). These projectors are mutually orthogonal:

$$\Pi_i \Pi_j = \delta_{ij} \Pi_i \tag{14}$$

and are complete:

$$\sum_a \Pi_a = \mathbb{I}. \tag{15}$$

Proof. The idempotency of projectors covers the $i = j$ case in Eq. (14), and if $i \neq j$, then the expression is zero as eigenvectors of an observable A corresponding to distinct eigenvalues are orthogonal. The completeness relation is merely a restatement of the resolution of the identity in terms of projectors. □

We are now ready to state our axiom of quantum measurement.

Axiom: Quantum measurement

Let $A = \sum_a a\Pi_a$ be the observable (a Hermitian operator) being measured, where the as are the eigenvalues of A and $\{\Pi_a = \sum_{a_i=a} |a_i\rangle\langle a_i|\}_a$ are the associated projectors onto the eigenvalue-a subspaces. Let $|\psi\rangle$ be the pre-measurement state.

Dirac postulate: If outcome a is measured, then the post measurement state is given by:

$$|\psi\rangle \mapsto \frac{1}{\sqrt{\langle\psi|\,\Pi_a\,|\psi\rangle}} \Pi_a\,|\psi\rangle\,. \tag{16}$$

Born rule: The probability of measuring outcome a is given by:

$$p(a) = \langle\psi|\,\Pi_a\,|\psi\rangle\,. \tag{17}$$

Example: spin-1 particle.

$$S_z = \hbar(|+\rangle\langle+| - |-\rangle\langle-|) \tag{18}$$

So then

$$S_z^2 = \hbar^2 (|+\rangle\langle+| + |-\rangle\langle-|) \tag{19}$$

has a degenerate eigenvalue, with both $|+\rangle$ and $|-\rangle$ being eigenstates with eigenvalue $+\hbar^2$. To deal with this degeneracy, we can use our new projector formalism of measurement. The projector corresponding to the $\hbar^2$ subspace is given by:

$$\Pi_{\hbar^2} = |+\rangle\langle+| + |-\rangle\langle-| \tag{20}$$

while the projector corresponding to the eigenvalue 0 subspace is given by:

$$\Pi_0 = |0\rangle\langle 0| . \tag{21}$$

So, if we wanted to find the probability of measuring $S_z^2 = \hbar^2$ given a pre-measurement state $|\psi\rangle$, the probability would be given by:

$$p(\hbar^2) = \langle\psi| \Pi_{\hbar^2} |\psi\rangle = |\langle+|\psi\rangle|^2 + |\langle-|\psi\rangle|^2 \tag{22}$$

and the post measurement state would be given by:

$$|\psi\rangle \mapsto \frac{1}{\sqrt{\langle\psi| \Pi_{\hbar^2} |\psi\rangle}} \Pi_{\hbar^2} |\psi\rangle = \frac{1}{\sqrt{|\langle+|\psi\rangle|^2 + |\langle-|\psi\rangle|^2}} (\langle+|\psi\rangle |+\rangle + \langle-|\psi\rangle |-\rangle) . \tag{23}$$

2.4 Properties of the Bell state

Let's explore some properties of this state - let us begin by looking at what happens when we measure one of the two spins. In general, if we perform an operation on one subsystem (represented by the application of an operator A) of a composite system while doing nothing to the other parts, we can represent this by the composite operator consisting of applying A to the specific subsystem, tensored with the identity operation $\mathbb{I}$ on the other subsystems. In our case, we consider operators of the form $\Pi_A \otimes \mathbb{I}_B$ where Π_A is a projector acting on the first spin.
Let's suppose we measure the first spin in the $\{|\uparrow\rangle_A, |\downarrow\rangle_A\}$ basis. From the Born rule we find:

$$\begin{aligned} p(\uparrow) &= \langle B_{11}| \Pi_{\uparrow,A} \otimes \mathbb{I}_B |B_{11}\rangle \\ &= \frac{\langle\uparrow| \Pi_\uparrow |\uparrow\rangle \langle\downarrow| \mathbb{I} |\downarrow\rangle + \langle\downarrow| \Pi_\uparrow |\downarrow\rangle \langle\uparrow| \mathbb{I} |\uparrow\rangle - \langle\uparrow| \Pi_\uparrow |\downarrow\rangle \langle\downarrow| \mathbb{I} |\uparrow\rangle - \langle\downarrow| \Pi_\uparrow |\uparrow\rangle \langle\uparrow| \mathbb{I} |\downarrow\rangle}{2} \\ &= \frac{1 \cdot 1 + 0 \cdot 1 - 0 \cdot 0 - 0 \cdot 0}{2} \\ &= \frac{1}{2} \end{aligned} \tag{24}$$

and analogously $p(\downarrow) = \frac{1}{2}$. The Dirac postulate tells us that if we measure spin-up, then the post-measurement state is:

$$|B_{11}\rangle \to \frac{\Pi_{\uparrow,A} \otimes \mathbb{I}_B |B_{11}\rangle}{\sqrt{\langle B_{11}| \Pi_{\uparrow,A} \otimes \mathbb{I}_B |B_{11}\rangle}} = \frac{1}{\sqrt{\frac{1}{2}}} \frac{\Pi_{\uparrow,A} |\uparrow\rangle_A \otimes \mathbb{I}_B |\downarrow\rangle_B - \Pi_{\uparrow,A} |\downarrow\rangle_A \otimes \mathbb{I}_B |\uparrow\rangle_B}{\sqrt{2}} = |\uparrow\rangle_A \otimes |\downarrow\rangle_B \tag{25}$$

Analogously, it can be shown that if we measure the first spin to be spin-down, then the post-measurement state is:

$$|B_{11}\rangle \to |\downarrow\rangle_A \otimes |\uparrow\rangle_B \tag{26}$$

We note two things - it seems as though when we measure the first spin in the S_z eigenbasis that we have a 50/50 probability of measuring the first spin to be up or down, and that the second spin after the measurement points in the direction opposite that of which the first spin was measured to be.
Perhaps this interesting result is just a consequence of our choice of measurement basis $\{|\uparrow\rangle, |\downarrow\rangle\}$. Let us try then measuring in the S_x eigenbasis of $\{|\rightarrow\rangle = \frac{|\uparrow\rangle+|\downarrow\rangle}{\sqrt{2}}, |\leftarrow\rangle = \frac{|\uparrow\rangle-|\downarrow\rangle}{\sqrt{2}}\}$. Using that $|\uparrow / \downarrow\rangle = \frac{|\rightarrow\rangle \pm |\leftarrow\rangle}{\sqrt{2}}$, we can rewrite the $|B_{11}\rangle$ state in terms of the S_x eigenstates as:

$$\begin{aligned} |B_{11}\rangle &= \frac{\frac{|\rightarrow\rangle_A+|\leftarrow\rangle_A}{\sqrt{2}} \otimes \frac{|\rightarrow\rangle_B-|\leftarrow\rangle_B}{\sqrt{2}} - \frac{|\rightarrow\rangle_A-|\leftarrow\rangle_A}{\sqrt{2}} \otimes \frac{|\rightarrow\rangle_B+|\leftarrow\rangle_B}{\sqrt{2}}}{\sqrt{2}} \\ &= \frac{|\leftarrow\rangle_A \otimes |\rightarrow\rangle_B - |\rightarrow\rangle_A \otimes |\leftarrow\rangle_B}{\sqrt{2}} \end{aligned} \tag{27}$$

Up to a (physically irrelevant) global minus sign, the form of $|B_{11}\rangle$ expressed in terms of S_x eigenstates is identical to $|B_{11}\rangle$ expressed in terms of S_z eigenstates. So, if we were to measure the first spin in the S_x eigenbasis, just as before, we would find that we would have 50/50 probability of measuring spin right or spin left, and the post-measurement state would have the unmeasured spin pointing in the opposite direction as the measured one.
In fact we could go through with the above calculation for an arbitrary measurement basis, and find the same result.

Proposition: $|B_{11}\rangle$ is non-local and anti-correlated in every direction

Consider the Bell state $|B_{11}\rangle$:

$$|B_{11}\rangle = \frac{|\uparrow\rangle_A \otimes |\downarrow\rangle_B - |\downarrow\rangle_A \otimes |\uparrow\rangle_B}{\sqrt{2}} \tag{28}$$

and consider an arbitrary ONB (for a spin-1/2 particle):

$$\mathcal{B}(\alpha, \beta) = \{|\mathbf{r}_{\alpha,\beta}\rangle := \alpha |\uparrow\rangle + \beta |\downarrow\rangle, |\bar{\mathbf{r}}_{\alpha.\beta}\rangle := \beta^* |\uparrow\rangle - \alpha^* |\downarrow\rangle\} \tag{29}$$

where $\alpha, \beta \in \mathbb{C}$ and $|\alpha|^2 + |\beta|^2$. Then:

1. $|B_{11}\rangle$ has no local properties; that is, whatever parameters $\alpha, \beta \in \mathbb{C}$ the measurement of particle A in the basis $\mathcal{B}(\alpha, \beta)$ leads to a 50/50 distribution of outcome.

2. The measurement of particle A leads to the post-measurement states:

$$\begin{aligned} \text{outcome “+”:}\ & |B_{11}\rangle \to |\mathbf{r}_{\alpha,\beta}\rangle_A \otimes |\bar{\mathbf{r}}_{\alpha,\beta}\rangle_B \\ \text{outcome “-”:}\ & |B_{11}\rangle \to |\bar{\mathbf{r}}_{\alpha,\beta}\rangle_A \otimes |\mathbf{r}_{\alpha,\beta}\rangle_B \end{aligned} \tag{30}$$

   That is, after measurement, the spin states of particles A/B are perfectly anti-correlated, irrespective of the measurement outcome and measurement basis.

The above demonstrates how quantum entanglement can give rise to "stronger-than-classical" correlations. In classical mechanics, it is possible for measurements to be correlated in certain measurement bases, but not all.

Proof. Left as a homework exercise. □

2.5 The No-Cloning Theorem

Consider now a thought experiment where, upon preparing two spin-1/2 particles in a Bell state $|B_{11}\rangle$, we flew out one pair to the moon while we kept one on Earth. Suppose we were to measure the particle on Earth; then the formalism of quantum mechanics would tell us that the particle on the moon would instantaneously collapse to the spin state pointing opposite to that which was measured on Earth. This phenomenon, coined "spooky action at a distance" by Einstein, seems quite troubling; it seems as information is travelling faster than the speed of light when the measurement is made! Could we harness this quantum-mechanical effect to communicate superliminally (and therefore - break the laws of special relativity)?
There is no need to fear, as the answer is no, as we will prove this in full generality in the latter half of this chapter. However, it will be of interest to consider a specific example of a protocol which does not work, as the reason for its failure is highly interesting. The (non) protocol for superluminal communication goes as follows:

(Non)-Protocol: Superluminal communication

Objective: Transmit one bit b of information (where $b \in \mathbb{Z}_2 = \{0, 1\}$) from A to B.
Protocol:

1. In advance of the actual communication, the parties share a Bell state $|B_{11}\rangle$ between themselves.

2. If $b = 0$, then A measures their particle of $|B_{11}\rangle$ in the eigenbasis of S_z, i.e. in $\mathcal{B}_z = \{|\uparrow\rangle, |\downarrow\rangle\}$. If $b = 1$, then A measures their particle of $|B_{11}\rangle$ in the eigenbasis of S_x, i.e. in $\mathcal{B}_x = \{|\rightarrow\rangle, |\leftarrow\rangle\}$.

3. Party B copies their state a large number of times, i.e.:

$$\begin{aligned} |\uparrow\rangle &\rightarrow |\uparrow\rangle \otimes |\uparrow\rangle \otimes |\uparrow\rangle \otimes \dots \\ |\downarrow\rangle &\rightarrow |\downarrow\rangle \otimes |\downarrow\rangle \otimes |\downarrow\rangle \otimes \dots \\ |\rightarrow\rangle &\rightarrow |\rightarrow\rangle \otimes |\rightarrow\rangle \otimes |\rightarrow\rangle \otimes \dots \\ |\leftarrow\rangle &\rightarrow |\leftarrow\rangle \otimes |\leftarrow\rangle \otimes |\leftarrow\rangle \otimes \dots \end{aligned} \tag{31}$$

4. B identifies the state received in the transmission, through the measurement of the multiple copies (doing a large number of measurements in the S_z and S_x eigenbases, until they are confident that the state transmitted is one of $|\uparrow\rangle / |\downarrow\rangle$ or $|\rightarrow\rangle / |\leftarrow\rangle$). If the state received was $|\uparrow\rangle$ or $|\downarrow\rangle$, B outputs $b = 0$. If the state received was $|\leftarrow\rangle$ or $|\rightarrow\rangle$, B outputs $b = 1$.

Where does this protocol fail? The first and second steps are fine; there is no problem with creating a Bell state, then taking them far apart from each other, and measuring one of the two spins in a particular basis. The fourth step is also fine; if we are given a large number of identical states, by

doing a sufficiently large number of measurements (in different bases), we can be confident about the state that we have (and hence obtain the correct output). The failure of the protocol comes in step 3 - namely, *unknown quantum states cannot be copied.* This is known as the *No-cloning theorem*, which is simple to prove but has profound implications.

Theorem: No-cloning

Let $|\psi\rangle \in \mathcal{H}_d$ be an unknown quantum state, and let $|0\rangle \in \mathcal{H}_d$ be a fixed known quantum state[a]. Then, the copying (cloning) operation C defined by:

$$C : |\psi\rangle \otimes |0\rangle \to |\psi\rangle \otimes |\psi\rangle , \quad \forall |\psi\rangle \in \mathcal{H}_d \tag{32}$$

cannot be realized in quatnum mechanics.

[a]Note: $|0\rangle$ is *not* the null ket

The proof of the above theorem rests on the linearity of quantum mechanics.

Lemma: Linearity of quantum mechanics

Unitary evolution according to the Schrödinger equation:

$$i\hbar \frac{\partial}{\partial t} |\psi(t)\rangle = H |\psi(t)\rangle \tag{33}$$

is linear; that is, if $|\psi_1(t)\rangle$ and $|\psi_2(t)\rangle$ solve Eq. (33), then:

$$|\psi_{a,b}(t)\rangle = a |\psi_1(t)\rangle + b |\psi_2(t)\rangle \tag{34}$$

is also a solution to Eq. (33). In addition, measurement according to the Dirac projection postulate is linear up to normalization:

$$\Pi_i(a |\psi_1(t)\rangle + b |\psi_2(t)\rangle) = a\Pi_i |\psi_1(t)\rangle + b\Pi_i |\psi_2(t)\rangle . \tag{35}$$

Proof. Plugging in $|\psi_{a,b}(t)\rangle$ in to the LHS of Eq. (33), we find:

$$\begin{aligned} i\hbar \frac{\partial}{\partial t} |\psi_{a,b}(t)\rangle &= i\hbar \frac{\partial}{\partial t}(a |\psi_1(t)\rangle + b |\psi_2(t)\rangle) \\ &= ai\hbar \frac{\partial}{\partial t} |\psi_1(t)\rangle + bi\hbar \frac{\partial}{\partial t} |\psi_2(t)\rangle \\ &= aH |\psi_1(t)\rangle + bH |\psi_2(t)\rangle \\ &= H(a |\psi_1(t)\rangle + b |\psi_2(t)\rangle) \\ &= H |\psi_{a,b}(t)\rangle \end{aligned} \tag{36}$$

where in the second equality we use the linearity of the derivative, in the third equality we use that $|\psi_1(t)\rangle , |\psi_2(t)\rangle$ are individually solutions to Eq. (33), and in the fourth equality we use the linearity of the Hamiltonian operator H. We have thus shown that $|\psi_{a,b}(t)\rangle$ is also a solution to Eq. (33).
Next, the Dirac projection postulate tells us that:

$$|\psi\rangle \to \frac{\Pi_i |\psi\rangle}{\sqrt{\langle\psi| \Pi_i |\psi\rangle}} \tag{37}$$

So neglecting the normalization factor:

$$|\psi\rangle \to \propto \Pi_i |\psi\rangle \tag{38}$$

Therefore since projectors are linear:

$$\begin{aligned} |\psi_{a,b}\rangle \to &\propto \Pi_i |\psi_{a,b}\rangle \\ &= a\Pi_i |\psi_1\rangle + b\Pi_i |\psi_2\rangle \end{aligned} \tag{39}$$

which proves the claim. □

Having made this observation about linearity, we can proceed to the proof of the no-cloning theorem.

Proof. Assume for the sake of contradiction that C exists. We have $|0\rangle \in \mathcal{H}_d$ the fixed/reference quantum state, and let $|1\rangle \in \mathcal{H}_d$ be some state orthogonal to $|0\rangle$. By assumption, C clones $|0\rangle$ and $|1\rangle$, so:

$$\begin{aligned} C(|0\rangle \otimes |0\rangle) &\propto |0\rangle \otimes |0\rangle \\ C(|1\rangle \otimes |0\rangle) &\propto |1\rangle \otimes |1\rangle \end{aligned} \tag{40}$$

furthermore, defining $|+\rangle = \frac{|0\rangle+|1\rangle}{\sqrt{2}}$, C should also clone this state:

$$C(|+\rangle \otimes |0\rangle) \propto |+\rangle \otimes |+\rangle = \frac{|0\rangle \otimes |0\rangle + |0\rangle \otimes |1\rangle + |1\rangle \otimes |0\rangle + |1\rangle \otimes |1\rangle}{2} \tag{41}$$

However, since C is some quantum-mechanical operation, it must be linear (as we have shown that evolution in quantum mechanics is linear in general). Therefore:

$$\begin{aligned} C(|+\rangle \otimes |0\rangle) &= \frac{1}{\sqrt{2}} \left[C(|0\rangle \otimes |0\rangle) + C(|1\rangle \otimes |0\rangle)\right] \\ &= \frac{a}{\sqrt{2}} |0\rangle \otimes |0\rangle + \frac{b}{\sqrt{2}} |1\rangle \otimes |1\rangle \end{aligned} \tag{42}$$

where in the second equality we invoke Eq. (40), and a, b are the proportionality constants. However, the results in Eqs. (41), (42) are not equal (or proportional); contradiction. Therefore C cannot exist. □

Note that the No-cloning theorem does not forbid cloning in a fixed basis - try for example constructing a quantum mechanical protocol that can clone $|\uparrow\rangle, |\downarrow\rangle$ states. What it does forbid is the cloning of arbitrary states, i.e. cloning in an arbitrary basis; this is where the linearity in the above proof kicks in to derive a contradiction. We leave the reader to ponder why the existence of classical copying machines is not in contradiction with the No-cloning theorem.

2.6 Superdense Coding and Quantum Teleportation

Now that we've looked at a non-application of entanglement, let's start to study some actual applications!

To set up our discussion; we introduce the notion of a quantum bit, or qubit, which generalizes the notion of a classical bit to a quantum setting. A bit is the fundamental unit of information classically, taking on one of two states, 0 or 1. A qubit as the quantum-mechanical fundamental unit of information can take on any complex superposition of the 0 and 1 states:

$$|\psi\rangle = \alpha |0\rangle + \beta |1\rangle . \tag{43}$$

$|0\rangle, |1\rangle$ are orthogonal basis states (like the classical bit) that span the Hilbert space $\mathcal{H} = \mathbb{C}^2$ that the qubit lives in. So really when we have been discussing spin-1/2 systems we have been talking about qubits all along! We have done nothing really more than re-label $|\uparrow\rangle$ with $|0\rangle$ and $|\downarrow\rangle$ with $|1\rangle$.
What is different compared to the classical setting is that qubits can be in superposition states, and they can be entangled with each other (if we have multiple of them). We now ask - what can we do with qubits? This question in generality is still a research field, with quantum algorithms being a exciting research area (some examples of which we will discuss later in this chapter). But to begin, maybe let's start with a slightly easier question; what is a qubit worth, relative to a classical bit?
As we have discussed before, to specify a qubit state, we have two complex coefficients α, β, but with the normalization constraint $\langle\psi|\psi\rangle = 1 \implies |\alpha|^2 + |\beta|^2 = 1$ and the irrelevance of global phase $|\psi\rangle \sim e^{i\phi}|\psi\rangle$ a single qubit state is uniquely specified by two real numbers. To specify a real number, we require an infinite number of bits, so is the answer that a qubit is worth an infinite number of classical bits?
Well no, not quite; although a general qubit state is in fact specified by two real numbers, when we measure the qubit in some basis, we only ascertain one of two outcomes - in other words, we can only measure one bit worth of information from a qubit. So, is the answer that a qubit is worth exactly one classical bit?

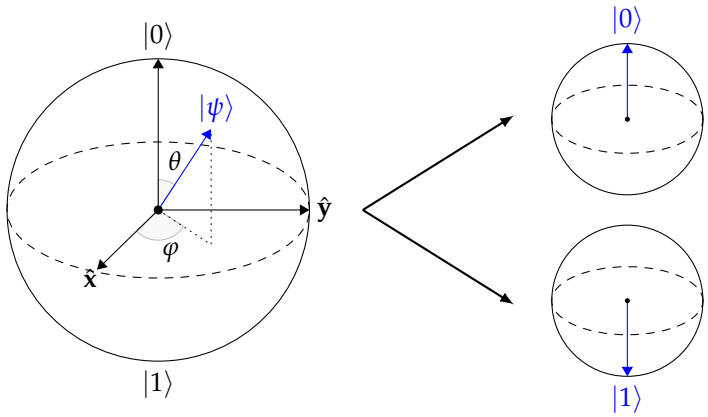

Figure 2: When we measure a qubit in the S_z eigenbasis (in quantum information lingo, called the computational basis) of $\{|0\rangle, |1\rangle\}$, we only find one of two outcomes, and the post measurement-state is one of $|0\rangle, |1\rangle$ - one of two states, just like the classical bit (this is true regardless of what single-qubit measurement basis we choose; the possible post-measurement states will be some two antipodal points on the Bloch sphere).

The answer to this question is illuminated by the discussion of our first (real!) quantum protocol - known as superdense coding. We will see in this protocol that it is possible to encode *two* classical bits in a single qubit; provided we make use of entanglement.
To discuss this protocol, we introduce the Bell basis - we have discussed the Bell state $|B_{11}\rangle$ already, but in fact there are four Bell states, which form a orthonormal basis (check!) for the Hilbert space

of two qubits $\mathcal{H} = \mathbb{C}^2 \otimes \mathbb{C}^2$:

$$\begin{aligned} |B_{00}\rangle &= \frac{|0\rangle \otimes |0\rangle + |1\rangle \otimes |1\rangle}{\sqrt{2}} \\ |B_{01}\rangle &= \frac{|0\rangle \otimes |0\rangle - |1\rangle \otimes |1\rangle}{\sqrt{2}} \\ |B_{10}\rangle &= \frac{|0\rangle \otimes |1\rangle + |1\rangle \otimes |0\rangle}{\sqrt{2}} \\ |B_{11}\rangle &= \frac{|0\rangle \otimes |1\rangle - |1\rangle \otimes |0\rangle}{\sqrt{2}} \end{aligned} \tag{44}$$

Recalling that $\sigma_z = |0\rangle\langle 0| - |1\rangle\langle 1|$ and $\sigma_x = |0\rangle\langle 1| + |1\rangle\langle 0|$, we make the observation that the 01/10/11 Bell states are related to the 00 Bell state via the application of Paulis on the first qubit:

$$\begin{aligned} |B_{01}\rangle &= \sigma_{z,1} |B_{00}\rangle \\ |B_{10}\rangle &= \sigma_{x,1} |B_{00}\rangle \\ |B_{11}\rangle &= \sigma_{z,1}\sigma_{x,1} |B_{00}\rangle \end{aligned} \tag{45}$$

Which we may summarize with the relation:

$$|B_{ab}\rangle = (\sigma_{z,1})^b (\sigma_{x,1})^a |B_{00}\rangle . \tag{46}$$

Note that we could very well have applied the Paulis to the second qubit, i.e.:

$$|B_{ab}\rangle = (\sigma_{z,2})^b (\sigma_{x,2})^a |B_{00}\rangle . \tag{47}$$

Although we will not need it for the superdense coding protocol (we will need it for the following teleportation protocol), it will be useful to note one more property of $|B_{00}\rangle$, namely that it is the $+1$ eigenvalue of $\sigma_z \otimes \sigma_z$ and $\sigma_x \otimes \sigma_x$ (check!):

$$\begin{aligned} \sigma_z \otimes \sigma_z |B_{00}\rangle &= |B_{00}\rangle \\ \sigma_x \otimes \sigma_x |B_{00}\rangle &= |B_{00}\rangle \end{aligned} \tag{48}$$

Physically, this means that S_z and S_x measurements on the two qubits of $|B_{00}\rangle$ are perfectly correlated. It is also worth noting that $|B_{00}\rangle$ is uniquely specified by the property that it is so-called *stabilized* by $\sigma_z \otimes \sigma_z$ and $\sigma_x \otimes \sigma_x$ - this is probably the first time you have seen states described in this manner, but if you go on to do more courses/research in quantum information theory (and in particular quantum error correction) you will see this method of specifying states (via the operators they are stabilized by) come up time and time again through the *stabilizer formalism*.
With this, we now have all the tools available to understand the superdense coding protocol, which we now lay out here.

Protocol: Superdense coding

Objective: Transmit two bits of information a, b from A to B.
Protocol:

1. In advance of the actual communication, the parties share a bell state $|B_{00}\rangle$ between themselves.
2. Sender A applies $(\sigma_z)^b(\sigma_x)^a$ to their qubit, encoding $(a, b) \in \mathbb{Z}_2 \times \mathbb{Z}_2$.
3. A sends their qubit to B.
4. B measures their qubit in the Bell basis[a]. Depending on the outcome, they recover a, b.

[a]Formally, they can measure some observable $O = \lambda_{00}|B_{00}\rangle\langle B_{00}| + \lambda_{01}|B_{01}\rangle\langle B_{01}| + \lambda_{10}|B_{10}\rangle\langle B_{10}| + \lambda_{11}|B_{11}\rangle\langle B_{11}|$, and since the state they have is one of the four Bell states, depending on which outcome λ_{ab} they measure they can recover the two bits a, b.

So given the above protocol, is the answer that one qubit is worth two classical bits? The answer is not so clear cut - we would not have been able to transmit two bits of information had we only sent over an unentangled qubit. Indeed, the entanglement here played a role, and in communicating the two bits of information, we have used up one bit of entanglement.
Let us also discuss the counterpart to the superdense coding protocol - namely, the quantum teleportation protocol. In superdense coding, we wanted to communicate two bits worth of information and so we physically sent a qubit; in the teleportion protocol things are reversed; we will communicate/teleport a qubit state, and in order to do so physically send two classical bits worth of information. The protocol is as follows.

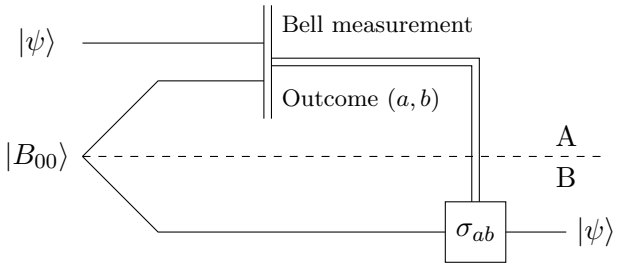

Figure 3: Graphical depiction of the quantum teleportation protocol.

Protocol: Quantum Teleportation

Objective: Transmit a qubit state $|\psi\rangle \in \mathbb{C}^2$ from A to B.
Protocol:

1. In advance of the actual communication, the parties share a bell state $|B_{00}\rangle$ between themselves.
2. A prepares the state $|\psi\rangle$ she wants to transmit (she now has two qubits; one from the shared Bell pair, and one for $|\psi\rangle$).
3. A performs a measurement in the Bell basis on their two qubits, and obtains the two-bit outcome (a, b).
4. A transmits the two-bit measurement outcome (a, b) to B.
5. B applies the correction operator $\sigma_{ab} = (\sigma_x)^a(\sigma_z)^b$ to their qubit. The resulting state of B's qubit is $|\psi\rangle$.

Proof of Correctness. Let $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ be the state that A wants to transmit to B. Let us label this qubit as qubit 1, A's half of the Bell pair as qubit 2, and B's half of the Bell pair as qubit 3. The initial state is then $|\psi\rangle_1 \otimes |B_{00}\rangle_{23}$. Now, we consider the action of the Bell measurement (with outcome (a, b)) on Alice's two qubits. Up to normalization, this has the effect of applying the projector:

$$\Pi_{ab,12} \otimes \mathbb{I}_3 = |B_{ab}\rangle\langle B_{ab}|_{12} \otimes \mathbb{I}_3 \tag{49}$$

onto the state. Using Eq. (45), we can write:

$$|B_{ab}\rangle_{ij} = (\sigma_{z,i})^b(\sigma_{x,i})^a |B_{00}\rangle_{ij} \tag{50}$$

and so we can rewrite the projector as:

$$\Pi_{ab,12} \otimes \mathbb{I}_3 = |B_{ab}\rangle\langle B_{00}|_{12} (\sigma_{x,2})^a(\sigma_{z,2})^b \otimes \mathbb{I}_3 \tag{51}$$

So applying this to the initial state, we find:

$$\begin{aligned}\Pi_{ab,12} \otimes I_3 (|\psi\rangle_1 \otimes |B_{00}\rangle_{23}) &= \left(|B_{ab}\rangle\langle B_{00}|_{12} (\sigma_{x,2})^a(\sigma_{z,2})^b \otimes \mathbb{I}_3\right) ((\alpha|0\rangle_1 + \beta|1\rangle_1) \otimes |B_{00}\rangle_{23}) \\ &= (|B_{ab}\rangle\langle B_{00}|_{12} \otimes \mathbb{I}_3) \left((\alpha|0\rangle_1 + \beta|1\rangle_1) \otimes ((\sigma_{x,2})^a(\sigma_{z,2})^b \otimes \mathbb{I}_3) |B_{00}\rangle_{23}\right)\end{aligned} \tag{52}$$

Now using that $|B_{00}\rangle$ is stabilized by $\sigma_z \otimes \sigma_z$ and $\sigma_x \otimes \sigma_x$ (Equation (48)) we can write:

$$|B_{00}\rangle_{23} = (\sigma_{z,2} \otimes \sigma_{z,3})^b |B_{00}\rangle_{23} = (\sigma_{z,2} \otimes \sigma_{z,3})^b(\sigma_{x,2} \otimes \sigma_{x,3})^a |B_{00}\rangle_{23} \tag{53}$$

And now using that $\sigma_i^2 = \mathbb{I}$ for each of the Pauli matrices:

$$\begin{aligned}((\sigma_{x,2})^a(\sigma_{z,2})^b \otimes \mathbb{I}_3) |B_{00}\rangle_{23} &= ((\sigma_{x,2})^a(\sigma_{z,2})^b \otimes \mathbb{I}_3)(\sigma_{z,2} \otimes \sigma_{z,3})^b(\sigma_{x,2} \otimes \sigma_{x,3})^a |B_{00}\rangle_{23} \\ &= ((\sigma_{x,2})^a \otimes \mathbb{I}_3)((\sigma_{z,2})^2 \otimes \sigma_{z,3})^b(\sigma_{x,2} \otimes \sigma_{x,3})^a |B_{00}\rangle_{23} \\ &= ((\sigma_{x,2})^a \otimes \mathbb{I}_3)(\mathbb{I}_2 \otimes \sigma_{z,3})^b(\sigma_{x,2} \otimes \sigma_{x,3})^a |B_{00}\rangle_{23} \\ &= (\mathbb{I}_2 \otimes \sigma_{z,3})^b((\sigma_{x,2})^2 \otimes \sigma_{x,3})^a |B_{00}\rangle_{23} \\ &= (\mathbb{I}_2 \otimes \sigma_{z,3})^b(\mathbb{I}_2 \otimes \sigma_{x,3})^a |B_{00}\rangle_{23} \\ &= \mathbb{I}_2 \otimes (\sigma_{z,3})^b(\sigma_{x,3})^a |B_{00}\rangle_{23}\end{aligned} \tag{54}$$

So then Eq. (52) becomes:

$$\begin{aligned}
&\Pi_{ab,12} \otimes I_3 \left(|\psi\rangle_1 \otimes |B_{00}\rangle_{23}\right) \\
&= (\sigma_{z,3})^b (\sigma_{x,3})^a \left(|B_{ab}\rangle\langle B_{00}|_{12} \otimes \mathbb{I}_3\right) \left(|\psi\rangle_1 \otimes |B_{00}\rangle_{23}\right) \\
&= (\sigma_{z,3})^b (\sigma_{x,3})^a \left(\frac{|B_{ab}\rangle\langle 00|_{12} + |B_{ab}\rangle\langle 11|_{12}}{\sqrt{2}} \otimes \mathbb{I}_3\right) \left(\frac{\alpha |000\rangle_{123} + \beta |100\rangle_{123} + \alpha |011\rangle_{123} + \beta |111\rangle_{123}}{\sqrt{2}}\right) \\
&= (\sigma_{z,3})^b (\sigma_{x,3})^a \frac{1}{2} \left(|B_{ab}\rangle_{12}\right) \otimes \left(\alpha |0\rangle_3 + \beta |1\rangle_3\right) \\
&= \frac{1}{2} |B_{ab}\rangle_{12} \otimes \left((\sigma_{z,3})^b (\sigma_{x,3})^a |\psi\rangle_3\right)
\end{aligned} \tag{55}$$

So indeed we have teleported $|\psi\rangle$ to the third (B's) qubit, up to applying a correction operator of $\sigma_{ab} = (\sigma_{x,3})^a (\sigma_{z,3})^b$.

□

Note that although the name might suggest some kind of superluminal communication, the teleportation protocol is fully consistent with special relativity - in order to recover the correct state $|\psi\rangle$ at the end, A must transmit to B the two-bit measurement outcome of the Bell basis measurement; this communication cannot exceed light speed.
A tangent to conclude this section that is certainly beyond the scope of this lecture but nevertheless interesting; a variant of the teleportation protocol (half-teleportation) forms the backbone of measurement-based quantum computation, where a computation is carried out solely via a sequence of local (adaptive) measurements on an initial resource state. You can read more about it here, among other places.

3 Manifolds

3.1 Metric Spaces

We begin by attempting to motivate the definition of a sort of abstract space, called a "metric space" which will later prove important to our definition of a manifold. For the sake of motivating this definition, begin by recalling the definition of a continuous function on $\mathbb{R}$:

Definition 3.1 (Continuity on $\mathbb{R}$)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be some function, and let $x \in \mathbb{R}$. We say that f is* continuous at *x if, for all $\varepsilon > 0$, there exists some $\delta > 0$ such that for all $y \in \mathbb{R}$, $|x - y| < \delta$ implies $|f(x) - f(y)| < \varepsilon$.*

Observe, in particular, that a huge majority of the structure of $\mathbb{R}$ is entirely irrelevant to this definition. In particular, the only part of the structure of $\mathbb{R}$ which we really required to make this definition was the function $d(x, y) = |x - y|$, which gives us a way of talking about the distance between two points in $\mathbb{R}$. Based on this observation, we may begin to suspect that we may be able to define continuity on a much broader class of spaces, namely those which have some coherent notion of distance. This rough idea of a "coherent notion of distance" is central to the definition of a metric space:

Definition 3.2 (Metric Space)**.** *A* metric space *(X, d) is a set X along with a function $d : X \to \mathbb{R}$ (this function is sometimes called a metric, but in the context of manifolds this terminology becomes somewhat confounding, so we avoid it) satisfying the following properties for all $x, y, z \in X$:*

- *$d(x, y) \geq 0$, with $d(x, y) = 0$ if and only if $x = y$*

- $d(x, y) = d(y, x)$
- $d(x, z) \leq d(x, y) + d(y, z)$.

These properties are called positive-definiteness, symmetry, and the triangle inequality, respectively. We will sometimes speak of "the metric space X" if confusion about d is unlikely to arise, or if d is irrelevant.

While the importance of positive-definiteness and symmetry in the above definition are hopefully fairly clear in their importance to a sensible definition of distance, the importance of the triangle inequality is perhaps somewhat less obvious. The general idea of the triangle inequality is that metric spaces ought to satisfy the classical Euclidean notion that the shortest distance between two points is a straight line. In other words, it should always be shorter to go directly from x to z than to take a detour through y along the way. If theorems happen to motivate you more than intuition does, the triangle inequality is also necessary for continuous functions (as we will shortly define between metric spaces) to have their usual closure properties, namely that the sums and products of continuous functions are continuous. With the definition of a metric space under our belt, we can proceed to providing a definition of continuity for functions between metric spaces:

Definition 3.3 (Continuity)**.** *Let (X, d_X) and (Y, d_Y) be metric spaces, let $f : X \to Y$ be some function, and let $x \in X$. We say that f is* continuous at x *if, for all $\varepsilon > 0$, there exists some $\delta > 0$ such that for all $y \in X$, $d_X(x, y) < \delta$ implies $d_Y(f(x), f(y)) < \varepsilon$.*

Observe that this definition is nearly identical to our previous definition of continuity, just with the distance between two points in $\mathbb{R}$ replaced with the appropriate distance function for our metric space. Essentially all of the usual theorems about continuous functions on $\mathbb{R}$ carry over wholesale with similar minor modifications, so we skip over this process of porting over theorems, although they may be a good exercise for increasing your comfort in reasoning about metric spaces. Finally, aside from continuity, there is a property of subsets of metric spaces which we will make heavy use of, so we define it here:

Definition 3.4 (Open Set)**.** *Let (X, d) be a metric space, and let $U \subseteq X$. U is called* open *if, for all $x \in U$, there exists some $\varepsilon > 0$ such that $d(x, y) < \varepsilon$ implies $y \in U$.*

3.2 Topological Manifolds & Smooth Manifolds

As shown in the previous section, the definition of a metric space allowed us to hugely broaden the class of spaces upon which we have a notion of continuity. We might then hope that we can similarly extend all of our constructions in calculus, such as differentiability, to metric spaces in a similar manner. Unfortunately, such a project will fail. A brief examination of the definition of differentiability of a function on $\mathbb{R}$ (or, for that matter $\mathbb{R}^n$) reveals that the definition of differentiability depends in a much more integral way on the particular structure of $\mathbb{R}$ than did the definition of continuity. As such, if we wish to extend the notion of differentiability to a broader class of spaces, it proves necessary to impose more structure on metric spaces. The critical observation turns out to be that for a space to admit a coherent notion of differentiability, it is necessary that our spaces be "locally Euclidean" in some sense. Making this notion of a locally Euclidean metric space precise will require one important preliminary definition:

Definition 3.5 (Homeomorphism)**.** *Let X and Y be metric spaces. $f : X \to Y$ is called a* homeomorphism *if f is invertible, f is continuous, and f^{-1} is continuous. If there exists a homeomorphism between X and Y, then they are called* homeomorphic*.*

This idea of homeomorphism is important in that it allows us to speak of two spaces being the same up to some kind of well-behaved deformation. Consider the Earth as an example. Globally speaking, the earth is a sphere, which is not homeomorphic to $\mathbb{R}^2$, so we can claim that, as metric spaces, the surface of the Earth is somehow fundamentally different from a plane (in that they are not homeomorphic). That being said, a local patch of the Earth in your immediate vicinity certainly appears flat. While this patch of Earth around you is almost not completely without curvature (indeed, some long bridges must account for the local curvature of the Earth or risk being off in their measurements by several centimetres or even metres), it is still very much like $\mathbb{R}^2$. In particular, it is homeomorphic to $\mathbb{R}^2$. This property of being locally homeomorphic to $\mathbb{R}^n$ is so important that we make the following definition:

Definition 3.6 (Topological Manifold)**.** *Let X be a metric space. X is called an* n-dimensional topological manifold *if, for all $x \in X$, there exists some open set $U \subseteq X$ such that $x \in U$ and such that U is homeomorphic to $\mathbb{R}^n$. If $\varphi : U \to \mathbb{R}^n$ is the homeomorphism in question, then (U, φ) is called a* chart*. A collection of charts covering M is called an* atlas*.*

With the definition of a topological manifold complete, let us see if we can come up with a definition of differentiability for functions between topological manifolds. For the sake of simplicity going forward, we will concern ourselves with smooth functions, which is to say those which are infinitely differentiable, rather than simply differentiable functions. It is still perfectly acceptable to talk about functions which are differentiable some finite number of times, but it introduces a lot more complexity for very little benefit. With a definition of a topological manifold, we define smoothness as follows:

Definition 3.7 (Smoothness?)**.** *Let M and N be m- and n-dimensional topological manifolds, respectively, let $f : M \to N$ be a function, and let (U, φ) and (V, ψ) be charts on M and N respectively such that $f(U) \subseteq V$. We say that f is* smooth on U *if the function $\psi \circ f \circ \varphi^{-1} : \mathbb{R}^m \to \mathbb{R}^n$ is smooth.*

This definition of smoothness seems good, but unfortunately it comes with a major caveat. We want our notion of smoothness to be independent of our choice of local coordinates for our topological manifold, because smoothness should simply be a property of the function, which is independent of a particular choice of coordinates. Suppose that, in the above definition, $(\tilde{U}, \tilde{\varphi})$ is another chart on M such that $f(\tilde{U}) \subseteq V$, and $\tilde{U} \cap U \neq \varnothing$. If we wish to know whether or not f is smooth on $\tilde{U} \cap U$, there are two ways that we could attempt to answer this question. We could ask if $\psi \circ f \circ \varphi^{-1}$ is smooth, but we could just as well ask if $\psi \circ f \circ \tilde{\varphi}^{-1}$ is smooth. For this definition of smoothness to be coherent, we would need to know that $\psi \circ f \circ \varphi^{-1}$ is smooth if and only if $\psi \circ f \circ \tilde{\varphi}^{-1}$ is smooth, but a topological manifold does not have enough structure for us to make this guarantee. Observe, however, that $\psi \circ f \circ \varphi^{-1} = \psi \circ f \circ \tilde{\varphi}^{-1} \circ (\tilde{\varphi} \circ \varphi^{-1})$ and $\psi \circ f \circ \tilde{\varphi}^{-1} = \psi \circ f \circ \varphi^{-1} \circ (\varphi \circ \tilde{\varphi}^{-1})$. With this in mind, we make a new definition:

Definition 3.8 (Smooth Compatibility)**.** *Let M be an n-dimensional topological manifold, and let (U, φ) and $(\tilde{U}, \tilde{\varphi})$ be charts on M such that $\tilde{U} \cap U \neq \varnothing$. The charts are called* smoothly compatible *if the* transition functions *$\tilde{\varphi} \circ \varphi^{-1} : \varphi(\tilde{U} \cap U) \to \tilde{\varphi}(\tilde{U} \cap U)$ and $\varphi \circ \tilde{\varphi}^{-1} : \tilde{\varphi}(\tilde{U} \cap U) \to \varphi(\tilde{U} \cap U)$ are both smooth. An atlas composed of charts which are all pairwise smoothly compatible is called a* smooth atlas*.*

Armed with this definition, we can amend our previous attempt at defining smooth functions between topological manifolds. In particular, we observe that topological manifolds do not quite have enough structure to define smoothness in a coherent way, so it is necessary to introduce a new definition encoding the additional structure which we require:

Definition 3.9 (Smooth Manifold)**.** *An n-dimensional* smooth manifold *$(M, \mathscr{A})$ is composed of an n-dimensional topological manifold M along with a smooth atlas $\mathscr{A}$ on M. When $\mathscr{A}$ is clear from context or irrelevant, we will sometimes refer to just M as a smooth manifold.*

From this point on in the notes, when we refer simply to a "manifold", with no qualifiers, it should always be taken to mean a smooth manifold. Finally, with the definition of a manifold in hand, we can amend our previous definition of smoothness:

Definition 3.10 (Smoothness)**.** *Let M and N be m- and n-dimensional manifolds, respectively, let $f : M \to N$ be a function, and let (U, φ) and (V, ψ) be charts on M and N respectively such that $f(U) \subseteq V$. We say that f is* smooth on U *if the function $\psi \circ f \circ \varphi^{-1} : \mathbb{R}^m \to \mathbb{R}^n$ is smooth.*

Notation 3.11. *We will denote the set of all smooth functions from a manifold M to $\mathbb{R}$ by $C^\infty(M)$. The set of all smooth functions from $\mathbb{R}^n$ to $\mathbb{R}$ will usually just be denoted C^∞, unless confusion about n is likely to arise.*

3.3 Tangent Spaces

Having defined manifolds, we are now able to talk about what it means for a function between manifolds to be smooth. Despite this, we do not immediately have a way of saying what it means to compute the derivative of a function on a manifold. To this end, we define an analogue of the directional derivative on arbitrary manifolds:

Definition 3.12 (Derivation)**.** *Let M be a manifold, and let $p \in M$. A* derivation at p *is a function $v : C^\infty(M) \to \mathbb{R}$ satisfying the following properties for all $f, g \in C^\infty(M)$ and $a, b \in \mathbb{R}$:*

- $v(af + bg) = av(f) + bv(g)$
- $v(fg) = f(p)v(g) + g(p)v(f)$.

The properties are called linearity and the Leibniz rule, respectively.

From which a new definition immediately follows:

Definition 3.13 (Tangent Space)**.** *Let M be a manifold, and let $p \in M$. The* tangent space of M at p *is the set of all derivations at p, which we denote T_pM. The tangent space has a natural vector space structure once the identification $(av + bu)(f) = av(f) + bu(f)$ for all $v, u \in T_pM$ and $a, b \in \mathbb{R}$.*

One way of thinking about the tangent space is, as the name suggests, as the vector space composed of all vectors which are tangent to M at some fixed point p. The behaviour of our derivations then corresponds to taking the directional derivative of functions in the direction of these tangent vectors. While this definition of the tangent space is entirely usable, for practical purposes it is valuable to be able to express elements of the tangent space in terms of local coordinates on the manifold. Functions on M are, naturally, usually written in terms of some kind of local coordinates, so for the sake of actually computing the result of applying a derivation to a function it is valuable to have a means of expressing derivations in local coordinates as well. In particular, let M be some n-dimensional manifold, let (U, φ) be some chart on M, and let p be some point in the chart. Suppose that $\varphi = (x^1, \ldots, x^n)$, so that the $x^i : M \to \mathbb{R}$ are the component functions of φ. This choice of coordinates induces a natural basis on T_pM, called the coordinate basis, whose elements are denoted $\partial/\partial x^1\big|_p, \ldots, \partial/\partial x^n\big|_p$. Observe that the tangent space at any point of an n-dimensional manifold

is then an n-dimensional vector space. The action of the elements of the coordinate basis on some $f \in C^\infty(M)$ is then defined by

$$\left.\frac{\partial}{\partial x^i}\right|_p (f) = \left.\frac{\partial (f \circ \varphi^{-1})}{\partial e^i}\right|_{\varphi(p)},$$

where e^i is the i-th coordinate in $\mathbb{R}^n$. One way of visualizing this is as φ^{-1} mapping the axes in $\mathbb{R}^n$ to so-called coordinate curves in M. The action of $\partial/\partial x^i\big|_p$ then corresponds to taking the derivative of f along the i-th coordinate curve.

Notation 3.14. *When confusion about the precise local coordinates in question is unlikely to arise, or if the coordinates are irrelevant, we will generally write $\partial_i|_p$ instead of $\partial/\partial x^i\big|_p$ for brevity.*

If $v \in T_pM$, then we can write v in terms of the coordinate basis. In particular, suppose that $v = \sum_i a^i \, \partial_i|_p$. If we observe that

$$\begin{aligned}
\partial_i|_p (x^j) &= \left.\frac{\partial (x^j \circ \varphi^{-1})}{\partial e^i}\right|_{\varphi(p)} \\
&= \left.\frac{\partial e^j}{\partial e^i}\right|_{\varphi(p)} \\
&= \delta_i^j,
\end{aligned}$$

then it immediately follows that

$$\begin{aligned}
v(x^j) &= \sum_i a^i \, \partial_i|_p (x^j) \\
&= \sum_i a^i \delta_i^j \\
&= a^j,
\end{aligned}$$

so we have $a^i = v(x^i)$, and the expression for v in local coordinates is simply $v = \sum_i v(x^i) \, \partial_i|_p$.

Notation 3.15 (Einstein Summation Convention)**.** *When an index is repeated in some expression, once in subscript and once in superscript, summation over all values of that index is implied. For the previously computed expression for v in terms of local coordinates, for instance, we could suppress the summation sign, and rather than writing $v = \sum_i v(x^i) \, \partial_i|_p$ simply write $v = v(x^i) \, \partial_i|_p$, with summation over i being implied by the Einstein summation convention.*

We have now defined taking directional derivatives of functions in $C^\infty(M)$, but the question remains of what it means to take the derivative of some smooth function $F : M \to N$, where N is some arbitrary manifold. This question leads us to the final definition in this section:

Definition 3.16 (Differential Map)**.** *Let M and N be manifolds, let $p \in M$, and let $F : M \to N$ be a smooth function. The* differential of F at p *is the function $dF_p : T_pM \to T_{F(p)}N$ defined by $(dF_p(v))(f) = v(f \circ F)$.*

We conclude by examining the form which the differential map takes on in coordinates. Suppose that we fix some charts (U, φ) and (V, ψ) on M and N respectively, such that $p \in U$ and $F(p) \in V$.

Suppose further that $\varphi = (x^1, \ldots, x^m)$ and $\psi = (y^1, \ldots, y^n)$, then if $v \in T_pM$ and $f \in C^\infty(N)$, by definition we have

$$\begin{aligned}
(\mathrm{d}F_p(v))(f) &= \left(\mathrm{d}F_p\left(v(x^i)\left.\frac{\partial}{\partial x^i}\right|_p\right)\right)(f) \\
&= \left(v(x^i)\left.\frac{\partial}{\partial x^i}\right|_p\right)(f \circ F) \\
&= v(x^i)\left.\frac{\partial}{\partial x^i}\right|_p (f \circ F) \\
&= v(x^i)\left.\frac{\partial(f \circ F \circ \varphi^{-1})}{\partial e^i}\right|_{\varphi(p)} \\
&= v(x^i)\left.\frac{\partial(f \circ \psi^{-1} \circ \psi \circ F \circ \varphi^{-1})}{\partial e^i}\right|_{\varphi(p)} \\
&= v(x^i)\left.\frac{\partial(\psi \circ F \circ \varphi^{-1})^j}{\partial e^i}\right|_{\varphi(p)} \left.\frac{\partial(f \circ \psi^{-1})}{\partial e^j}\right|_{\psi(F(p))} \\
&= v(x^i)\left.\frac{\partial(\psi \circ F \circ \varphi^{-1})^j}{\partial e^i}\right|_{\varphi(p)} \left.\frac{\partial}{\partial y^j}\right|_{F(p)} (f),
\end{aligned}$$

where $(\psi \circ F \circ \varphi^{-1})^j$ denotes the j-th component function of $\psi \circ F \circ \varphi^{-1}$. We then conclude that the differential map of F at p is given in coordinates by

$$\mathrm{d}F_p\left(v(x^i)\left.\frac{\partial}{\partial x^i}\right|_p\right) = v(x^i)\left.\frac{\partial(\psi \circ F \circ \varphi^{-1})^j}{\partial e^i}\right|_{\varphi(p)} \left.\frac{\partial}{\partial y^j}\right|_{F(p)}.$$

Observe that, with $\mathrm{d}F_p$ being a linear map, this takes on the form of matrix multiplication. In particular, the partial derivatives $\partial(\psi \circ F \circ \varphi^{-1})^j/\partial e^i\big|_{\varphi(p)}$ are the elements of the matrix representation of $\mathrm{d}F_p$ in terms of the selected coordinates on M and N. This matrix representation of $\mathrm{d}F_p$ is usually called the Jacobian matrix (or just the Jacobian) of F.

3.4 Tangent Bundles

Before we are able to turn to applying the theory of manifolds to Lagrangian mechanics, we require one final theoretical notion: the tangent bundle. As defined in the previous section, at any given point p on an n-dimensional manifold M, we have an n-dimensional vector space T_pM composed of all of the derivations at p. One issue with this definition is that, at least a priori, we have a massive number of tangent spaces at every point on our manifold, and no way to discuss all of the tangent spaces on a manifold as being part of a larger entity. To this end, we define the tangent bundle, which for the purposes of this lecture may be thought of as effectively being a bookkeeping tool.

Definition 3.17 (Tangent Bundle)**.** *Let M be a manifold, then the* tangent bundle *of M, denoted TM is the set defined by*

$$TM = \bigcup_{\substack{p \in M \\ v \in T_pM}} (p, v).$$

A central reason for theoretical interest in the tangent bundle is the fact that the tangent bundle of a manifold is, once equipped with a smooth atlas in an appropriate manner, a manifold in its own right. For our purposes, however, this is largely unimportant, and we will treat tangent bundles as simple bookkeeping tools.

3.5 Lagrangian Mechanics on Manifolds

Having defined and examined many of the central concepts in the theory of manifolds, we turn finally to an application of differential geometry to physics: Lagrangian mechanics on manifolds. While this topic is perhaps less famous as an application of differential geometry than, say, Hamiltonian mechanics or general relativity, each of these applications require the development of substantially more differential geometry (in particular, core to the study of Hamiltonian mechanics is the idea of a symplectic manifold, and symplectomorphisms between symplectic manifolds, and core to the study of general relativity is the idea of a Riemannian manifold). As such, we examine an application of the theory of manifolds to a problem in Lagrangian mechanics, namely that of a bead on a helical wire.

We begin by defining a helical wire of radius R and pitch k, which we take to be a subset of $\mathbb{R}^3$ defined by $M = \{(R\cos(\theta), R\sin(\theta), k\theta/2\pi) \in \mathbb{R}^3 : \theta \in \mathbb{R}\}$. This naturally defines a smooth atlas for M consisting of the single smooth map $\varphi : \mathbb{R} \to M$ defined by

$$\varphi(\theta) = \begin{bmatrix} R\cos(\theta) \\ R\sin(\theta) \\ k\theta/2\pi \end{bmatrix}.$$

With this map in hand, we may now compute its differential, which is given in coordinates, by the formula presented previously, by the Jacobian matrix

$$\mathrm{d}\varphi(\theta) = \begin{bmatrix} -R\sin(\theta) \\ R\cos(\theta) \\ k/2\pi \end{bmatrix},$$

which is a map of the form $\mathrm{d}\varphi : T\mathbb{R} \to TM$. Given these, we proceed to define two functions, called the kinetic energy and the potential energy. We begin with the potential energy, which is a map $U : M \to \mathbb{R}$. In this case, we take U to be defined in coordinates by $U(\theta) = mgk\theta/2\pi$, which you may recognize as corresponding to the potential energy of a bead with mass m threaded onto the rod, under the influence of gravity. The definition of the kinetic energy is somewhat more subtle, as it relies on the existence of a function called a *metric* on M. Defining metrics in full detail would take more time than we have available, so I will commit here what I hope is the first and only instance of hand-waving in this lecture. There is a standard metric on $\mathbb{R}^3$, called the Euclidean metric which is a function $g_{\mathbb{R}^3} : T\mathbb{R}^3 \times T\mathbb{R}^3 \to \mathbb{R}$, defined by $g_{\mathbb{R}^3}(\partial_i, \partial_j) = \delta_{ij}$. This defines $g_{\mathbb{R}^3}$ on all of $T\mathbb{R}^3 \times T\mathbb{R}^3$, as metrics are necessarily linear in both of their arguments. I claim that using this metric I may obtain an expression in coordinates for a corresponding metric on M. In particular, I claim that the coordinate representation of g_M, the metric on M, is given by

$$g_M(\partial_1, \partial_1) = g_{\mathbb{R}^3}(\mathrm{d}\varphi(\partial_1), \mathrm{d}\varphi(\partial_1)) = R^2 + \frac{k^2}{4\pi^2}.$$

Using this metric on M, we may define the kinetic energy, which is a function $T : TM \to \mathbb{R}$ defined by

$$T(\partial_i) = \frac{1}{2} m g(\partial_i, \partial_i).$$

In particular, on M, the kinetic energy is given, in coordinates, by

$$T(v) = \frac{1}{2} m g_M(v, v) = \frac{1}{2} m \left(R^2 + \frac{k^2}{4\pi^2} \right) \dot{\theta}^2.$$

Having defined the potential and kinetic energies, we may compute the Lagrangian as usual, which yields a function $\mathcal{L} : TM \to \mathbb{R}$ defined by

$$\mathcal{L}(\theta, v) = T(v) - U(\theta) = \frac{1}{2}m\left(R^2 + \frac{k^2}{4\pi^2}\right)\dot{\theta}^2 - \frac{mgk\theta}{2\pi}.$$

We may now apply the usual Euler-Lagrange equations, which, in this case, tells us that

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathcal{L}}{\partial \dot{\theta}} = m\left(R^2 + \frac{k^2}{4\pi^2}\right)\ddot{\theta} = -\frac{mgk}{2\pi} = \frac{\partial \mathcal{L}}{\partial \theta}.$$

Rearranging this yields a straightforward expression for $\ddot{\theta}$, namely

$$\ddot{\theta} = -\frac{gk}{2\pi R^2 + k^2/2\pi},$$

which is indeed the correct equation of motion for a bead of mass m on a helical wire of radius R and pitch k, subject to the force of gravity.

4 4-Vectors and Similar Objects

4.1 Tensors as Multilinear Maps

The view of tensors used in quantum mechanics is fairly unique in physics. In both classical and relativistic mechanics, you're more likely to see the following definition.

Definition 4.1. *Let V be a vector space over $\mathbb{R}$. A (k, ℓ)-Tensor T is a map $T : V'^k \times V^\ell \to \mathbb{R}$ which is multilinear (i.e. linear in each argument).*

A $(0, \ell)$-tensor is often called a covariant ℓ-tensor, and a $(k, 0)$-tensor a contravariant k-tensor. The correspondence here to our original definition is not totally immediate, but with a bit of work we can tease it out. First, we'll look at the *tensor product* in this context.

Definition 4.2. *Let V be a vector space over $\mathbb{R}$, T a (k, ℓ)-Tensor and S a (k', ℓ')-Tensor. Then tensor product $T \otimes S$ of these two maps is the multilinear $(k + k', \ell + \ell')$-Tensor defined in the following manner (where we use $'$ to indicated dual vectors)*

$$(T \otimes S)(\vec{v}'_1, \ldots, \vec{v}'_{k+k'}, \vec{u}_1, \ldots, \vec{u}_{\ell+\ell'}) = T(\vec{v}_1, \ldots, \vec{v}'_k, \vec{u}_1, \ldots, \vec{u}_k)S(\vec{v}'_{k+1}, \ldots, \vec{v}'_{k+k'}, \vec{u}_{\ell+1}, \ldots, \vec{u}_{\ell+\ell'})$$

One can check that this is, in fact, multilinear, and using it we can start to build up our correspondence with the quantum mechanical view of the tensor product. Let V be an n-dimensional vector space over $\mathbb{R}$, with basis $\{\vec{u}_1, \ldots, \vec{u}_n\}$. We'll take it as given (this was covered last term), that this corresponds to a dual basis $\{\vec{u}'_1, \ldots, \vec{u}'_n\}$ for V' with the following property

$$\vec{u}'_i(\vec{u}_j) = \delta_{ij}$$

Let T be a contravariant 1-tensor. This is just a linear map from $V \to \mathbb{R}$, which can be written as (for some $a_i \in \mathbb{R}$

$$T\Big(\sum_{i=1}^n b_i\vec{u}_i\Big) = \sum_{i=1}^n b_i a_i = \sum_{i=1}^n a_i\vec{u}'_i\Big(\sum_{i=1}^n b_j\vec{u}_j\Big) = \Big(\sum_{i=1}^n a_i\vec{u}'_i\Big)\Big(\sum_{i=1}^n b_j\vec{u}_j\Big)$$

So in reality, $T = \sum_{i=1}^{n} a_i \vec{u}_i'$ is just an element of the dual space! We could use the same trick to find that covariant 1-tensors are just vectors. Now, let's look at what happens with the tensor product. In particular, we'll focus on the tensor product of a covariant and contravariant 1-tensor (T and S), but these ideas will generalize without issues. We know that

$$T(\vec{u}', \vec{v}) = T(\vec{u}')S(\vec{v})$$

Now let $\vec{w}$ be the vector corresponding to T and $\vec{x}'$ the dual vector corresponding to S. Then

$$T(\vec{u}')S(\vec{v}) = \vec{u}'(\vec{w})\vec{x}'(\vec{v}) = (\vec{u}' \otimes \vec{x}')(\vec{w} \otimes \vec{v})$$

where the tensor product written here is that from the quantum mechanics section. Now, let's notice that for any $a \in \mathbb{R}$

$$a(\vec{u}' \otimes \vec{x}')(\vec{w} \otimes \vec{v}) = ((a\vec{u}') \otimes \vec{x}')(\vec{w} \otimes \vec{v}) = (\vec{u}' \otimes \vec{x}')(\vec{w} \otimes (a\vec{v}))$$

and that if $\vec{v} = \vec{b} + \vec{c}$

$$(\vec{u}' \otimes \vec{x}')(\vec{w} \otimes \vec{v}) = (\vec{u}' \otimes \vec{x}')(\vec{w} \otimes \vec{b}) + (\vec{u}' \otimes \vec{x}')(\vec{w} \otimes \vec{c})$$

with a similar result happening when splitting up $\vec{u}'$. That is, the pair $\vec{u}', \vec{v}$ actually meets all the axioms for a tensor product, so our map $T \otimes S$ can be viewed as a vector in $V' \otimes V$. We can then extend this idea, noting that we get a similar result with 2 covariant/contravariant 1-tensors and that the tensor product is commutative and associative to get that (k, ℓ)-Tensors can be viewed as elements of $\otimes_{i=1}^{k} V' \otimes_{i=1}^{\ell} V$. The reason we prefer to look at tensors as multilinear maps in classical mechanics and general relativity is simply because this is the only form in which they show up, so it makes sense to avoid any abstract algebra entirely and work only in linear algebra by using multilinear maps.

4.2 Products

When dealing with vector spaces, there are a variety of products between vectors that we are concerned about. In physics, we often learn about two of these products very well, which we so fondly refer to as the **dot product** (or **scalar product**) and the **cross product** (or **vector product**). There are of course other types of products, such as the **wedge product**, but here we are only concerned with two general types of products, the **inner product** and the **outer product**. These are products that we have seen and used before but we may not yet have given a name to the procedure. As such, let us look at this more in depth.

To begin this exploration, we can begin by looking at vectors and vector spaces. For a vector space where each element of the vector $a \in \mathbb{R}$, the **inner product** of two vectors is often thought of as the **dot product**. That is, if I have two column vectors $\vec{x} \in \mathbb{R}^n$ and $\vec{y} \in \mathbb{R}^n$, then the inner product is found to be:

$$\langle \vec{x} \, | \vec{y} \rangle = \left\langle \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \middle| \middle| \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \right\rangle = \vec{x}^{\mathrm{T}} \cdot \vec{y} = \sum_{i=1}^{n} x_i y_i$$

The inner product and by extension the dot product does have some interesting properties as a result of this. Notably, we have the following:

$$\langle \vec{x} \,|\vec{y}\rangle = \overline{\langle \vec{y} \,|\vec{x}\rangle}$$

$$\langle a\vec{y} + b\vec{z} \,|\vec{x}\rangle = a \langle \vec{y} \,|\vec{x}\rangle + b \langle \vec{z} \,|\vec{x}\rangle$$

$$\left\langle \vec{0} \,|\vec{x} \right\rangle = \left\langle \vec{x} \,\middle|\vec{0} \right\rangle = 0$$

$$\langle \vec{x} \,|\vec{x}\rangle \in \mathbb{R}, \quad \langle \vec{x} \,|\vec{x}\rangle \geq 0$$

$$\langle \vec{x} \,|\vec{x}\rangle = 0 \text{ iff } \vec{x} = \vec{0}$$

$$\langle \vec{x} \,|a\vec{y} + b\vec{z}\rangle = \overline{a} \langle \vec{x} \,|\vec{y}\rangle + \overline{b} \langle \vec{x} \,|\vec{z}\rangle$$

where $\overline{a}$ denotes the complex conjugate of a. The outer product on the other hand is something a little more extreme. Once again, starting with two column vectors $\vec{x} \in \mathbb{R}^m$ and $\vec{y} \in \mathbb{R}^n$, their outer product defines some matrix A such that

$$\vec{x} \otimes \vec{y} = \vec{x}\vec{y}^{\mathrm{T}} = \mathbf{A} = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_n \\ x_2y_1 & x_2y_2 & \cdots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_my_1 & x_my_2 & \cdots & x_my_n \end{bmatrix}$$

Note that while the inner product requires $\vec{x}$ and $\vec{y}$ to be the same length (dimension), the outer product is applicable to vectors of different dimensions.
This conception extends into quantum mechanics as well. The inner product between two states is well defined and for two particular states $|\psi_1\rangle$ and $|\psi_2\rangle$, we represent the inner product as follows

$$\langle \psi_1 \,|\psi_2 \rangle$$

which should be a familiar quantity to us. Note that this satisfies all the above properties of the inner product. The outer product is also well defined as an operator is this case of the following form:

$$|\psi_1 \rangle\langle \psi_2|$$

which has some uses, particularly when projecting onto a specific basis.
While this may seem a bit of a tangent for now, I assure you that this will be applicable in the near future.

4.3 Notation

The second big thing I wanted to talk about here is the notation. To be more specific, we will review Einstein notation. Einstein notation is especially useful for us as taking the time to write out all possible combination of two vector elements can often be tedious to perform. This notational style helps to keep things organized and clutter free when working with these vector-like objects. The first notion of this appears when considering the dot product. Recall that for the dot product, we can write out the expression as a sum of the product of vector components. That is, if we have two arbitrary vectors $\vec{x} \in \mathbb{R}^n$ and $\vec{y} \in \mathbb{R}^n$, then the dot product is expressed as follows.

$$\langle \vec{x} \,|\vec{y}\rangle = \vec{x} \cdot \vec{y} = \sum_{i=1}^{n} x_i y_i = x_1y_1 + x_2y_2 + \cdots + x_ny_n$$

However, with Einstein notation, we can write the components of our vectors as x_i and y_j and conversely express the dot product with an implied sum as follows:

$$\langle \vec{x} \, | \vec{y} \rangle = \vec{x} \cdot \vec{y} = x_i y^i = y_i x^i = \vec{y} \cdot \vec{x} = \langle \vec{y} \, | \vec{x} \rangle$$

I have briefly and accidentally introduced upper and lower indices here, but I kindly ask you to ignore this for the time being until we get to talking about it in the next paragraph. Regardless, it is important to note that **for each unique index in Einstein notation, there is an implied sum over all possible values of that index.** This additionally means that we may have multiple different indices to sum over at the same time. For example, the following two expressions are not equivalent

$$x_i y^i = \sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$

$$x_i y^j = \sum_{j=1}^{n} \sum_{i=1}^{n} x_i y_j = \sum_{j=1}^{n} (x_1 + x_2 + \cdots + x_n) \, y_j = (x_1 + x_2 + \cdots + x_n) \, (y_1 + y_2 + \cdots + y_n)$$

since we have one sum in the first and two in the second.

The second important aspect of Einstein notation regards the upper and lower indices that I have been making use of. To fully understand the indices, we must understand, briefly, the concept of **contravariant** versus **covariant** vectors (and transformation). For the most part, one can ignore parts of this section if they are more concerned with understanding how to calculate things with Einstein notation, but an understanding of this may benefit an understanding of the invariant. We can of course begin by understanding what we mean by contravariant as opposed to covariant. In the majority of our previous courses, we have often dealt with basis vectors that do not change. Yet, our choice in basis vectors is not unique and as such, we may be able to transform them. It should then be obvious that if we change the basis vectors, the vectors also need to be changed to remain constant.

This is where our new terminology comes into play. **A contravariant vector (which belongs to the vector space in question) transforms inversely with respect to how the basis vectors transform. Conversely, a covariant vector (which belongs to the dual vector space) transforms similarly with respect to how the basis vectors transform.** This new terminology may seem a bit strange, but we can introduce an example to help solidify the concepts. Consider some space in which we have defined some position vector along with some wavenumber vector. To make this easier, let us suppose that we have also assigned units to our basis vectors and by extension our position and wavenumber vectors. There is nothing preventing us from doing this and doing so will make this example somewhat clearer. If we say that our basis vectors are in meters and then transform them into kilometers by multiplying our original values by 1000, we would have to equivalently divide our position vector by 1000 to retain the same meaning, which gives that our position vector is a contravariant vector. Furthermore, we would have to multiply our wavevector by 1000 to retain the same meaning, which shows that wavevector is a covariant vector.

The final product of all of this comes down to representation. In the end, when working with Einstein notation, **contravariant vectors are represented as having upper indices while covariant vectors are represented as having lower indices.**

Before we move onto higher-rank objects and how they interact with these vectors, it is important to know how to convert between a contravariant and a covariant vector. The details of this conversion would require a deeper dive into the behavior of dual spaces and such which are unfortunately not covered well in some linear algebra courses. Such a deep dive is a bit unwarranted here so I will

keep this short. In essence, when working in certain spaces, we have some object known as the metric tensor which defines distances and angles in our space. In spacetime, our metric tensor is the Minkowski space metric defined as follows.

$$g^{\mu\nu} = \begin{cases} 0 & \text{if } \mu \neq \nu \\ 1 & \text{if } \mu = \nu = 0 \\ -1 & \text{if } \mu = \nu \neq 0 \end{cases}$$

The conversion between contravariant and covariant vectors is simply done by applying the metric tensor onto the covariant and contravariant vectors, respectively.

$$v_\mu = g_{\mu\nu} v^\nu$$

$$v^\mu = v_\nu g^{\mu\nu}$$

Since we have already started talking about it, it only seems natural for us to have a small discussion regarding tensors and matrices. We will be working with many of these objects although some only distantly so a deep understanding is not completely necessary. To begin, we can have a discussion of rank. The rank of a tensor is the number of lower and upper indices that the tensor has. This is often represented as a tuple. That is a tensor with p upper indices and q lower indices is known as a (p, q)-tensor.

$$T^{i_1, i_2, \ldots, i_p}_{j_1, j_2, \ldots, j_q}$$

We can also consider the total rank or order of the tensor, in which we say that it is a rank $(p+q)$ tensor. Although not particularly accurate, many often describe matrices as rank-2 tensors. Matrices are, in our case, the more important object for us to work with. A matrix transforms a vector $\vec{w}$ into a vector $\vec{v}$ as

$$v_j = w_i A^i_j$$

$$v^j = A^j_i w^i$$

It is important to note that matrices are specifically bastardized as $(1, 1)$-tensors, we will often write them with either two upper or two lower indices. We are mostly only concerned with the total rank of the matrix rather than the contravariance or the covariance of the matrix. Thus, we may sometimes write them as

$$v^j = A_{ij} w^i$$

$$v_j = A^{ij} w_i$$

The culmination of Einstein notation finally reveals itself when we can use it to write different products. I had previously written the inner product in this notation. For arbitrary vectors $\vec{x}$ and $\vec{y}$, the inner product is defined as

$$s = x_i y^i = y_i x^i,$$

the outer product is defined as

$$A^i_j = A^{ij} = A_{ij} = x^i y_j \neq y^i x_j,$$

and we can even define the cross product using the Levi-Civita symbol

$$(\vec{x} \times \vec{y})^i = \varepsilon^i_{jk} x^j y^k = \delta^{il} \varepsilon_{ljk} x^j y^k$$

where ε_{ljk} is the **Levi-Civita symbol** and δ^{il} is the **Kronecker delta.**

The last portion of Einstein notation does not have any formal or mathematical backing to it, but is simply a convention that is typically used to distinguish between certain vectors. Specifically, in regards to indices, Greek letters are used for indices when dealing with 4-vectors, in which case these indices can have values of 0, 1, 2, or 3. Conversely, Latin letters are used for indices when dealing with three vectors, in which case these indices can have values of 1, 2, or 3.

4.4 4-Vectors

It seems like an apt time to return to this, although I should be honest in that I've been quite torn-up with how to present this section. A natural place to start would be to define what a 4-vector is and how it differs from a normal Cartesian vector. It is easy to say that the 4-vector is simply a four dimensional vector. That is, a three dimensional Cartesian vector is one with three real elements in it. The 4-vector is the same with four elements instead of three, meaning that it must exist in four-dimensional space rather than three-dimensional space. However, there is a bit of a caveat to this.

The vectors that we are used to working with in two- and three-dimensions transform under the same metrics. Recall that the metric tensor is used to convert a contravariant vector into a covariant vector or vice versa. Thus, the metric tensor for a space is in some way a description of how the vector transforms. For our standard two- and three-dimensional Cartesian spaces, the metric tensors, known as the **Euclidean metrics**, are the same and are extensions of each other.

$$g_{2D} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad g_{3D} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

4-vectors by extension do not transform by the four-dimensional form of the Euclidean matrix. Instead, they transform using the **Minkowski space metric**.

$$g_{Minkowski} = \eta = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

Note that this metric tensor is not unique and it is also fine to take the negative of this tensor. However, we will only use this form of the Minkowski metric for our purposes.

Because we must have that our 4-vectors transform in this particular way, we cannot only have four real elements in our 4-vector. For example, a 4-vector cannot be the combination of temperature (corrected for units) and the three elements of a position vector, since temperature and position do not share the same relationship as defined by the Minkowski space metric. By extension, the Dirac spinors that we use do not transform in this way and are not 4-vectors. They aren't quite vectors either and I will devote a portion of this at the end to go a bit more in depth into what spinors are.

It should be explained what it means for a vector to transform in this way. The metric in this way ensures that, should the contravariant and covariant vectors be transformed, then their inner product would remain constant. For example, consider some matrix S that is the scaling matrix, where if applied to the contravariant vector scales all elements by some set scalar s.

$$S^b_a v^a = s v^b$$

The respectively scaled covariant vector can be found by applying the inverse of the matrix to the covariant vector

$$w_b \left(S^{-1}\right)^b_a = \frac{1}{s} w_a$$

It is clear to see that all three of the metric tensors provided maintains that the inner product of v^a and w_b are equal both before and after the scaling.

$$\langle Sw \mid Sv \rangle = w_b \left(S^{-1}\right)^b_a S^c_d v^d = \frac{1}{s} w_a s v^c = \frac{s}{s} w^a g_{ac} v^c = w^a g_{ac} v^c = \langle w \mid v \rangle$$

While I admit that this should work in every case, as all vectors should be eigenvectors of the scaling matrix, more complicated transformations, such as shearing, requires specific metrics to maintain the invariance of the inner product. In terms of our Minkowski metric, the metric tensor guarantees that the inner product of our 4-vectors are invariant under a Lorentz transformation. In other words, regardless of what frame the 4-vectors are expressed in, the inner product, using the Minkowski metric, is the same.

The typical example for a 4-vector is the **four-position**, which you may recall appears as follows:

$$x^\mu = \begin{bmatrix} x^0 \\ x^1 \\ x^2 \\ x^3 \end{bmatrix} = \begin{bmatrix} ct \\ x \\ y \\ z \end{bmatrix}$$

where the given representation is in Cartesian coordinates. In the four-position 4-vector, each element of the vector is a scalar. However, the elements of vectors do not need to be scalar values. In fact, even before we breach the topic of 4-vectors, there is a commonly used vector that uses non-scalar elements: the gradient. Recall that the gradient is a vector such that

$$\vec{\nabla} = \begin{bmatrix} \partial_x \\ \partial_y \\ \partial_z \end{bmatrix}$$

where the partials are distinctly operators and not scalars. Thus, it stands that we would be able to make a 4-vector using partials (or other non-scalar elements) in the same way. It then stands that the **four-derivative**, or the **four-gradient**, is the 4-vector whose elements are the derivatives with respect to each element of the four-position. In other words

$$\partial^\mu = \begin{bmatrix} \partial^0 \\ -\partial^1 \\ -\partial^2 \\ -\partial^3 \end{bmatrix} = \begin{bmatrix} \frac{1}{c}\partial_t \\ -\partial_x \\ -\partial_y \\ -\partial_z \end{bmatrix}$$

These two 4-vectors can be used to find a variety of other 4-vectors. For example, the **four-velocity** is found by taking the derivative of the four-position with respect to the proper time. In order to perform this calculation, it is useful to note that $dt = \gamma(\vec{u})\, d\tau$, where t is the time of the frame in which the four-position is taken in, $\gamma(\vec{u})$ is the Lorentz factor of a frame moving at a velocity $\vec{u}$, and τ is the proper time. Then, we express the four-velocity as follows

$$U^\mu = \partial_\tau x^\mu = \partial_t x^\mu \frac{\partial t}{\partial \tau} = \gamma(\vec{u}) \begin{bmatrix} c \\ u^1 \\ u^2 \\ u^3 \end{bmatrix} = \gamma(\vec{u}) \begin{bmatrix} c \\ u_x \\ u_y \\ u_z \end{bmatrix}$$

The **four-momentum** follows from the four-velocity by multiplying the quantity by the rest mass of the object.

$$P^\mu = m_0 U^\mu = \begin{bmatrix} \gamma(\vec{u})\, m_0 c \\ \gamma(\vec{u})\, m_0 u_x \\ \gamma(\vec{u})\, m_0 u_y \\ \gamma(\vec{u})\, m_0 u_z \end{bmatrix}$$

Then using the fact that $E = \gamma m_0 c^2$ and $p^a = \gamma m_0 v^a$, we can rewrite the four-momentum in more familiar quantities.

$$P^\mu = \begin{bmatrix} P^0 \\ P^1 \\ P^2 \\ P^3 \end{bmatrix} = \begin{bmatrix} E/c \\ p_x \\ p_y \\ p_z \end{bmatrix}$$

Other 4-vectors can be found, such as the four-force, the four-potential, the four-frequency, and so on, following the same logic and applying our typical classical physics ideas to the manipulation of 4-vectors. Doing so in this message would be tedious and not useful. However, it may be useful to discuss one final important 4-vector. The Dirac matrices or the gamma matrices are found when analyzing the Dirac equation for relativistic particles. Using these matrices, the Dirac equation can be written as

$$\left(i\hbar\left(\gamma^0\partial_0 + \gamma^1\partial_1 + \gamma^2\partial_2 + \gamma^3\partial_3\right) - mc\right)\psi = 0$$

It is clear that the individual gamma matrices are then tied to each element of the four-position. Thus, it would make sense that the four gamma matrices would then form a 4-vector of their own. To which, we get the following

$$\gamma^4 = \begin{bmatrix} \gamma^0 \\ \gamma^1 \\ \gamma^2 \\ \gamma^3 \end{bmatrix}$$

which may occur from our previous discussion that the elements of the 4-vectors do not have to be scalars.

4.5 Properties of 4-vectors

It should be noted that even after all of this, the 4-vectors still behave like vector objects. That is, they still belong to a vector space as must obey the following axioms:

1. Associativity under vector addition
   - $(u^\mu + (v^\mu + w^\mu) = (u^\mu + v^\mu) + w^\mu)$
2. Commutativity under vector addition
   - $(u^\mu + v^\mu = v^\mu + u^\mu)$
3. Existence of identity element of vector addition
   - $(v^\mu + 0^\mu = v^\mu)$
4. Existence of inverse elements of vector addition
   - $(v^\mu + (-v^\mu) = 0^\mu)$
5. Compatibility of scalar multiplication with field multiplication
   - $(a(bv^\mu) = (ab)v^\mu)$
6. Existence of identity element of scalar multiplication
   - $(1v^\mu = v^\mu)$

7. Distributivity of scalar multiplication with respect to vector addition
   - $(a\,(u^\mu + v^\mu) = au^\mu + av^\mu)$
8. Distributivity of scalar multiplication with respect to field addition
   - $((a+b)\,u^\mu = au^\mu + bu^\mu)$

As well as other aspects such as not being able to commute with matrices in general. One could still find matrices through which vectors may be able to commute by converting the vector to its covariant form, but the process of commutation is not guaranteed. However, one of the most important properties that we will use is the **invariance property**. In general, the inner product of two 4-vectors is simply invariant and does not hold physical meaning. Consider the four-position for example. The inner product of the four-position, which appears as follows

$$x_\mu x^\mu = x_0 x^0 - x_1 x^1 - x_2 x^2 - x_3 x^3 = c^2 t^2 - \vec{x}^2$$

may tell you whether or not the particular spacetime separation is space-like or time-like, but otherwise does not hold any more specific physical information. There are a number of inner products that do otherwise contain extra meaning or use. For example, the inner product of the four-gradient gives another operator which is

$$\partial_\mu \partial^\mu = \frac{1}{c^2}\partial_t^2 - \partial_x^2 - \partial_y^2 - \partial_z^2 = \frac{1}{c^2}\partial_t^2 - \nabla^2 = \Box$$

This operator, denoted as $\Box$, is known as the D'Alembertian operator. The four-velocity and the four-momentum have more physically significant and rather constant invariants. The inner product of the four-velocity is always the square of the speed of light.

$$U_\mu U^\mu = \gamma^2(\vec{u})\left(c^2 - u_x^2 - u_y^2 - u_z^2\right) = \gamma^2(\vec{u})\left(c^2 - \vec{u}^2\right) = \frac{c^2}{(c^2 - \vec{u}^2)}\left(c^2 - \vec{u}^2\right) = c^2$$

Meanwhile, the inner product of the four-momentum is always square of the rest mass of the particle, with correct units of scaling.

$$P_\mu P^\mu = m_0^2 U_\mu U^\mu = m_0^2 c^2$$

And can be used to derive the energy relationship with mass and momentum

$$P_\mu P^\mu = E^2/c^2 - \vec{p}^2 = m_0^2 c^2$$

Lastly, we can show that the gamma matrix 4-vector still holds as such an object by looking at the inner product of the vector:

$$\gamma_\mu \gamma^\mu = \left(\gamma^0\right)^2 - \left(\gamma^1\right)^2 - \left(\gamma^2\right)^2 - \left(\gamma^3\right)^2 = I_4 - (-I_4) - (-I_4) - (-I_4) = 4I_4$$

And since the identity matrix should not change with the frame, the inner product is indeed Lorentz invariant.

5 Computational Methods in Physics

Everything in this section is based on the excellent and quite accessible textbook *Numerical Methods for Ordinary Differential Equations : Initial Value Problems* by D.F. Griffiths and D.J. Higham. A PDF copy is available through the UBC library.

5.1 Introduction and Euler's Method

From a mathematical perspective, there are generally two steps to any given physics problem. You first model the problem using some set of equations, then attempt to solve those equations. Where the physics really occurs is in the first step, but some of the most pressing research questions actually concern the second. Almost no equations in physics are explicitly solvable, so it's up to us to come up with "reasonable" approximations. Of course, this raises many questions, primarily what "reasonable" means and how its evaluated. As an example of this, let's look at the simplest possible case, beginning with precisely defining the problems of interest.

Definition 5.1. *Let x, f be smooth real-valued functions. An ordinary differential equation (ODE) is one of the form*

$$x'(t) = f(t, x) \tag{56}$$

An initial value problem (IVP) is an ordinary differential equation, along with an initial condition of the form $x(t_0) = a$, where $a \in \mathbb{R}$.

At first glance, this definition seems extremely narrow. It excludes differential equations such as $x''(t) = x(t)$, ones that are simple to solve and certainly studied in your average ODEs course. However, as we shall later see, any higher order ODE can be turned into a first order *system* of ODEs, therby dodging this issue.

We could of course be in an excellent situation where we can exactly solve for $x(t)$, but the chances of this are quite low. This is where our numerical methods come in, starting with the simplest possible approximation.

Definition 5.2. *Let*

$$\begin{cases} x'(t) = f(t, x) \\ x(t_0) = \eta \end{cases}$$

be an initial value problem. An Euler method is an approximation of the solution of this IVP of the following form. Set $x_0 = \eta$, $f_0 = f(t_0, x_0)$ and fix some $h > 0$. For each $n \in \mathbb{N}$, define x_n, t_n, f_n recursively by

$$\begin{cases} x_{n+1} = x_n + hf_n \\ t_{n+1} = t_n + h \\ f_{n+1} = f(t_{n+1}, x_{n+1}) \end{cases}$$

Our approximation here is that $x(t_n) \approx x_n$, and $f(t_n, x) \approx f_n$.

Stated formally its a little complicated, but this is the basic Euler method you're familiar with. The reason for this complication is that we'll re-use this notation over and over to look at future numerical methods. To characterize how good our approximation is at any given time, we'll define the *global error* (GE) of the approximation by $e_n = x(t_n) - x_n$. Of course we can't calculate the global error; if we could then we would necessarily know the exact solution, so approximating would be rather silly. Instead, we want to put bounds on this error. Specifically, since $e_0 = 0$, we'd like to describe how this error grows or shrinks depending on our step size h, which intuitively should control the accuracy of our approximation. To do this, we borrow some notation from the dreaded computer scientists.

Definition 5.3. *Let f be a real-valued function. We say that $f \in O(x^p)$, where $p > 0$, if $\exists t_0, C > 0$ such that $|f(t)| \leq Ct^p$ for all $t \in (0, t_0)$.*

You may notice that this notation is actually reversed from the one used in computer science, but its for a good reason. While computer scientists care about behaviour for very large numbers, we instead care about behaviour for very small numbers. Using this, we can finally define what it means for our approximation method to be some sort of "baseline" of good.

Definition 5.4. *A numerical method converges to the solution $x(t)$ of an IVP at $t = t^*$ if, setting $t_n = t^*$, we get that $|e_n| \to 0$ as $h \to 0$. It converges at a p-th order if $e_n = O(h^p)$.*

A method that isn't *convergent* is essentially worthless, and all the methods we'll talk about are (at least usually) convergent. There's two important things to note here.

1. Converging at p-th order provides us an idea of how quickly the method converges. The bigger the value of $p > 0$, the quicker the convergence is.
2. Numerical methods do not converge or diverge universally, whether they converge is entirely dependent on the IVP given.

As a basic example of the second point, a numerical method cannot converge for IVPs with no unique solution. Euler methods are particularly good in this regard, they converge for any IVP with a unique solution! Depending on the form of the given ODE, we may also be able to specify the rate at which methods converge, the conditions on certain parameters in the ODE for convergence, on what interval of time the method converges, and so on.

5.2 Linear Methods

Euler's method is quite nice, but of course if it were the end of the story there would be no reason for me to bother writing these notes. The most natural extension of Euler's method is something called a *Linear Multistep Method* (LMM). The notation can quickly get dense, so we'll start with the simples of these methods.

Definition 5.5. *A 2-Step LMM is a numerical method of the form*

$$x_{n+2} + \alpha_1 x_{n+1} + \alpha_0 x_n = h(\beta_2 f_{n+2} + \beta_1 x_{n+1} + \beta_0 x_n) \tag{57}$$

Example 5.6. *Simpson's rule (which you should have seen in first year calculus) is a 2-step LMM with $\alpha_1 = 0, \alpha_0 = -1, \beta_2 = \beta_0 = \frac{1}{3}, \beta_1 = \frac{4}{3}$.*

There's a lot of small details to unpack in this equation, so let's do so one step at a time.

1. This method needs 2 initial conditions, one at t_0 and one at t_1. To use this on a standard IVP, the general method is to first approximate the initial conditions at t_1 using the Euler method, then proceed using the 2-step LMM.
2. If $\beta_2 \neq 0$, we run into an issue. This equation is nominally solving for x_{n+2}, but f_{n+2} depends on x_{n+2}. This leads to what we call an *implicit* (rather than *explicit*) method, x_{n+2}, f_{n+2} must be solved for simultaneously. This brings both advantages and disadvantages, which we'll go over later.

When analyzing these methods, we like to use some more advanced tools.

Definition 5.7. *The linear difference operator of a 2-step LMM is given by*

$$\mathcal{L}_h(x) = x(t+2h) + \alpha_1 x(t+h) + \alpha_0 x(t) - h(\beta_2 x'(t+2h) + \beta_1 x'(t+h) + \beta_0 x'(t)) \tag{58}$$

Note that we've switched out f for x' here, just to make the equation a little less messy. Essentially, this linear operator gives us a measurement of how bad our approximation failed at any given step. Like the GE, it cannot be computed exactly unless the exact solution was known. Also like our GE, we want to put a bound on this "local" error using the step size h.

Definition 5.8. *A two-step LMM is said to be consistent of order $p > 0$ if $\mathcal{L}_h(x(t)) = O(h^{p+1})$ for all sufficiently smooth[14] functions x.*

The nice thing about all these definitions is that they actually generalize quite well to any LMM method, which we define in general now.

Definition 5.9. *For any $k \in \mathbb{N}$ a k-step LMM is one of the form*

$$x_{n+k} + \sum_{j=0}^{k-1} \alpha_j x_{n+j} = h \sum_{j=0}^{k} \beta_j f_{n+j} \tag{59}$$

where $\alpha_j, \beta_j \in \mathbb{R}$.

Again, a method is *implicit* if $\beta_k \neq 0$, and *explicit* otherwise. Let's take some time to go over examples of these methods.

Example 5.10. *Euler's method is a 1-step LMM with $\alpha_0 = -1, \beta_1 = 0, \beta_0 = 1$. The reverse Euler method (an implicit variant) is a 1-step LMM with $\alpha_0 = -1, \beta_1 = 1, \beta_0 = 0$.*

Example 5.11. *The Trapezoid rule is a 2-step LMM with $\alpha_1 = 0, \alpha_0 = -1, \beta_2 = \beta_0 = 1, \beta_1 = 0$.*

Example 5.12. *There is a family of explicit LMMs called the Adams-Bashford (AB) methods. These are of the form*

$$x_{n+k} - x_{n+k-1} = h \sum_{j=0}^{k-1} \beta_j f_{n+j} \tag{60}$$

where the β_j are chosen according to a certain rule which will be explained in the next section. The first AB method is just the Euler method, the second is

$$x_{n+2} = x_{n+1} + h\Big(\frac{3}{2} f_{n+1} - \frac{1}{2} f_n\Big)$$

and the third

$$x_{n+3} = x_{n+2} + h\Big(\frac{23}{12} f_{n+2} - \frac{16}{12} f_{n+1} + \frac{5}{12} f_n\Big)$$

As you can see, these quickly become unruly.

Example 5.13. *There is a family of implicit LMMs called the Milne-Simpson methods. These are of the form*

$$x_{n+k} - x_{n+k-2} = h \sum_{j=0}^{k} \beta_j f_{n+j} \tag{61}$$

where again the β_j are chosen according to certain rules explained in the next section. The first of these is just Simpson's rule.

Again, it's good to note that a k-step LMM will need k initial conditions, which we generally get by using the Euler method on the given initial condition.

[14] Don't you dare ask what exactly sufficiently smooth means. You're a physicist, everything is smooth!

5.3 Convergence, Consistency, and Stability

In this section, we'll dive more into how we evaluate the quality of LMMs. Perhaps the best feature of LMMs, besides being (relatively) easy to write down and compute, is how well-characterized its performance is. Let's start by defining how we'll evaluate our methods.

Definition 5.14. *A LMM with well-approximated*[15] *initial values is called convergent if for all IVPs with a unique solution $x(t)$ on $t \in [t_0, t_f]$*

$$\lim_{\substack{h\to 0 \\ nh=t^*-t_0}} x_n = x(t^*), \forall t^* \in [t_0, t_f] \tag{62}$$

This is essentially a generalization of our condition from the introduction. We can consider this a characterization of the GE, while consistency is a characterization of what we call the *local truncation error* (LTE). Our first result of this section will be focused on relating convergence and consistency. To do this, we need to start with the following definitions.

Definition 5.15. *The first and second characteristic polynomials of a k-step LMM are*

$$\rho(r) = r^k + \sum_{j=0}^{k-1} \alpha_j r^j \tag{63}$$

$$\sigma(r) = \sum_{j=0}^{k} \beta_j r^j \tag{64}$$

At first these seem completely random, but they do connect to what we've been looking at so far. Indeed, if we take a look at the linear difference operator for a k-step LMM

$$\mathcal{L}_h(x) = x(t+kh) + \sum_{j=0}^{k-1} \alpha_k x(t+jh) - h\sum_{j=0}^{k} \beta_j x'(t+jh)$$

we can Taylor expand all the terms with arguments of the form $(t+jh)$ about t to get

$$\mathcal{L}_h(x) = \sum_{j=0}^{p+1} C_j h^j \frac{d^j x}{dt^j}(t) + O(h^{p+2})$$

for any $p \in \mathbb{N}$. In this expansion, one can get that $C_0 = \rho(1), C_1 = \rho'(1) - \sigma(1)$, so the characteristic polynomials actually arise from this expansion![16] We can pull another conclusion from this expansion as well : a LMM is consistent of order p if and only if $C_0 = C_1 = \cdots = C_p = 0$. In this case, we call C_{p+1} the *error constant*.

Example 5.16. *The coefficients for the AB methods outlined in the previous section are chosen such that they have consistency order k, and the coefficients in the Milne-Simpson methods such that the maximum possible consistency order is achieved.*

Now, back to arcane polynomial definitions.

Definition 5.17. *A polynomial satisfies the root condition if all of its root lie on or within the unit circle in $\mathbb{C}$, with those on the circle having multiplicity 1. A LMM is zero-stable (or stable) if its first characteristic polynomial satisfies the root condition.*

[15] I'm sure there's a precise definition of this somewhere

[16] This may not be an entirely satisfactory explanation for many of you, but it's the best I can give here.

You can maybe start to see where we're going here, there's two (I'm told quite remarkable) theorems that relates the stability, consistency, and convergence of a LMM.

Theorem 5.18. *A LMM is convergent if and only if it is consistent and stable.*

Theorem 5.19. *The consistency order p of a stable k-step LMM satisfies*

1. *$p \leq k+2$ if k is even*
2. *$p \leq k+1$ if k is odd*
3. *$p \leq k$ if the method is explicit*

This gives us our first hint of the benefits of implicit methods : they can be have a higher level of consistency than explicit ones. The first of these theorems also provides us with a sort of interpretation of zero-stability : it's the guarantee that the LTE (local error) propagates in such a way that the GE (global error) remains small.

We take a break now to travel down a seemingly very silly and specific road, which will turn out to give a quite important result. In particular here, we're going to focus on ODEs of the form

$$x'(t) = \lambda x(t) \tag{65}$$

where $\lambda \in \mathbb{C}$ has a negative real part. Now of course we know the explicit solution to this, it's simply

$$x(t) = Ce^{\lambda t}$$

So why bother at all? Well, this is a useful model system for us to measure how good our approximations are in the long term, since we know that $x \to 0$ as $t \to \infty$. With that being said, we make the following definition.

Definition 5.20. *A LMM is absolutely stable if, when applied to $x'(t) = \lambda x(t)$ of the above for form a given "complex" step $\hat{h} = h\lambda$, where $h > 0$, its solution tends to zero as $n \to \infty$ for any initial conditions.*

Essentially, it needs to perform well no matter what initial conditions its given. Another way of looking at this (although technically not identical) is that the GE is damped over large time scales.

Definition 5.21. *The set of values R in the complex $\hat{h}$ plane for which a LMM is absolutely stable is called its region of absolute stability. Its region of absolute stability is $R \cap (-\infty, 0)$.*

With this, we can define a much stronger form of stability.

Definition 5.22. *A numerical method is A-stable if its region of stability is the entire left half-plane.*

The following theorem demonstrates just how severe (and strict) this condition is.

Theorem 5.23.

1. *There is no A-stable explicit LMM*
2. *An implicit A-stable LMM cannot have order $p > 2$*
3. *The order-two A-stable LMM with the error constant of smallest magnitude is the trapezoid rule*

That's right, the trapezoid rule was actually the most stable method this entire time. In fact, A-stable methods seem to generally perform better than all other methods, even on non-linear problems. This provides further motivation for ever using an implicit method over an explicit method (despite the computational cost), they come with a large amount of added stability.

5.4 Systems of ODEs

Definition 5.24. *Let $\vec{x}(t)$ be a smooth function from $\mathbb{R}$ to $\mathbb{R}^n$. A system of ODEs is an equation of the form*

$$\frac{d\vec{u}}{dt} = A(t, \vec{u})\vec{u} \tag{66}$$

where $A(t, \vec{u}) : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^{n \times n}$ is a matrix. If A is constant, we call the system linear.

We can also convert any ordinary differential equation of any order into a system, such at $x'' + x' = x$, by setting $u_i = \frac{d^{(i)}x}{dt^i}$. All of the LMMs we've talked about so far can be applied to systems of ODEs, simply by applying them to each of the equations in the system. Generally, these get written as

$$\vec{x}_{n+k} + \sum_{j=0}^{k-1} \alpha_j \vec{x}_{n+j} = h \sum_{j=0}^{k} \beta_j \vec{f}_{n+j} \tag{67}$$

where our notation here follows that on the previous uses of LMMs. However, with larger systems, we run into problems such as stiffness.

Definition 5.25. *A first order system of ODEs is called stiff if a small perturbation to the initial conditions (letting "initial conditions" be at any point in the domain of the problem) causes a large change in the solution.*

You'll note that this is an imprecise definition, and this is entirely on purpose. Stiffness in general has no widely accepted definition, it's more a word for expressing when systems of differential equations need to be approximated with a very small step size, blowing up otherwise. To handle these types of systems, we'd ideally like to use implicit methods. The problem is that implicit methods tend to accumulate more rounding errors and become less and less efficient as systems grow. As a sort of compromise, we introduce the *predictor-corrector* methods.

Definition 5.26. *A predictor-corrector method uses a pair of LMMS, one implicit*

$$\vec{x}_{n+k} + \sum_{j=0}^{k-1} \alpha_j \vec{x}_{x+j} = h \sum_{j=0}^{k} \beta_j \vec{f}_{n+j} \tag{68}$$

and one explicit

$$\vec{x}_{n+k} + \sum_{j=0}^{k-1} \alpha_j^* \vec{x}_{x+j} = h \sum_{j=0}^{k} \beta_j^* \vec{f}_{n+j} \tag{69}$$

and proceeds via the following steps.

1. *Predict $\vec{x}_{n+k}$ using Equation 69, call this $\vec{x}^{[0]}_{n+k}$.*
2. *Calculate $\vec{f}_{n+k}$ using $\vec{f}(t, \vec{x})$ and $\vec{x}^{[0]}_{n+k}$, call it $\vec{f}^{[0]}_{n+k}$.*
3. *Predict $\vec{x}_{n+k}$ using $\vec{f}^{[0]}_{n+K}$ and Equation 68.*
4. *Calculate $\vec{f}_{n+k}$ using $\vec{f}(t, \vec{x})$ and $\vec{x}_{n+k}$.*
5. *Repeat.*

5.5 Runge-Kutta Methods

There's one more common method of approximation that we haven't yet talked about, *Runge-Kutta Methods.*

Definition 5.27. *An s-stage Runge-Kutta (RK(s)) method is written in the form*

$$x_{n+1} = x_n + h\sum_{i=1}^{s} b_i k_i \tag{70}$$

where

$$k_i = f\Big(t_n + c_i h, x_n + h\sum_{j=1}^{s} a_{i,j}k_j\Big), c_i = \sum_{j=1}^{s} a_{i,j} \tag{71}$$

These are often represented in the form of a Butcher array

$$\begin{array}{c|cccc} c_1 & a_{1,1} & a_{1,2} & \cdots & a_{1,s} \\ c_2 & a_{2,1} & a_{2,2} & \cdots & a_{2,s} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_s & a_{s,1} & a_{s,2} & \cdots & a_{s,s} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array}$$

This is an extremely obtuse definition, so let's look at some common examples.

Example 5.28. *The Euler method is a 1-stage RK method with Butcher array*

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

Example 5.29. *The backwards Euler method is a 1-stage RK method with Butcher array*

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

It can be written as

$$x_{n+1} = x_n + hk_1, k_1 = f(t+h, x_n + hk_1)$$

we can note that this method is extremely implicit.

Example 5.30. *A general 2-stage explicit RK method is given by the Butcher array*

$$\begin{array}{c|cc} 0 & 0 & 0 \\ a & a & 0 \\ \hline & b_1 & b_2 \end{array}$$

It can be written as

$$\begin{cases} k_1 = f(t_n, x_n) \\ k_2 = f(t_n + ah, x_n + hak_1) \\ x_{n+1} = x_n + h(b_1k_1 + b_2k_2) \end{cases}$$

In general, RD methods are explicit if and only if their Butcher arrays are *strictly upper triangular*, that is upper triangular with a zero diagonal.

Example 5.31. *The RK4 method is a 4-stage RK method with Butcher array*

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

It can be written as

$$\begin{cases} x_{n+1} = x_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) \\ k_1 = f(t_n, x_n) \\ k_2 = f\left(t_n + \frac{h}{2}, x_n + \frac{hk_1}{2}\right) \\ k_3 = f\left(t_n + \frac{h}{2}, x_n + \frac{hk_2}{2}\right) \\ k_4 = f(t_n + h, x_n + hk_3) \end{cases}$$

This is probably the most commonly used of the non-Euler RK methods.

With all these examples in pocket, let's define what *consistency* means for RK methods.

Definition 5.32. *The LTE of an RK method is defined as $T_{n+1} = x(t_{n+1}) - x_{n+1}$. An RK method is said to be consistent of order p, $p > 0$, if $T_{n+1} = O(h^{p+1})$.*

This lines up with our prior definition of consistency, but unlike in LMMs it's extremely difficult to determine which RK methods are consistent, and to what order. We do have the following results.

Theorem 5.33. *All RK methods with 4 steps or less have at most a consistency order of the number of steps in the method. In particular.*

1. *The only consistent 1-stage RK method is the Euler method.*
2. *Explicit RK(2) methods are consistent of order 1 if and only if $b_1 + b_2 = 1$. They are consistent of order 2 if and only if $b_1 + b_2 = 1$ and $ab_2 = \frac{1}{2}$.*
3. *The RK4 method is a consistent RK(4) method of order 4.*

Results for higher than order/stage 4 are generally not know, and if you encounter a problem that needs that complicated of a method you should probably be simplifying your problem anyways. We can also note something interesting from the second point : an explicit, consistent, RK(2) method of maximal order can be entirely specified by a choice of $a \neq 0$. Finally, we note that we can of course apply all of these methods to systems of ODEs as well, extending them in the same way as we did LMMs.

5.6 Adaptive Steps

The final technical thing we'll talk about here is adaptive step size. The motivation for this is fairly clear : systems of equations may be quite stiff at some times and less stiff at later times. Keeping the small step size the whole way through would be wasting quite a bit of time and computing power, so it'd be better if we could adapt our step size to be only as small as is strictly required. There's no one algorithm for doing this, and many of the ones used are quite informal. I'll be

presenting one here that's worked well for me before, and works well with the numerical methods we've been talking about.
Assume that we have a numerical method consistent of order p. The procedure we use is as follows.

1. Choose an initial step size, which we call h_{new}, and a tolerance we call tol.
2. Set $h_n = h_{new}$, and calculate provisional values of $x_{n+1}, t_{n+1} = t_n + h_n$ using the numerical method.
3. Estimate a value of T_{n+1}, which we'll call $\hat{T}_{n+1}$, using the values from the previous step.
4. Set $h_{new} = h_n \left| \frac{\text{tol}}{\hat{T}_{n+1}} \right|^{\frac{1}{p_1}}$.
5. If $|\hat{T}_{n+1}| > \text{tol}$, return to step 2 and re-try with the current h_{new}. Otherwise, return to step 2 and calculate the next step using the current h_{new}.

There's still a couple things to deal with here. First, the tolerance. This is chosen (arbitrarily) by the user. A lower tolerance requires error to be lower to avoid re-sizing steps. Second, there's the question of how to estimate the error in our calculation. There's many ways to do this, but I'll present some standard ones for common methods here. For the Euler method, it's common to take the approximation

$$\hat{T}_{n+1} = \frac{h_n}{2}(x'_{n+1} - x'_n)$$

For RK methods which are consistent of order p, it's common to run the calculation again with an RK method consistent of order $p+1$, and take

$$\hat{T}_{n+1} = x_{n+1}^{(p+1)} - x_n^{(p)}$$

5.7 Practical Tips

I'll end this section by giving some tips on how to use numerical methods in physics.

1. Be lazy, use the simplest possible method to make your approximation.
2. Be careful of rounding errors using adaptive steps : in particular, put a lower bound on how small your step size can get. Otherwise, the simulation may never end for particularly stiff systems.
3. When in doubt, use RK4.
4. If that doesn't work, use a predictor-corrector method.
5. If your system is very stiff, don't believe in it blindly to model average behaviour. The stiffness of the system is often an indication that actual outcomes can vary wildly about the average.