Hyperbolic Trigonometry and Special Relativity

Timothy Magee

Hollins University Summer Term 2025

1 Background

1.1 The Euclidean Distance Formula

In usual, everyday geometry – the sort you teach to high school students – if we want to describe the distance between a pair of points in a Cartesian coordinate system, we use the *Euclidean distance formula.* In two dimensions, the Euclidean distance formula is just the Pythagorean theorem. Let's call our points $\mathbf{p}$ and $\mathbf{q}$ and write $\mathbf{p} = p_1\mathbf{e}_1 + p_2\mathbf{e}_2$ and $\mathbf{q} = q_1\mathbf{e}_1 + q_2\mathbf{e}_2$. Here, $\{\mathbf{e}_1, \mathbf{e}_2\}$ is the basis we've chosen for our Cartesian coordinate system, and their coefficients p_i and q_i are the *components* of $\mathbf{p}$ and $\mathbf{q}$ in this coordinate system. Then the Euclidean distance between $\mathbf{p}$ and $\mathbf{q}$ is given by

$$\mathbf{d}_{\mathrm{E}}(\mathbf{p}, \mathbf{q})^2 = (q_1 - p_1)^2 + (q_2 - p_2)^2.$$

See Figure 1.

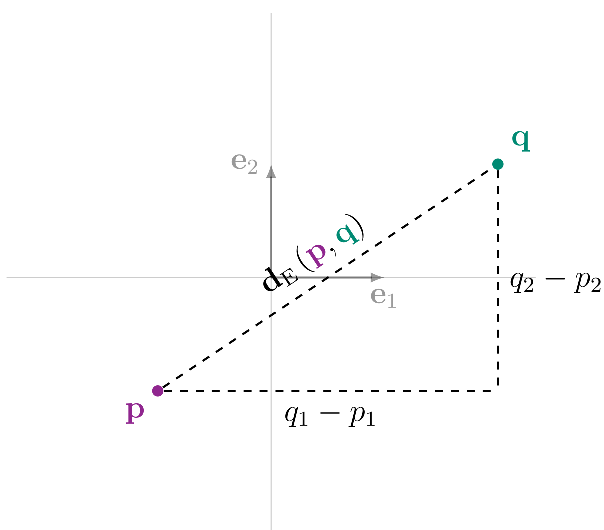

Figure 1: The Euclidean distance between a pair of points $\mathbf{p}$ and $\mathbf{q}$ in 2-space is given by the Pythagorean theorem, with side lengths as shown above.

Notice that horizontal and vertical side lengths in Figure 1 are themselves Euclidean distances. Let $\mathbf{r}$ be the intersection of the horizontal and vertical sides, *i.e.* $\mathbf{r} = q_1\mathbf{e}_1 + p_2\mathbf{e}_2$. Then $\mathbf{d}_\mathrm{E}(\mathbf{p}, \mathbf{r}) = q_1 - p_1$ and $\mathbf{d}_\mathrm{E}(\mathbf{q}, \mathbf{r}) = q_2 - p_2$. Compare to Figure 2.

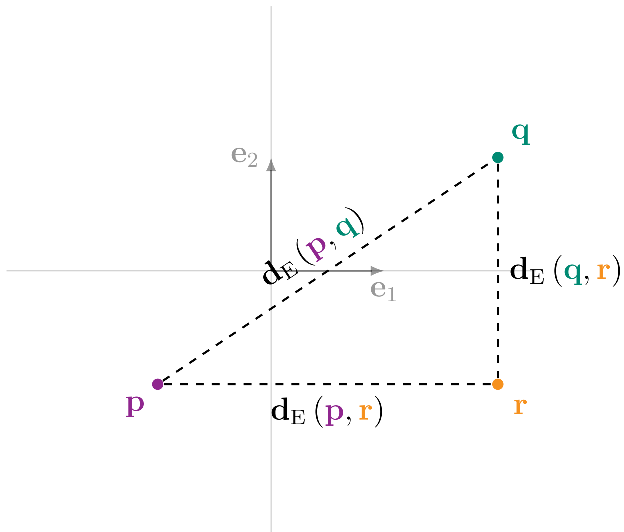

Figure 2: *All* side lengths are Euclidean distances.

In higher dimensions, the Euclidean distance formula just has more terms on the right hand side – each a square of a difference of $\mathbf{q}$ and $\mathbf{p}$ components.

Definition 1. In dimension n, the Euclidean distance formula is given by

$$\mathbf{d}_\mathrm{E}(\mathbf{p}, \mathbf{q})^2 = (q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2 = \sum_{i=1}^{n} (q_i - p_i)^2.$$

You can think of this as repeated application of the Pythagorean theorem.

Question 2. Why?

Hint: *Start with the $n = 3$ case, where you should be able to picture things. Then try to reduce the $n = d$ case to the $n = d - 1$ case for arbitrary d.*

Euclidean distance and spheres are intimately related.

Definition 3. The $(n-1)$-sphere of radius R centered at a point $\mathbf{p}$ in $\mathbb{R}^n$ is

$$\{\mathbf{q} \in \mathbb{R}^n \mid \mathbf{d}_\mathrm{E}(\mathbf{p}, \mathbf{q}) = R\}.$$

We will mostly work in two dimensions in this course. In this case, we are dealing with 1-*spheres*, better known as circles.

1.2 Euclidean Transformations

We chose a particular Cartesian coordinate system in order to write the Euclidean distance formula. This may seem weird. The length of a string shouldn't depend on our coordinate system, right?[1] For instance, what if we were to take our original coordinate system and rotate it, or re-center it so a different point is viewed as the origin? Surely this should not affect $\mathbf{d}_{\mathrm{E}}(\mathbf{p}, \mathbf{q})$?

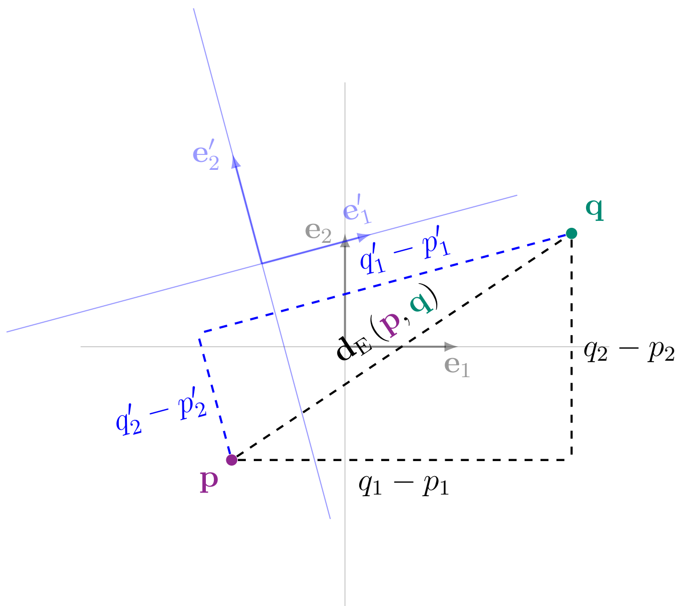

Figure 3: The Euclidean distance $\mathbf{d}_{\mathrm{E}}(\mathbf{p}, \mathbf{q})$ is the same in both of these coordinate systems.

We will make this idea precise below with a more detailed discussion of what goes into the Euclidean distance formula.

First, when we discuss Cartesian coordinate systems, we think of the defining basis vectors as *unit vectors*, and we think of distinct basis vectors as being *orthogonal* to each other. That is, we view $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ as an *orthonormal basis*. So, although we simply wrote $\mathbb{R}^n$ for our ambient space in Section 1.1, we were intrinsically assuming this ambient space had enough structure for the concept of an orthonormal basis to be meaningful. We need it to come equipped with an *inner product*.

Definition 4. Let V be a real vector space. An *inner product* on V is a map

$$\langle \,\cdot\, , \,\cdot\, \rangle : V \times V \to \mathbb{R}$$

such that for all $\mathbf{p}$, $\mathbf{q}$, $\mathbf{r}$ in V and a, b in $\mathbb{R}$:

[1] We'll have to revisit this thought later, but let's stay in the good old world of Newtonian mechanics for now.

(1) $\langle \mathbf{p}, \mathbf{q} \rangle = \langle \mathbf{q}, \mathbf{p} \rangle$ *(symmetry)*

(2) $\langle a\mathbf{p} + b\mathbf{q}, \mathbf{r} \rangle = a\langle \mathbf{p}, \mathbf{r} \rangle + b\langle \mathbf{q}, \mathbf{r} \rangle$ *(linearity)*

(3) $\langle \mathbf{p}, \mathbf{p} \rangle > 0$ if $\mathbf{p} \neq \mathbf{0}$ *(positive-definiteness)*

Example 5. The dot product on an n dimensional real vector space V is an inner product. Recall that that if V comes with an orthonormal basis $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$, then given $\mathbf{p} = \sum_{i=1}^n p_i \mathbf{e}_i$ and $\mathbf{q} = \sum_{i=1}^n q_i \mathbf{e}_i$, we have

$$\mathbf{p} \cdot \mathbf{q} = \sum_{i=1}^n p_i q_i.$$

Alternatively, if we wish to express the dot product in a coordinate-free way – no basis needed at all – we can describe the dot product as follows:

$$\mathbf{p} \cdot \mathbf{q} = \|\mathbf{p}\| \|\mathbf{q}\| \cos(\theta),$$

where θ is the angle between $\mathbf{p}$ and $\mathbf{q}$.

Let $\mathbf{p} = \sum_{i=1}^n p_i \mathbf{e}_i$, $\mathbf{q} = \sum_{i=1}^n q_i \mathbf{e}_i$, and $\mathbf{r} = \sum_{i=1}^n r_i \mathbf{e}_i$ be in V and let a and b be in $\mathbb{R}$. Then

$$\begin{aligned} \mathbf{p} \cdot \mathbf{q} &= \sum_{i=1}^n p_i q_i \\ &= \sum_{i=1}^n q_i p_i \\ &= \mathbf{q} \cdot \mathbf{p}. \end{aligned}$$

Here, the blue equality is by commutativity of multiplication in $\mathbb{R}$. So, the dot product is symmetric. Next,

$$\begin{aligned} (a\mathbf{p} + b\mathbf{q}) \cdot \mathbf{r} &= \sum_{i=1}^n (ap_i + bq_i) r_i \\ &= \sum_{i=1}^n (ap_i r_i + bq_i r_i) \\ &= a \sum_{i=1}^n p_i r_i + b \sum q_i r_i \\ &= a\mathbf{p} \cdot \mathbf{r} + b\mathbf{q} \cdot \mathbf{r}, \end{aligned}$$

by distributivity, which shows linearity. Finally, if $\mathbf{p} \neq \mathbf{0}$, then

$$\mathbf{p} \cdot \mathbf{p} = \sum_{i=1}^n p_i^2.$$

Since $\mathbf{p} \neq 0$, some component p_i must be non-zero. If $p_i \neq 0$, then $p_i^2 > 0$. It follows that

$$\sum_{i=1}^{n} p_i^2 > 0,$$

and the dot product is positive-definite.

We can now revisit the Euclidean distance formula using the dot product. First, we can make sense of our previous description in terms of the *orthonormal basis* $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$. When we say that the $\mathbf{e}_i$'s are *unit vectors* in a real vector space V equipped with an inner product $\langle \cdot\, , \, \cdot \rangle$, we mean that $\langle \mathbf{e}_i, \mathbf{e}_i \rangle = 1$ for all i.[2] When we say that $\mathbf{e}_i$ and $\mathbf{e}_j$ are orthogonal to each other, we mean $\langle \mathbf{e}_i, \mathbf{e}_j \rangle = 0$. In particular, we can talk about orthonormal bases for real vector spaces that come equipped with the dot product.

Next, we can also use the dot product to give a coordinate-free description of the Euclidean distance formula.

Definition 6. Let $(V, \cdot)$ be a real vector space together with the dot product on V. For any pair of points $\mathbf{p}$, $\mathbf{q}$ in V, the Euclidean distance $\mathbf{d}_{\mathrm{E}}(\mathbf{p}, \mathbf{q})$ is given by

$$\mathbf{d}_{\mathrm{E}}(\mathbf{p}, \mathbf{q}) = \sqrt{(\mathbf{q} - \mathbf{p}) \cdot (\mathbf{q} - \mathbf{p})}.$$

The linear maps $V \to V$ that preserve the dot product on V are called *orthogonal transformations*:

Definition 7. Let $(V, \cdot)$ be a real vector space together with the dot product on V. A linear map $f : V \to V$ is called an *orthogonal transformation* if for every pair of vectors $\mathbf{p}$, $\mathbf{q}$ in V, we have

$$f(\mathbf{p}) \cdot f(\mathbf{q}) = \mathbf{p} \cdot \mathbf{q}.$$

The set of orthogonal transformations on V is denoted $O(V)$.

In fact, $O(V)$ is not just a set – it is a *group* under composition.

Definition 8. A *group* is a non-empty set G together with a *binary operation* on G, which will denote $\cdot$,

$$\cdot : G \times G \to G$$

satisfying the following properties.

(1) For all $g_1, g_2, g_3 \in G$, we have $g_3 \cdot (g_2 \cdot g_1) = (g_3 \cdot g_2) \cdot g_1$.
(Associativity)

[2] Technically, we really mean $\sqrt{\langle \mathbf{e}_i, \mathbf{e}_i \rangle} = 1$, but it amounts to the same thing.

(2) There exist an element $\mathrm{id}_G \in G$, such that for all $g \in G$, we have $\mathrm{id}_G \cdot g = g = g \cdot \mathrm{id}_G$. *(Identity element)*

(3) For all $g \in G$, there is an element $g^{-1} \in G$ satisfying $g \cdot g^{-1} = \mathrm{id}_G = g^{-1} \cdot g$. *(Inverses)*

If both f and g preserve the dot product on V, clearly their composition does as well. That is, composition is binary operation on $O(V)$:

$$\circ : O(V) \times O(V) \to O(V).$$

Function composition is associative – given three functions f, g, and h from V to V, we have $h \circ (g \circ f) = (h \circ g) \circ f$.[3] The identity map on V is clearly also the identity for function composition: $\mathrm{id}_V \circ f = f = f \circ \mathrm{id}_V$. Moreover, if f preserves the dot product on V, it must be invertible. If f were not invertible, it would have some non-trivial kernel. That is, f would map some non-zero $\mathbf{p}$ to $\mathbf{0}$. But $\mathbf{p}$ has non-zero dot product with elements of V (for instance, with $\mathbf{p}$ itself), while the dot product of $\mathbf{0}$ with any element of V is 0. So the dot-product-preserving linear maps on V must be invertible. And the inverse must also preserve the dot product:

$$f(\mathbf{p}) \cdot f(\mathbf{q}) = \mathbf{p} \cdot \mathbf{q}$$

implies

$$f^{-1}(f(\mathbf{p})) \cdot f^{-1}(f(\mathbf{q})) = \mathbf{p} \cdot \mathbf{q} = f(\mathbf{p}) \cdot f(\mathbf{q}).$$

This is *almost* the set of transformations of V that preserve Euclidean distance. Note that translations do not affect Definition 6, since $(\mathbf{q} + \mathbf{x}) - (\mathbf{p} + \mathbf{x}) = \mathbf{q} - \mathbf{p}$.

Definition 9. Let $(V, \cdot)$ be a real vector space together with the dot product on V. We say a transformation $f : V \to V$ (not necessarily linear) is a *Euclidean transformation* if for every pair of points $\mathbf{p}$, $\mathbf{q}$ in V, we have

$$\mathbf{d}_{\mathrm{E}}\left(f(\mathbf{p}), f(\mathbf{q})\right) = \mathbf{d}_{\mathrm{E}}\left(\mathbf{p}, \mathbf{q}\right).$$

The set of Euclidean transformations of V is denoted $E(V)$.

As with the orthogonal transformations, $E(V)$ is not just a set – it is a group under composition. Euclidean transformations can all be expressed as a composition of rotations, reflections, and translations.

[3] If this isn't completely familiar, it's a valuable exercise to show that associativity follows from the definition of function composition.

1.3 Complex Numbers and Trigonometry

A basic understanding of complex numbers is very useful for understanding trigonometry, and a basic understanding of trigonometry is essential for understanding complex numbers. So, in this section we will review these two subjects together and highlight their interplay.

1.3.1 Complex Numbers

The real numbers have a serious short-coming that leads a lot of mathematicians to work with other number systems. The problem is that $\mathbb{R}$ is not *algebraically closed*. This means that there are (non-constant) polynomials with coefficients in $\mathbb{R}$ that have no real roots.

Example 10. Consider the polynomial $f(x) = x^2 + 1$. A real number α is a *root* of f if $f(\alpha) = 0$. This would mean $\alpha^2 = -1$, which isn't satisfied for any real number α.

This causes no end of headaches in fields like algebraic geometry, where solution sets of polynomial equations are studied. To resolve this issue, algebraic geometers tend to work over the *algebraic closure* of $\mathbb{R}$ – the *complex numbers* $\mathbb{C}$ – rather than working over $\mathbb{R}$. To form $\mathbb{C}$, we define i to be a root of the polynomial f of Example 10, and we *adjoin* it to $\mathbb{R}$. That is, we define

$$\mathbb{C} := \mathbb{R}[i] = \{a + ib \mid a, b \in \mathbb{R}\},$$

where i is a root of the polynomial $f(x) = x^2 + 1$. The fact that $\mathbb{C}$ is algebraically closed is a very important result in mathematics. It is known as the *Fundamental Theorem of Algebra*, and it has many different flavors of proof – some algebraic, some analytic, some topological, some geometric.

Note that we have made a choice in the definition of $\mathbb{C}$ above. We defined i to be a root of $f(x) = x^2 + 1$, but this polynomial has two complex roots. If i is a root, then so is $-i$. The complex number $a - ib$ is called the *complex conjugate* of $a + ib$, and it is useful for describing the magnitude of a complex number. We have constructed $\mathbb{C}$ as a copy of $\mathbb{R}^2$ generated by a real axis and an imaginary axis which are orthogonal to each other. The magnitude of $a + ib$, denoted $\|a + ib\|$, is $\sqrt{(a + ib)(a - ib)} = \sqrt{a^2 + b^2}$. Note that this is the Euclidean distance between $a + ib$ and the origin.

Notation 11. If c is a complex number, its complex conjugate is denoted $\bar{c}$.

1.3.2 Trigonometry

Trigonometry is the study of relationships among side lengths of triangles and arc lengths of circles. For instance, consider the following figure.

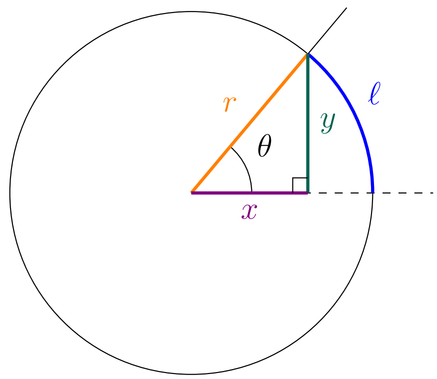

Figure 4: The diagram used to define trigonometric functions.

Here the angle θ (in radians) is defined to be the ratio of lengths $\theta = \frac{\ell}{r}$. The three most commonly used trigonometric functions are sine, cosine, and tangent. These are defined as follows:

- $\sin(\theta) = \frac{y}{r} = \frac{\text{opp}}{\text{hyp}}$
- $\cos(\theta) = \frac{x}{r} = \frac{\text{adj}}{\text{hyp}}$
- $\tan(\theta) = \frac{y}{x} = \frac{\text{opp}}{\text{adj}}$

1.3.3 Interplay of Complex Numbers and Trigonometry

Definition 12. Let $\Bbbk$ be a field (perhaps $\mathbb{R}$ or $\mathbb{C}$) and consider a smooth function $f : \Bbbk \to \Bbbk$. The *Taylor series of f at a* is

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

If f can be represented as a power series in a neighborhood of a, then it *must* be this series. To see this, suppose f can be represented as a power series in a neighborhood of a and write

$$f(x) = \sum_{k=0}^{\infty} c_k (x-a)^k$$

in this neighborhood for some yet-to-be-determined coefficients c_k. Evaluating both the left hand side and right hand side at $x = a$ gives $f(a) = c_0$. Next, if we first differentiate both sides with respect to x and then evaluate at $x = a$, we find $f'(a) = c_1$. More generally, if we were to take the k^{th} derivative of both sides and evaluate at $x = a$, we would obtain $f^{(k)}(a) = k!\, c_k$. We can rearrange this as $c_k = \frac{f^{(k)}(a)}{k!}$.

Definition 13. A smooth function $f : \Bbbk \to \Bbbk$ is $\Bbbk$-*analytic on an open set* $U \subset \Bbbk$ if for any $a \in U$, the Taylor series of f at a converges to f on U. We say f is *globally* $\Bbbk$-*analytic*, or simply $\Bbbk$-*analytic*, if it is analytic on $U = \Bbbk$.

Question 14. Let $c \in \mathbb{C}$. The function e^{cx} is $\mathbb{C}$-analytic and the functions $\sin(x)$ and $\cos(x)$ are $\mathbb{R}$-analytic.

(1) Compute the Taylor series of each at 0.

(2) Consider your answer to (1). Can you express e^{ix} in terms of $\sin(x)$ and $\cos(x)$ for real x?

If you've done Question 14 correctly and compare to Sections 1.3.1 and 1.3.2, you should now see that $e^{i\,\cdot}$ sends $\mathbb{R}$ to the unit circle in $\mathbb{C}$. If $\|a+ib\| = 1$, then $a + ib = e^{i\theta}$ for some $\theta \in \mathbb{R}$. This θ is the angle with respect to the real axis, while $a = \cos(\theta)$ and $b = \sin(\theta)$. More generally, if $\|a+ib\| = \sqrt{(a+ib)(a-ib)} = \sqrt{a^2+b^2} = R$, then $a + ib = Re^{i\theta}$, $a = R\cos(\theta)$, and $b = R\sin(\theta)$. So, we have both polar and Cartesian representations of complex numbers.

Question 15. Let $c = Re^{i\theta}$. What is the polar representation of $\overline{c}$?

Question 16. We've just described complex exponentials in terms of trigonometric functions. It's very useful to be able to express trigonometric functions in terms of complex exponentials as well.

(1) If $c = a + ib$, compute $c + \overline{c}$ and $c - \overline{c}$.

(2) Can you use your answer to 1 to express the real functions $\sin(\theta)$ and $\cos(\theta)$ in terms of complex exponentials?

(3) Verify that your answer to 2 is compatible with your answer to Question 14. That is, verify that Taylor series of your expressions in 2 match the sin and cos Taylor series you computed in Question 14.

If you've ever taught trigonometry, I'm sure you used the following identity quite a bit.

$$\sin^2(\theta) + \cos^2(\theta) = 1 \tag{1}$$

Using the definitions of $\sin(\theta)$ and $\cos(\theta)$, we can see that this is simply the Pythagorean theorem:

$$\begin{aligned}\sin^2(\theta) + \cos^2(\theta) &= \left(\frac{y}{r}\right)^2 + \left(\frac{x}{r}\right)^2 \\ &= \frac{y^2 + x^2}{r^2} \\ &= \frac{r^2}{r^2} \\ &= 1\end{aligned}$$

Alternatively, you could show this using your answer to Question 14, but it would be a bit like swatting a mosquito with a sledge hammer. Having said that, every other trigonometric identity that comes to mind follows pretty straightforwardly from Question 14 and Equation (1). Once you have the tools of Question 14, this is usually the simplest means for deriving trigonometric identities. Unfortunately, these tools generally aren't available in a typical trigonometry class – they rely on Taylor's theorem, which is usually taught in Calculus II.

Question 17. Use Question 14 to show the following identities:

(1) $\sin(\alpha + \beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta)$

(2) $\cos(\alpha + \beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)$

Note that you can derive the double angle and half-angle identities from the identities in Question 17. You could also derive them directly from your answer to Question 14 pretty efficiently.

2 Hyperbolic Trigonometry

We are now prepared to delve into the main topics of the course. We will approach hyperbolic trigonometry by analogy with usual Euclidean trigonometry, which will hopefully serve as a comfortable, familiar home base.

I will use some non-standard conventions and definitions here to make the analogy with Euclidean trigonometry more direct and, in my entirely objective opinion, make the discussion more elegant.

Essentially, I will take all distances used in hyperbolic trigonometry to be *Minkowski distances*, to be defined shortly. The standard approach in this subject is to take some distances to be Minkowski distances and others to be Euclidean distances. I will make an effort to point out where we stray from the norm to avoid potential confusion next time you find yourself in downtown Roanoke making small talk about hyperbolic trigonometric functions.

2.1 The Minkowski Distance Formula

We leave the world of Euclidean trigonometry for its hyperbolic cousin by making a single fundamental change – we work with Minkowski distances rather than Euclidean distances.

Definition 18. Let $V = V_- \oplus V_+$ be a decomposition of a real vector space as a direct sum of subspaces, which we will call the *negative subspace* and *positive subspace*, respectively. Assume both V_- and V_+ are equipped with the dot product, and let $\mathbf{p} = (\mathbf{p}_-, \mathbf{p}_+)$ and $\mathbf{q} = (\mathbf{q}_-, \mathbf{q}_+)$ be points in V. The *Minkowski distance* between $\mathbf{p}$ and $\mathbf{q}$ is given by

$$\begin{aligned}\mathbf{d}_{\mathrm{M}}(\mathbf{p}, \mathbf{q})^2 &= (\mathbf{q}_+ - \mathbf{p}_+) \cdot (\mathbf{q}_+ - \mathbf{p}_+) - (\mathbf{q}_- - \mathbf{p}_-) \cdot (\mathbf{q}_- - \mathbf{p}_-) \\ &= \mathbf{d}_{\mathrm{E}}(\mathbf{p}_+, \mathbf{q}_+)^2 - \mathbf{d}_{\mathrm{E}}(\mathbf{p}_-, \mathbf{q}_-)^2 .\end{aligned}$$

You may object that this doesn't satisfy the usual properties of a distance. Given distinct $\mathbf{p}$ and $\mathbf{q}$ in V, if $\mathbf{d}_{\mathrm{E}}(\mathbf{p}_-, \mathbf{q}_-)^2 = \mathbf{d}_{\mathrm{E}}(\mathbf{p}_+, \mathbf{q}_+)^2$, we will have $\mathbf{d}_{\mathrm{M}}(\mathbf{p}, \mathbf{q}) = 0$. Even more unsettlingly, if $\mathbf{d}_{\mathrm{E}}(\mathbf{p}_-, \mathbf{q}_-)^2 > \mathbf{d}_{\mathrm{E}}(\mathbf{p}_+, \mathbf{q}_+)^2$, then $\mathbf{d}_{\mathrm{M}}(\mathbf{p}, \mathbf{q})$ will be imaginary. And we surely can't make sense of the triangle inequality given the previous point.

You would be correct to voice these concerns – the Minkowski distance is *not* a distance in the traditional sense. It is a useful notion though. It plays a role in hyperbolic trigonometry that is completely analogous to the Euclidean distance in Euclidean trigonometry. And we will see later that it is the "correct" notion of spacetime distance in Special Relativity. While the $(n-1)$-sphere we visited in Definition 3 was the locus of points a fixed Euclidean distance from some center point, the locus of points a fixed Minkowski distance from some center point in an n-dimensional real vector space is an $(n-1)$-dimensional *hyperboloid*. In the special case of a two dimensional ambient space, we have a one dimensional hyperboloid, more commonly known as a hyperbola.

In the Euclidean setting, we could describe the Euclidean distance formula in terms of the dot product. We have a completely analogous description of the Minkowski distance formula.

Definition 19. Let $V = V_- \oplus V_+$ be a decomposition of a real vector space into a sum of negative and positive subspaces, and let $\mathbf{p} = (\mathbf{p}_-, \mathbf{p}_+)$ and $\mathbf{q} = (\mathbf{q}_-, \mathbf{q}_+)$ be arbitrary points in V. The *Minkowski pairing on* V is the bilinear map

$$\begin{aligned}\eta : V \times V &\to \mathbb{R} \\ (\mathbf{p}, \mathbf{q}) &\mapsto \mathbf{p}_+ \cdot \mathbf{q}_+ - \mathbf{p}_- \cdot \mathbf{q}_- .\end{aligned}$$

The *Minkowski distance* between $\mathbf{p}$ and $\mathbf{q}$ is

$$\mathbf{d}_{\mathrm{M}}(\mathbf{p}, \mathbf{q}) = \sqrt{\eta(\mathbf{p} - \mathbf{q}, \mathbf{p} - \mathbf{q})}.$$

2.2 Hyperbolic Trigonometric Functions

Let $V = V_- \oplus V_+$ be a decomposition of a real vector space into a sum of negative and positive subspaces, and let $\mathbf{p} = (\mathbf{p}_-, \mathbf{p}_+)$ be a point in V. To discuss such a hyperboloid, we will call

$$\mathbf{H}(\rho, \mathbf{p}) := \left\{\mathbf{q} \in V \;\middle|\; \mathbf{d}_{\mathrm{M}}(\mathbf{p}, \mathbf{q})^2 = \rho^2\right\} \tag{2}$$

the *hyperboloid in* V *of Minkowski radius* ρ *centered at* $\mathbf{p}$. Since Minkowski distance may be imaginary, so may the Minkowski radius.

Question 20. Let $V = V_- \oplus V_+$ be a decomposition of a two dimensional real vector space as a sum of one dimensional negative and positive subspaces. Can you sketch the hyperbolas $\mathbf{H}(1, \mathbf{0})$ and $\mathbf{H}(i, \mathbf{0})$? Treat V_- as the x-axis and V_+ as the y-axis.

To define the hyperbolic versions of the trigonometric functions sin, cos, and tan, we start with a hyperbola $\mathbf{H}(ir, \mathbf{0})$, where r is some real number. Now consider an angle subtending an arc of the hyperbola, as shown in Figure 5.

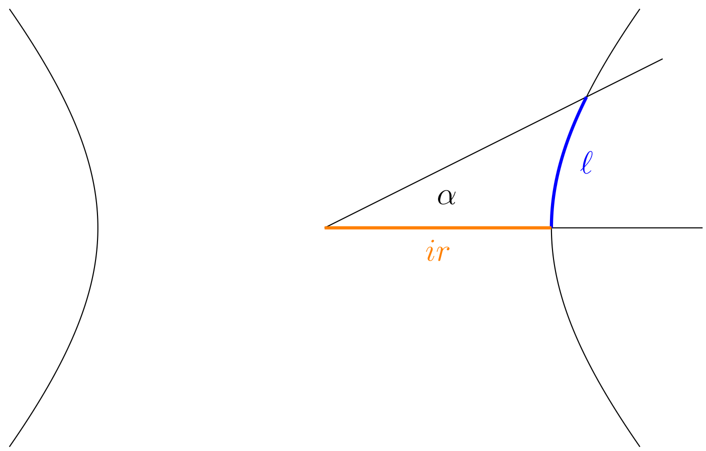

Figure 5: The angle α subtending an arc in $\mathbf{H}(ir, \mathbf{0})$.

The Minkowski angle α is the ratio $\alpha = \frac{\ell}{ir}$ where ℓ is the Minkowski arc length of the subtended hyperbolic arc. Next, consider a right triangle with hypotenuse a segment from the origin to some point $\mathbf{p}$ of $\mathbf{H}(ir, \mathbf{0})$. In the figure below, all lengths are Minkowski lengths and α is a Minkowski angle.

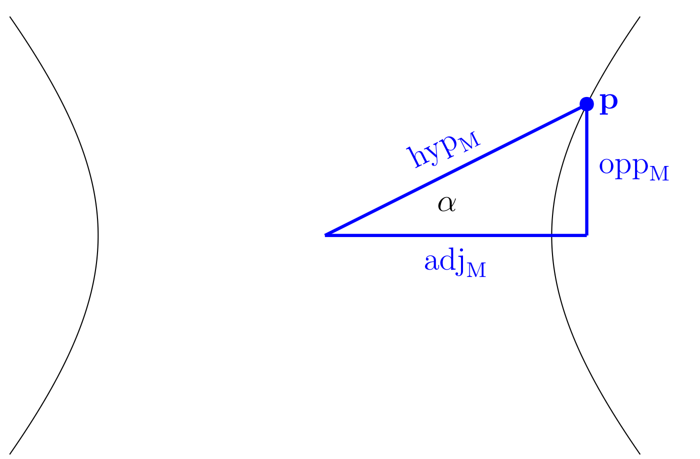

Figure 6: The right triangle used to define hyperbolic trigonometric functions. Note that $\mathrm{adj_M}$ and $\mathrm{hyp_M}$ are imaginary while $\mathrm{opp_M}$ is real.

We define the *Minkowski sine*, *Minkowski cosine*, and *Minkowski tangent* of α to be the ratios

$$\operatorname{msin}(\alpha) = \frac{\mathrm{opp_M}}{\mathrm{hyp_M}}, \qquad \operatorname{mcos}(\alpha) = \frac{\mathrm{adj_M}}{\mathrm{hyp_M}}, \qquad \text{and} \qquad \operatorname{mtan}(\alpha) = \frac{\mathrm{opp_M}}{\mathrm{adj_M}}.$$

Brief Aside. I am making some non-standard definitions here. First, rather than the imaginary Minkowski angle $\alpha = \frac{\ell}{ir}$, which is the ratio of the Minkowski arc length of the subtended hyperbolic arc and its Minkowski radius, the standard convention is to use the real *hyperbolic angle* $a = \frac{\ell}{r}$, which is the ratio of the Minkowski arc length of the subtended hyperbolic arc and the Euclidean distance between the hyperbola and its hyperbolic center.[4] Next, the usual hyperbolic trigonometric functions are called *hyperbolic sine*, *hyperbolic cosine*, and *hyperbolic tangent*. These functions take the hyperbolic angle a as their argument, and are denoted $\sinh(a)$, $\cosh(a)$, and $\tanh(a)$. They are defined, *in terms of Euclidean distances*, to be

$$\sinh(a) = \frac{\mathrm{opp_E}}{r}, \qquad \cosh(a) = \frac{\mathrm{adj_E}}{r}, \qquad \text{and} \qquad \tanh(a) = \frac{\mathrm{opp_E}}{\mathrm{adj_E}}.$$

Here is a dictionary between our conventions and the standard conventions:

$$\alpha = -ia, \qquad \mathrm{opp_M} = \mathrm{opp_E}, \qquad \mathrm{adj_M} = i\,\mathrm{adj_E},$$

$$\operatorname{msin}(\alpha) = -i\sinh(a), \qquad \operatorname{mcos}(\alpha) = \cosh(a), \qquad \text{and} \qquad \operatorname{mtan}(\alpha) = -i\tanh(a).$$

Both versions have their merits. I find the Minkowski version I've described more conceptually elegant. It makes the analogy with Euclidean trigonometry more direct – we are simply exchanging the Euclidean distance formula with the Minkowski distance formula. On the other hand, the relation between the standard conventions and real Cartesian coordinates is a bit more direct as the components of real Cartesian coordinates are *Euclidean* distances from coordinate hyperplanes, rather than Minkowski distances from these hyperplanes.

Question 21. In Euclidean trigonometry, we have the Pythagorean identity

$$\sin^2(\theta) + \cos^2(\theta) = 1.$$

Is there an analogous identity for $\operatorname{msin}(\alpha)$ and $\operatorname{mcos}(\alpha)$? What about $\sinh(a)$ and $\cosh(a)$?

At this point we have conceptual, geometric descriptions of Minkowski angles and trigonometric functions. It will be useful to have more explicit, calculational descriptions as well. As we have treated V_- as our x-axis and V_+ as our y-axis, let's call the coordinates on these

[4] Note that in this interpretation, only the horizontal line segment from Figure 5 would have length r. On the other hand, in the Minkowski distance interpretation, both this line segment and the diagonal line segment meeting the top of the hyperbolic arc ($\mathrm{hyp_M}$ in Figure 6) have Minkowski length ir. This also means that the denominator in the usual hyperbolic trigonometric functions is *not* the hypotenuse of a right triangle. If we wish to keep the hypotenuse interpretation, we can note that $r = -i(ir)$, so r is $-i$ times the Minkowski radius.

subspaces x and y respectively. We'll start by describing the Minkowski arc length ℓ as a function of y. Let's parametrize the arc by

$$y(t) = t \qquad \text{and} \qquad x(t) = \sqrt{t^2 + r^2} \qquad \text{for } t \geq 0.$$

Then

$$y'(t) = 1 \qquad \text{and} \qquad x'(t) = \frac{t}{\sqrt{t^2 + r^2}} \qquad \text{for } t > 0.$$

Recall that arc length is computed as an integral of hypotenuse lengths of infinitesimal triangles along the given path.

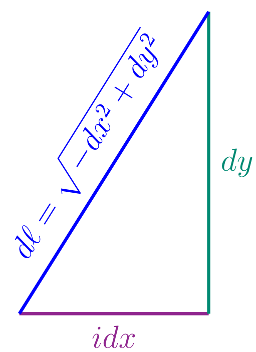

Figure 7: The Minkowski arc length ℓ is the integral of Minkowski lengths of hypotenuses of infinitesimal triangles of this form.

Then we have

$$\begin{aligned} \ell(y) &= \int_0^y \sqrt{1 - \frac{t^2}{t^2 + r^2}} dt \\ &= \int_0^y \sqrt{\frac{r^2}{t^2 + r^2}} dt. \end{aligned}$$

We can evaluate this integral using trig substitution. Set $t = r\tan(\theta)$. Then $dt = r\sec^2(\theta)d\theta$, and our integral reduces to an integral of $\sec(\theta)$:

$$\begin{aligned}
\ell(y) &= \int_0^{\tan^{-1}\left(\frac{y}{r}\right)} \sqrt{\frac{r^2}{r^2\tan^2(\theta)+r^2}}\, r\sec^2(\theta)d\theta \\
&= \int_0^{\tan^{-1}\left(\frac{y}{r}\right)} \sqrt{\frac{1}{\sec^2(\theta)}}\, r\sec^2(\theta)d\theta \\
&= r\int_0^{\tan^{-1}\left(\frac{y}{r}\right)} \sec(\theta)d\theta \\
&= r\ln\left(\sec\left(\tan^{-1}\left(\frac{y}{r}\right)\right)+\tan\left(\tan^{-1}\left(\frac{y}{r}\right)\right)\right) - r\ln\left(\sec(0)+\tan(0)\right) \\
&= r\ln\left(\frac{\sqrt{y^2+r^2}}{r}+\frac{y}{r}\right) - r\ln(1) \\
&= r\ln\left(\frac{\sqrt{y^2+r^2}}{r}+\frac{y}{r}\right)
\end{aligned} \tag{3}$$

Note that we could also write Equation (3) as

$$\ell(x,y) = r\ln\left(\frac{x}{r}+\frac{y}{r}\right). \tag{4}$$

Next, let's go in the other direction and express x and y in terms of ℓ. We'll want this in order to obtain explicit formulas for our hyperbolic trig functions. Equation (4) implies

$$e^{\frac{\ell}{r}} = \frac{x}{r}+\frac{y}{r}, \tag{5}$$

so,

$$\begin{aligned}
e^{\frac{2\ell}{r}} &= \left(\frac{x}{r}+\frac{y}{r}\right)^2 \\
&= \left(\frac{x}{r}\right)^2 + \left(\frac{y}{r}\right)^2 + 2\left(\frac{x}{r}\right)\left(\frac{y}{r}\right).
\end{aligned} \tag{6}$$

We will rewrite Equation (6) two ways.

$$\begin{aligned}
e^{\frac{2\ell}{r}} &= \left(\frac{x}{r}\right)^2 + \left(\frac{x}{r}\right)^2 - 1 + 2\left(\frac{x}{r}\right)\left(\frac{y}{r}\right) \\
&= 2\left(\frac{x}{r}\right)^2 + 2\left(\frac{x}{r}\right)\left(\frac{y}{r}\right) - 1 \\
&= 2\left(\frac{x}{r}\right)\left(\frac{x}{r}+\frac{y}{r}\right) - 1
\end{aligned} \tag{7}$$

$$\begin{aligned}
e^{\frac{2\ell}{r}} &= \left(\frac{y}{r}\right)^2 + 1 + \left(\frac{y}{r}\right)^2 + 2\left(\frac{x}{r}\right)\left(\frac{y}{r}\right) \\
&= 2\left(\frac{y}{r}\right)^2 + 2\left(\frac{x}{r}\right)\left(\frac{y}{r}\right) + 1 \\
&= 2\left(\frac{y}{r}\right)\left(\frac{x}{r}+\frac{y}{r}\right) + 1
\end{aligned} \tag{8}$$

Using Equation (5), Equation (7) becomes

$$e^{\frac{2\ell}{r}} = 2\left(\frac{x}{r}\right)e^{\frac{\ell}{r}} - 1 \tag{9}$$

and Equation (8) becomes

$$e^{\frac{2\ell}{r}} = 2\left(\frac{y}{r}\right)e^{\frac{\ell}{r}} + 1. \tag{10}$$

We can solve Equations (9) and (10) for $\frac{x}{r}$ and $\frac{y}{r}$ respectively to obtain

$$\begin{aligned} \frac{x}{r} &= \frac{e^{\frac{2\ell}{r}} + 1}{2e^{\frac{\ell}{r}}} \\ &= \frac{e^{\frac{\ell}{r}} + e^{-\frac{\ell}{r}}}{2} \end{aligned} \tag{11}$$

and

$$\begin{aligned} \frac{y}{r} &= \frac{e^{\frac{2\ell}{r}} - 1}{2e^{\frac{\ell}{r}}} \\ &= \frac{e^{\frac{\ell}{r}} - e^{-\frac{\ell}{r}}}{2}. \end{aligned} \tag{12}$$

So, we can now give explicit expressions for hyperbolic trig functions. Note that $\frac{\ell}{r}$ is the hyperbolic angle a, $\frac{x}{r}$ is $\cosh(a)$, and $\frac{y}{r}$ is $\sinh(a)$. Then Equations (11) and (12) tell us

$$\cosh(a) = \frac{e^a + e^{-a}}{2} \qquad \text{and} \qquad \sinh(a) = \frac{e^a - e^{-a}}{2},$$

or, in the language of our Minkowski trig functions,

$$\operatorname{mcos}(\alpha) = \frac{e^{i\alpha} + e^{-i\alpha}}{2} \qquad \text{and} \qquad \operatorname{msin}(\alpha) = \frac{e^{i\alpha} - e^{-i\alpha}}{2i}.$$

We can compare these expressions to the complex exponential expressions you found for $\cos(\theta)$ and $\sin(\theta)$ in Question 16, hopefully the following:

$$\cos(\theta) = \frac{e^{i\theta} + e^{-i\theta}}{2} \qquad \text{and} \qquad \sin(\theta) = \frac{e^{i\theta} - e^{-i\theta}}{2i}.$$

So, we gave different (but entirely analogous) definitions for the Euclidean and Minkowski trigonometric functions, but in the end we wound up with exactly the same expressions for both. Does this mean they are actually the same functions? Well, not yet. Note that the Euclidean trigonometric functions take real arguments while the Minkowski trigonometric functions take imaginary arguments – their domains do not match. However, suppose we extend the Euclidean trigonometric functions from $\mathbb{R}$ to $\mathbb{C}$. We can do this by *defining* the trigonometric functions in terms of either their complex exponential expressions or Taylor series, and interpreting the result as a map from $\mathbb{C}$ to $\mathbb{C}$ rather than $\mathbb{R}$ to $\mathbb{R}$. If we do this, then the real Euclidean trigonometric functions are the restrictions of the complex trigonometric functions to the real axis in $\mathbb{C}$, while the Minkowski trigonometric functions are the restrictions of the complex trigonometric functions to the imaginary axis in $\mathbb{C}$.

For reference, I've plotted some Minkowski angles below.

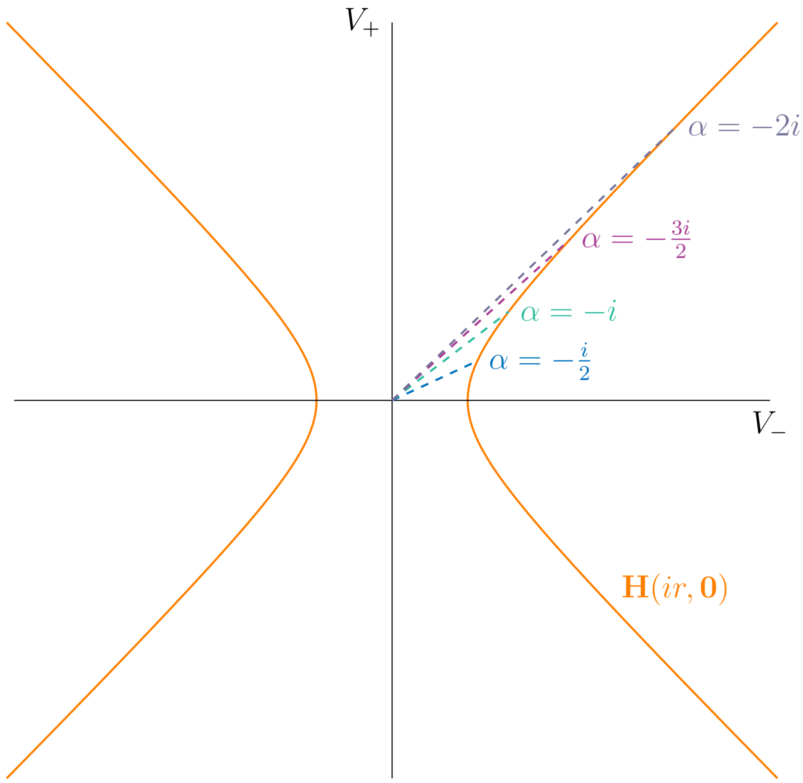

Figure 8: Plot of a few Minkowski angles.

Question 22. Compute the Taylor series at 0 of $\mathrm{msin}(\alpha)$, $\mathrm{mcos}(\alpha)$, $\sinh(a)$, and $\cosh(a)$.

Question 23. Compute double angle identities for $\mathrm{msin}(\alpha)$, $\mathrm{mcos}(\alpha)$, $\sinh(a)$, and $\cosh(a)$.

Question 24. In the Euclidean setting, a radius of a circle (such as the hypotenuse in Figure 4) meets the circle at a right angle. On the other hand, a Minkowski radius of a hyperbola (such as the hypotenuse in Figure 6) doesn't look like it meets the hyperbola at a right angle.

(1) What does it mean for a pair of lines or curves to meet at a right angle?

(2) What *should* be the analogue of meeting at a right angle in the hyperbolic setting?

(3) Does the hypotenuse in Figure 6 meet the hyperbola in the way you described in 2?

Two comments before proceeding. First, in the description preceding the Brief Aside, you may object that we have placed the starting point for our hyperbolic arc on the x-axis, while it is not necessary to do so in Euclidean trigonometry. In fact, in Euclidean trigonometry we could always rotate our original coordinate system to a new coordinate system in which our circular arc starts on the x-axis. The situation is actually similar here, but rotations are no longer the right transformation to consider – while they preserve Euclidean distances, rotations in dimension two do *not* preserve Minkowski distances.[5] The analogous transformation in the hyperbolic setting is called a *Lorentz boost*.[6] In fact, the analogue of *meeting at a right angle* you came up with in Question 24 will be relevant here. If we perform a Lorentz boost on our standard coordinate system, the resulting axes will be orthogonal in this sense rather than in the usual Euclidean sense. We will explore these transformations in Section 2.3.

Next, you may have noticed that the figures in both Section 1.3.2 and the current section are all two dimensional. Both Euclidean trigonometry and hyperbolic trigonometry are really two dimensional theories – the former involving Euclidean distances and circles, and the latter involving Minkowski distances and hyperbolas. These theories are very useful even though our world is more than two dimensional. The reason is that we can reduce many higher dimensional problems to two dimensional problems (or sequences of two dimensional problems), just like you did in Question 2. The two dimensional nature of these theories is reflected in the relevant transformations as well – no matter the dimension of the ambient space V, both rotations and Lorentz boosts fix a subspace of codimension 2.[7] For example, a rotation in dimension two fixes the point about which the rotation occurs (0-dimensional) and in three dimensions it fixes the axis of rotation (1-dimensional).

2.3 The Poincaré Group

Notation 25. In this section we will consider different bases for $V = V_- \oplus V_+$ that are related by Lorentz boosts. As such, it will be handy to introduce some notation for bases. If $\mathbf{v}$ is any vector in V with $\eta(\mathbf{v}, \mathbf{v}) \neq 0$, we will write

$$\hat{\mathbf{v}} := \frac{\mathbf{v}}{\|\sqrt{\eta(\mathbf{v}, \mathbf{v})}\|}.$$

Note that this may be subtly different from the unit vectors you are familiar with from Euclidean vector calculus – we may have $\eta(\hat{\mathbf{v}}, \hat{\mathbf{v}}) = -1$ rather than $+1$.

In the Euclidean setting, the distance-preserving linear transformations on V formed the orthogonal group $O(V)$, while the not-necessarily-linear distance-preserving transformations formed the Euclidean group $E(V)$. We have an analogous situation in the hyperbolic setting.

Definition 26. Let $V = V_- \oplus V_+$ be a decomposition of a real vector space as a direct sum of a negative and a positive subspace. The group of linear transformations $f : V \to V$ that preserve Minkowski distance is called the *Lorentz group*. The group of not-necessarily-linear transformations $f : V \to V$ that preserve Minkowski distance is called the *Poincaré group*. That is, such an f is in the Lorentz group or Poincaré group if for all $\mathbf{p}, \mathbf{q} \in V$,

$$\mathbf{d}_{\mathrm{M}}\left(f(\mathbf{p}), f(\mathbf{q})\right) = \mathbf{d}_{\mathrm{M}}\left(\mathbf{p}, \mathbf{q}\right).$$

[5] More generally, if $V = V_- \oplus V_+$ is of arbitrary dimension, a rotation will only preserve Minkowski distances in V if it preserves V_- and V_+, rather than mixing them.

[6] Or, less commonly but more suggestively, a *hyperbolic rotation*.

[7] Meaning, they act as the identity on a subspace of V whose dimension is $\dim(V) - 2$.

The Lorentz transformations on $V = V_- \oplus V_+$ can all be expressed as compositions of orthogonal transformations on V_- and V_+ and *Lorentz boosts*, to be described shortly. In the Poincaré group, we may also compose with translations in V.

Brief Aside. Generally, the terms *Lorentz group* and *Poincaré* group are reserved for the case in which V is four dimensional, with either V_- or V_+ three dimensional and the other one dimensional. We will see why when we get to the Special Relativity portion of the course. I am applying these terms to a slightly more general setting.

Just like a rotation $R_{\theta'}$ by θ' around the origin in two dimensions fix circles centered at the origin and simply adds θ' to the Euclidean angle associated to each point in the circle, a Lorentz boost $L_{\alpha'}$ by α' fixes hyperbolas centered at the origin and adds α' to the Minkowski angle associated to each point of the hyperbola. In more detail, let's adopt a hyperbolic version of polar coordinates. For a complete treatment, we would split V into four wedges and a pair of lines, but for concreteness let's restrict to the region shown below:

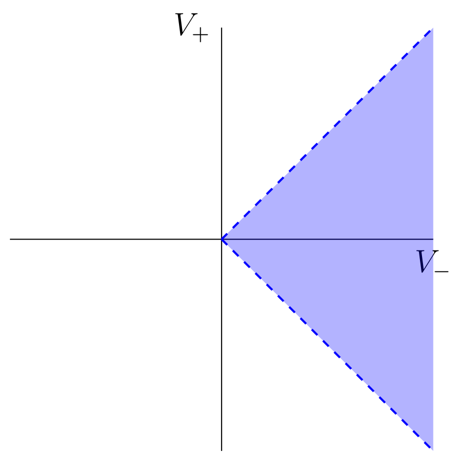

Figure 9: We will restrict to our attention to this region.

We can represent any point in the region indicated in Figure 9 in terms of a Minkowski radius ir and Minkowski angle α. So, just like we may write $\mathbf{p} = (R, \theta)$ in polar coordinates for a point in a two dimensional Euclidean space, we will write $\mathbf{p} = (ir, \alpha)$ in our hyperbolic version of polar coordinates for a point in the indicated region of Figure 9. The Lorentz boost $L_{\alpha'} : V \to V$ leaves the Minkowski radius ir unchanged but adds α' to the Minkowski angle α:

$$L_{\alpha'}\left((ir, \alpha)\right) = (ir, \alpha + \alpha').$$

In pictures, we have the following situation:

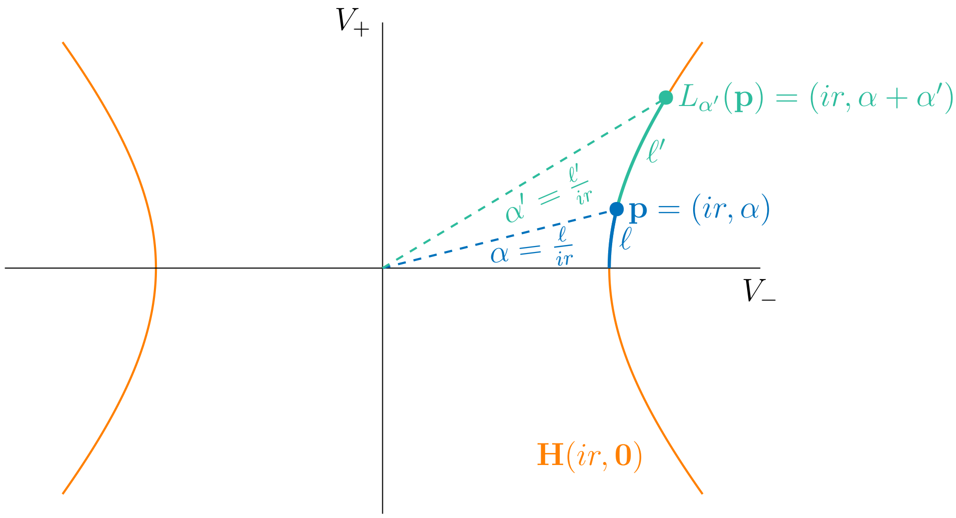

Figure 10: The Lorentz boost $L_{\alpha'}$ leaves the Minkowski radius ir unchanged, but adds α' to the Minkowski angle.

So, that was the most natural way we could possibly define a Minkowski analogue of a Euclidean rotation. But we've also said that the Lorentz boost should be a linear map. Is the map we've just defined linear (after extending from the specified region to all of V)? To answer this, let's see what form Lorentz boosts take in Cartesian coordinates. Using Figure 6, if $\mathbf{p} = (ir, \alpha)$ in hyperbolic polar coordinates, then in Cartesian coordinates we have

$$\mathbf{p} = (r \operatorname{mcos}(\alpha), ir \operatorname{msin}(\alpha)) .$$

So, in Cartesian coordinates the Lorentz boost $L_{\alpha'}$ is given by

$$L_{\alpha'} : (r \operatorname{mcos}(\alpha), ir \operatorname{msin}(\alpha)) \mapsto (r \operatorname{mcos}(\alpha + \alpha'), ir \operatorname{msin}(\alpha + \alpha')) .$$

We saw in Section 2.2 that the Minkowski trigonometric functions and Euclidean trigonometric functions have the same exponential form, and in fact are different restrictions of the same complex functions. So, they satisfy all of the same identities. In particular, we have

$$\operatorname{mcos}(\alpha + \alpha') = \operatorname{mcos}(\alpha) \operatorname{mcos}(\alpha') - \operatorname{msin}(\alpha) \operatorname{msin}(\alpha')$$

and

$$\operatorname{msin}(\alpha + \alpha') = \operatorname{msin}(\alpha) \operatorname{mcos}(\alpha') + \operatorname{mcos}(\alpha) \operatorname{msin}(\alpha').$$

Then

$$\begin{aligned} L_{\alpha'} : r \operatorname{mcos}(\alpha)\hat{\mathbf{x}} + ir \operatorname{msin}(\alpha)\hat{\mathbf{y}} \mapsto & r \left(\operatorname{mcos}(\alpha) \operatorname{mcos}(\alpha') - \operatorname{msin}(\alpha) \operatorname{msin}(\alpha')\right) \hat{\mathbf{x}} \\ & + ir \left(\operatorname{msin}(\alpha) \operatorname{mcos}(\alpha') + \operatorname{mcos}(\alpha) \operatorname{msin}(\alpha')\right) \hat{\mathbf{y}} \\ = & \left(r \operatorname{mcos}(\alpha) \operatorname{mcos}(\alpha') + ir \operatorname{msin}(\alpha)(i \operatorname{msin}(\alpha'))\right) \hat{\mathbf{x}} \\ & + \left(ir \operatorname{msin}(\alpha) \operatorname{mcos}(\alpha') + r \operatorname{mcos}(\alpha)(i \operatorname{msin}(\alpha'))\right) \hat{\mathbf{y}}. \end{aligned} \tag{13}$$

In other words,

$$L_{\alpha'} : x\hat{\mathbf{x}} + y\hat{\mathbf{y}} \mapsto (x \operatorname{mcos}(\alpha') + iy \operatorname{msin}(\alpha'))\,\hat{\mathbf{x}} + (ix \operatorname{msin}(\alpha') + y \operatorname{mcos}(\alpha'))\,\hat{\mathbf{y}}.$$

We can see that $L_{\alpha'}$ is indeed a linear map, at least on the specified region. We will extend $L_{\alpha'}$ linearly to obtain a linear map on all of V. In matrix form , we would write $L_{\alpha'}$ as

$$L_{\alpha'} = \begin{pmatrix} \operatorname{mcos}(\alpha') & i\operatorname{msin}(\alpha') \\ i\operatorname{msin}(\alpha') & \operatorname{mcos}(\alpha') \end{pmatrix}$$

so Equation (13) becomes

$$\begin{pmatrix} \operatorname{mcos}(\alpha') & i\operatorname{msin}(\alpha') \\ i\operatorname{msin}(\alpha') & \operatorname{mcos}(\alpha') \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x\operatorname{mcos}(\alpha') + iy\operatorname{msin}(\alpha') \\ ix\operatorname{msin}(\alpha') + y\operatorname{mcos}(\alpha') \end{pmatrix}. \tag{14}$$

Question 27. We have claimed that the set of Lorentz transformations forms a group under composition. So, each Lorentz transformation should have an inverse. What is ${L_{\alpha'}}^{-1}$?

Question 28. (1) What form would Equation (14) take if we were to use the hyperbolic angle a' and hyperbolic trigonometric functions $\sinh(a')$ and $\cosh(a')$ rather than the Minkowski versions α', $\operatorname{msin}(\alpha')$, and $\operatorname{mcos}(\alpha')$?

(2) Use the exponential expressions for $\sinh(a)$ and $\cosh(a)$ to derive hyperbolic angle addition formulas for these functions. That is, find expressions for $\sinh(a + a')$ and $\cosh(a + a')$. Is your result compatible with the equation you found in 1?

Question 29. Consider a coordinate system in which the coordinates x' and y' are related to our original coordinates x and y by the Lorentz boost $L_{\alpha'}$:

$$\hat{\mathbf{x}}' = L_{\alpha'}(\hat{\mathbf{x}}) \qquad \text{and} \qquad \hat{\mathbf{y}}' = L_{\alpha'}(\hat{\mathbf{y}}).$$

(1) Can you sketch the x'- and y'-axes in the original coordinate system?

(2) Are the new axes orthogonal to each other?

Hint: Think about your conclusions from Question 24.

(3) Let $\mathbf{p} = x\hat{\mathbf{x}} + y\hat{\mathbf{y}}$ in the original coordinate system and $\mathbf{p} = x'\hat{\mathbf{x}}' + y'\hat{\mathbf{y}}'$ in the new coordinate system. Express x' and y' in terms of x and y.

3 Special Relativity

3.1 Maxwell's Equations and Electromagnetic Waves

We'll begin our story of the development of the Theory of Special Relativity with James Clerk Maxwell. In his 1864 paper *A Dynamical Theory of the Electromagnetic Field*, Maxwell collected equations of electromagnetism (with some of his own fixes) and recognized that taken together, they supported a new concept. Waves, as understood at the time, were a phenomenon that required an ambient medium having an equilibrium state and a restorative force. If the system were ever out of equilibrium, the restorative force would push it in the equilibrium direction. Without dampening, it could overshoot equilibrium, leading to oscillations. Mathematically, in an idealized one dimensional system, the phenomenon is described by the *wave equation*

$$\frac{\partial^2 f(x,t)}{\partial t^2} = v^2 \frac{\partial^2 f(x,t)}{\partial x^2}. \tag{15}$$

Solutions have the form

$$f(x,t) = g(x+vt) + h(x-vt), \tag{16}$$

where g and h are functions of a single variable.

Question 30. (1) Can you verify that if g and h are smooth functions of a single variable, then Equation (16) is a solution to Equation (15)?

(2) If x represents position and t represents time, what do you think v represents?

Maxwell recognized that a change in the electric field at a location would induce a change in the magnetic field, and a change in the magnetic field would induce a change in the electric field. Let's look at the simplest setting available – we'll work in a vacuum, with no current or charges. Then in modern notation, this observation is usually expressed as follows:

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{17}$$

$$\nabla \times \mathbf{B} = \mu_0 \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \tag{18}$$

Here, $\mathbf{E}$ is the electric field, $\mathbf{B}$ is the magnetic field, and μ_0 and ϵ_0 are physical constants known as the *vacuum permeability* and *vacuum permittivity* respectively.[8] Meanwhile, the absence of electric and magnetic charges[9] is encoded by the equations

$$\nabla \cdot \mathbf{E} = 0 \tag{19}$$

and

$$\nabla \cdot \mathbf{B} = 0. \tag{20}$$

[8] If a material is subjected to a magnetic field, permeability describes the strength of the resulting magnetic field induced in the material. If it is subjected to an electric field, permittivity describes how polarized the material will become.

[9] For the record, magnetic charges (usually referred to as *magnetic monopoles*) aren't known to exist. Their theoretical merits are debated by physicists, but they've never been confirmed experimentally.

Together, Equations (17), (18), (19), and (20) imply[10]

$$\nabla^2 \mathbf{E} = \mu_0 \epsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} \tag{21}$$

and

$$\nabla^2 \mathbf{B} = \mu_0 \epsilon_0 \frac{\partial^2 \mathbf{B}}{\partial t^2}. \tag{22}$$

These are three dimensional versions of Equation 15, with

$$v^2 = \frac{1}{\mu_0 \epsilon_0}.$$

So, Maxwell concluded that there were *electromagnetic waves*, and he knew how the speed of these waves depended upon measurable physical quantities – equivalent to what we now describe as permeability and permittivity. At the time, these physical quantities had been experimentally measured for air, and based upon those experiments, he concluded that electromagnetic waves should travel at $310,740,000 \frac{\mathrm{m}}{\mathrm{s}}$ in air. He compared this to two different measured values of the speed of light in air and one of the speed of light in "the space surrounding the earth", and found good agreement (less than a 5% difference in each case). Along with some other properties Maxwell showed that light shared with electromagnetic waves, this agreement convinced Maxwell that light was an electromagnetic wave. I've included the relevant page from his paper below.

The results in this paper would require a revolutionary shift in fundamental physics that wouldn't be fully appreciated for another 40 years.

[10] By taking the curl $(\nabla \times \cdot)$ of Equations (17) and (18), then using the identity $\nabla \times (\nabla \times \cdot) = \nabla (\nabla \cdot \cdot) - \nabla^2 \cdot$ and Equations 19 and (20).

This wave consists entirely of magnetic disturbances, the direction of magnetization being in the plane of the wave. No magnetic disturbance whose direction of magnetization is not in the plane of the wave can be propagated as a plane wave at all.

Hence magnetic disturbances propagated through the electromagnetic field agree with light in this, that the disturbance at any point is transverse to the direction of propagation, and such waves may have all the properties of polarized light.

(96) The only medium in which experiments have been made to determine the value of k is air, in which $\mu=1$, and therefore, by equation (46),

$$\mathrm{V}=v. \qquad (72)$$

By the electromagnetic experiments of MM. WEBER and KOHLRAUSCH *,

$$v=310{,}740{,}000 \text{ metres per second}$$

is the number of electrostatic units in one electromagnetic unit of electricity, and this, according to our result, should be equal to the velocity of light in air or vacuum.

The velocity of light in air, by M. FIZEAU'S † experiments, is

$$\mathrm{V}=314{,}858{,}000\,;$$

according to the more accurate experiments of M. FOUCAULT ‡,

$$\mathrm{V}=298{,}000{,}000.$$

The velocity of light in the space surrounding the earth, deduced from the coefficient of aberration and the received value of the radius of the earth's orbit, is

$$\mathrm{V}=308{,}000{,}000.$$

(97) Hence the velocity of light deduced from experiment agrees sufficiently well with the value of v deduced from the only set of experiments we as yet possess. The value of v was determined by measuring the electromotive force with which a condenser of known capacity was charged, and then discharging the condenser through a galvanometer, so as to measure the quantity of electricity in it in electromagnetic measure. The only use made of light in the experiment was to see the instruments. The value of V found by M. FOUCAULT was obtained by determining the angle through which a revolving mirror turned, while the light reflected from it went and returned along a measured course. No use whatever was made of electricity or magnetism.

The agreement of the results seems to show that light and magnetism are affections of the same substance, and that light is an electromagnetic disturbance propagated through the field according to electromagnetic laws.

(98) Let us now go back upon the equations in (94), in which the quantities J and Ψ occur, to see whether any other kind of disturbance can be propagated through the medium depending on these quantities which disappeared from the final equations.

* Leipzig Transactions, vol. v. (1857), p. 260, or POGGENDORFF'S 'Annalen,' Aug. 1856, p. 10.

† Comptes Rendus, vol. xxix. (1849), p. 90. ‡ Ibid. vol. lv. (1862), pp. 501, 792.

3.2 Lorentz Invariance of Maxwell's Equations

The next major development we'll discuss is due to Hendrik Lorentz and Henri Poincaré – Lorentz got the ball rolling and Poincaré made some corrections and gave the elegant formulation still used today. They observed that Maxwell's equations are invariant under Lorentz transformations (name due to Poincaré) rather than the Galilean transformations of Newtonian mechanics. While their rotations and reflections are readily recognizable, it may not be immediately clear that Lorentz boosts as described by Lorentz and Maxwell align with those we described in Definition 26. The usual description of Lorentz boosts in physics is as follows. Let's take our original coordinate system S to have spacetime coordinates (x, y, z, ct), where c is the speed of light in vacuum. It may look odd that we have a factor of c in front of t. This allows us to use the same units for all coordinates – *yes, we will measure time in meters* – and it tends to make our analysis a bit cleaner. We'll denote the coordinates in the new system S' by (x', y', z', ct'), and take S' to be moving with speed v in the positive x direction with respect to S. Then

$$x' = \gamma\,(x - \beta ct)\,, \qquad y' = y, \qquad z' = z, \qquad \text{and} \qquad ct' = \gamma\,(ct - \beta x) \tag{23}$$

where $\beta = \frac{v}{c}$ and the *Lorentz factor* γ is $\gamma = \frac{1}{\sqrt{1-\beta^2}}$.

To understand this Lorentz transformation in terms of Definition 26, we can take the V_- to be a three dimensional subspace of V and V_+ to be one dimensional. The three spatial coordinates x, y, and z are coordinates on V_-, while the coordinate on V_+ is the time coordinate ct. Comparing Equation (23) to the answer you (hopefully) found for Question 29, part 3, you'll notice that we can relate the two by setting $\gamma = \mathrm{mcos}(\alpha')$ and $\gamma\beta = i\,\mathrm{msin}(\alpha')$, or, said differently, $\beta = i\,\mathrm{mtan}(\alpha')$.

Question 31. Can you verify that the above description of γ and β in terms of α' is compatible with the definition $\gamma = \frac{1}{\sqrt{1-\beta^2}}$?

Note that Lorentz and Poincaré's observation has some surprising consequences. For instance, two observers moving relative to each other will measure the same speed of light. At the time, the accepted theory was that there was a universal preferred reference frame, the rest frame of the hypothesized "ether". While Maxwell's analysis showed that a pair of fields could interact in a way that leads to wave motion, it was hard to throw off the idea that waves required some ambient medium. The ether was a hypothesized medium to support electromagnetic waves, and it was thought to provide a distinguished, universal reference frame. And to a large extent, this reticence to propose a theory of light with no ether made sense. Note that the propagation speed of electromagnetic waves depends on two medium-dependent properties – permeability and permittivity. It takes a bit of a philosophical leap to attribute these properties to empty space (rather than some as-yet-undetected ether medium) and recognize that velocity relative to empty space isn't meaningful. Poincaré and Einstein made the next major breakthrough in recognizing that only relative velocities were meaningful and we could throw out the idea of the ether – which various experiments had failed to detect.

In the interest of time – there are only a few classes left and more material I hope to cover – we will skip the proof of the key result that Maxwell's equations are Lorentz invariant. I'll just say that the modern approach to this involves rewriting Maxwell's equations in a clever, elegant way, typically in the language of either tensor calculus or differential forms. Lorentz invariance is easier to see in the rewritten equations. Taking this approach would require a detour to build up the mathematical background.

3.3 Spacetime Diagrams

When analyzing a question in Special Relativity, it can be very useful to draw *spacetime diagrams*. As before, we'll take the new coordinate system S' to be traveling with speed v in the positive x direction relative to the original coordinate system S. Since all motion is in the x direction, we will draw a two dimensional picture, where we ignore the y and z directions. If we match the origins of the two coordinate systems (so we are dealing with just a Lorentz boost, with no translation in spacetime), then we can draw the pair of coordinate systems as shown in Figure 11.

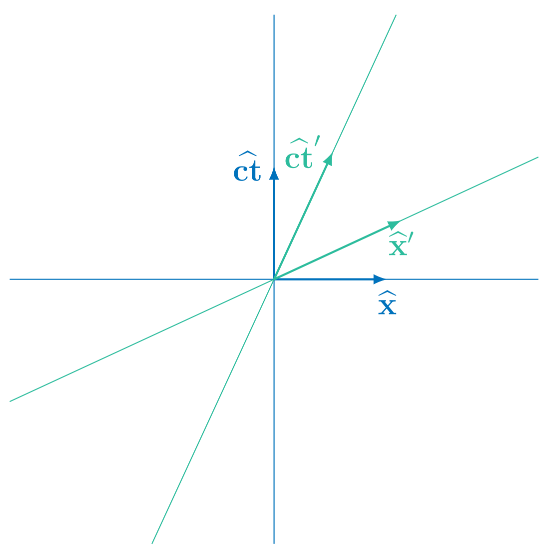

Figure 11: Spacetime diagram showing S' in the coordinate system S.

Now let's interpret Figure 11 a bit.

Question 32. (1) Can you describe the slopes of the x'- and ct'-axes of Figure 11 in terms of $\beta = \frac{v}{c}$?

(2) Can you sketch the possible paths of a photon of light?

Question 32 suggests the following observations.

- Any photon of light will be represented as a line with a slope of ± 1 on a spacetime diagram. If $\mathbf{p}$ and $\mathbf{q}$ are points on such a line, we will always have $\mathbf{d}_{\mathrm{M}}(\mathbf{p}, \mathbf{q}) = 0$. In this case, we say $\mathbf{p}$ and $\mathbf{q}$ have *lightlike separation*.

- If a pair of points $\mathbf{p}$ and $\mathbf{q}$ are connected by a line segment with slope greater than 1 in absolute value, then the time coordinate of one is greater than the time coordinate of the other in *every* reference frame, and it is possible to get from one point to the other with subluminal (slower than light) motion. In this case, $\mathbf{d}_{\mathrm{M}}(\mathbf{p}, \mathbf{q})$ is real, and we say $\mathbf{p}$ and $\mathbf{q}$ have *timelike separation*.

- If a pair of points $\mathbf{p}$ and $\mathbf{q}$ are connected by a line segment with slope less than 1 in absolute value, then there are some coordinate systems in which $\mathbf{p}$ occurs before $\mathbf{q}$ and others in which $\mathbf{q}$ occurs before $\mathbf{p}$. No information can be transmitted between $\mathbf{p}$ and $\mathbf{q}$ – they are not *causally connected* – as this would require superluminal information transfer (and produce all sorts of paradoxes). In this case, $\mathbf{d}_{\mathrm{M}}(\mathbf{p}, \mathbf{q})$ is imaginary, and we say $\mathbf{p}$ and $\mathbf{q}$ have *spacelike separation*.

Notice that for an observer in the S' reference frame, stationary objects will have spacetime trajectories parallel to the ct'-axis – corresponding to no motion in the x' direction. Meanwhile, if the line segment connecting a pair of events in spacetime is parallel to the x'-axis, this observer will interpret these events as occurring simultaneously – they have the same ct' coordinate. To reiterate, lines parallel to the ct'-axis correspond to a constant position in S', while lines parallel in the x'-axis correspond to a constant time in S'. This would clearly be interpreted differently by an observer in the S reference frame, where these lines are non-constant in both the spatial coordinate x and time coordinate ct. As a result, observers in the two reference frames will disagree about how to measure lengths and times, in the sense that a length or time interval in one frame will include changes in both length and time in the other. For instance, suppose we have an object in the S' reference frame, and we wish to measure its length. Since the object is in the S' frame, its trajectory traces out a region bounded by lines parallel to the ct' axis – it is only "moving" in the ct' direction. If we are measuring in the S' frame, this length measurement will correspond to a spacetime interval obtained by intersecting this region with a line parallel to the x'-axis. If we are measuring in the S frame, it will correspond to a spacetime interval obtained by intersecting the region with a line parallel to the x-axis. The analysis is completely analogous if our object is in the S frame instead of the S' frame. See Figure 12 for the relevant spacetime diagrams.

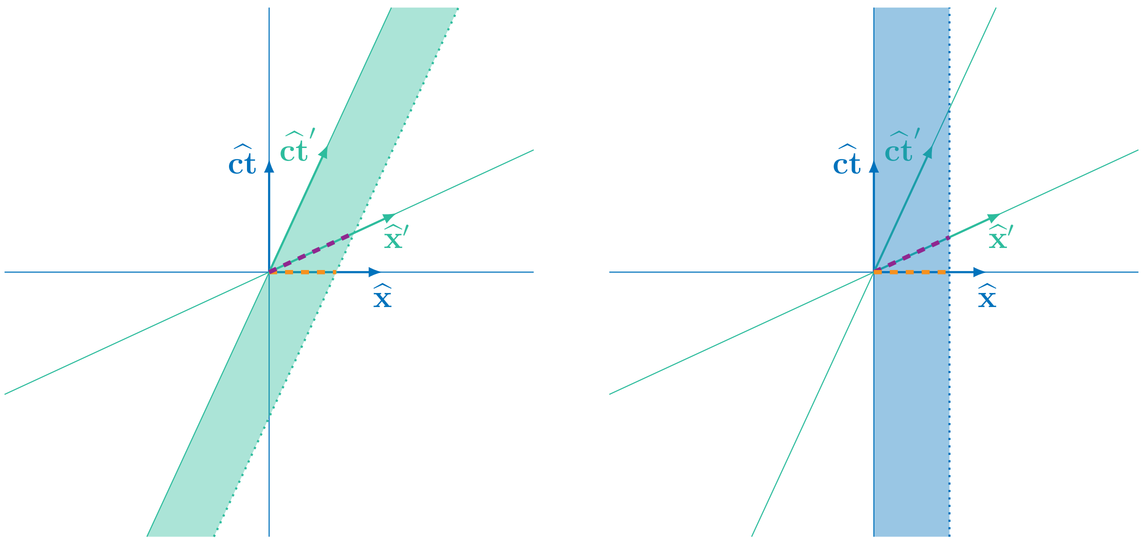

Figure 12: Intervals representing length measurements. On the left: The object is in the S' reference frame, and the dashed line segments represent a length measurement in S reference frame and a length measurement in S' reference frame. On the right: The object is in the S reference frame, and the dashed line segments represent a length measurement in S reference frame and a length measurement in S' reference frame.

The situation isn't perfectly symmetric if we exchange length measurements for time interval measurements. The difference amounts to how we interpret both situations. When we measure the length of an object, that object traces out a strip in spacetime, with sides parallel to the time axis of the object's reference frame. On the other hand, we keep track of time with clocks, and a single clock doesn't trace out a strip in spacetime with side lengths parallel to the x-axis of the clock's reference frame – the clock is localized in space, while the object travels through time.[11] So, suppose we have a clock in the S reference frame, and we observe from the S' reference frame as it records an elapsed time of $c\tau$ in the S frame. How much time $c\tau'$ does this take our S' coordinate system? Well, the clock itself is stationary in S, meaning it has the same x coordinate throughout. This time interval corresponds to a line segment parallel to the ct-axis, with the ends of the segment representing the beginning and end of the time interval. If we want to evaluate this line segment from S' frame, we can draw a *line of simultaneity* in S' – a line in which all points have the same ct' coordinate – for each endpoint of the line segment. These are lines parallel to the x' axis that intersect the line segment associated to the time interval at its endpoints. This pair of lines of simultaneity also intersects the ct'-axis, and so provides an interval on this segment as well, corresponding to elapsed time in S'. The analysis is completely analogous if our object is in the S frame instead of the S' frame. See Figure 13 for the relevant spacetime diagrams.

[11] Of course, we could imagine a situation in which a system of clocks is established to create such a strip. That isn't the situation we will address here, or the one that leads us to *time dilation*.

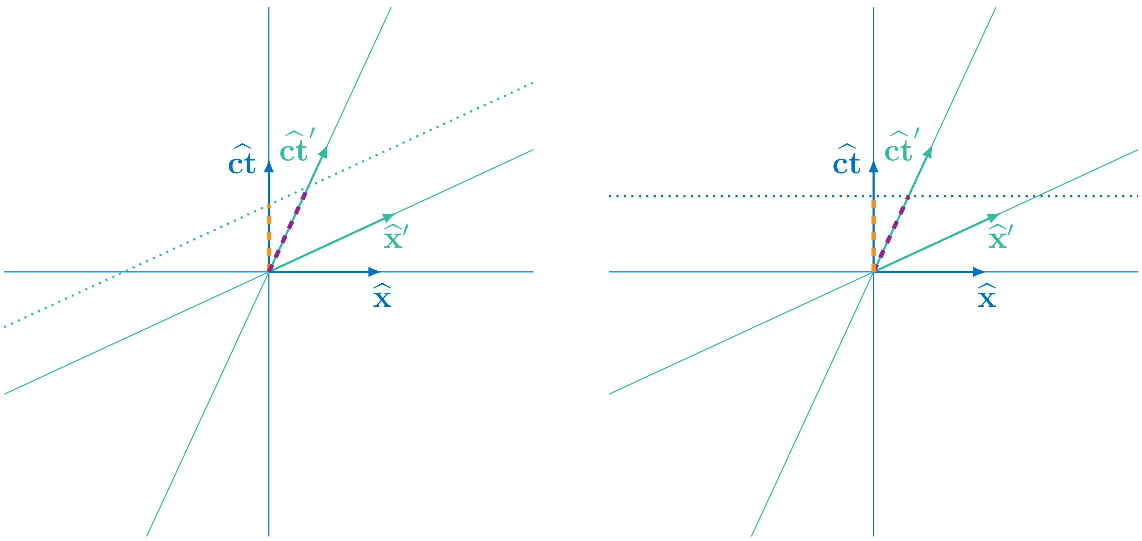

Figure 13: Intervals representing time measurements. On the left: The clock is in the S reference frame, and the dashed line segments represent a time measurement in S reference frame and a time measurement in S' reference frame. On the right: The clock is in the S' reference frame, and the dashed line segments represent a time measurement in S reference frame and a time measurement in S' reference frame.

However, even if the pair of observers were to agree on a spacetime interval to measure, they would not agree on the length or elapsed time of this spacetime interval. They would instead agree on the Minkowski length of the spacetime interval – after all, Minkowski lengths are what Lorentz transformations are meant to preserve.

Now let's compare the measured length and time measurements in S and S' algebraically. We'll first consider an object of length ℓ in the S reference frame and measure its length ℓ' in the S' reference frame. This corresponds to the spacetime diagram shown in Figure 14.

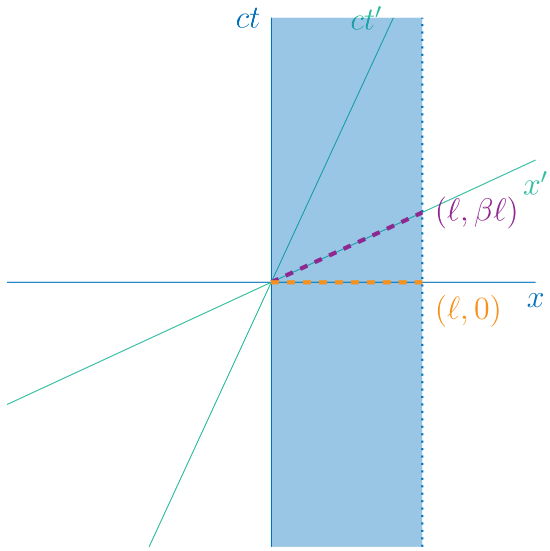

Figure 14: Spacetime diagram for an object of length ℓ in the S frame. Its length as measured in the S' frame is the norm of the Minkowski length of the dashed segment along the x'-axis.

Since Minkowski lengths are the same in both frames, we have

$$\begin{aligned} -\ell'^2 &= -\ell^2 + (\beta\ell)^2 \\ &= -\ell^2\left(1-\beta^2\right), \end{aligned}$$

so,

$$\ell' = \frac{\ell}{\gamma}. \tag{24}$$

Equation (24) is referred to as *length contraction*. Note that in hyperbolic trigonometry terms, Equation (24) can be rewritten as

$$\operatorname{mcos}(\alpha') = \frac{\ell}{\ell'},$$

which can be obtained immediately from Figure 14 and the definition of $\operatorname{mcos}(\alpha')$. The direction of motion (as in positive x direction as opposed to negative x direction) was entirely irrelevant to this calculation, so the roles of S and S' can be swapped in this analysis. The length of an object in its own reference frame is called its *proper length*.

Next let's consider a time interval $c\tau$ in the S reference frame and ask how a clock in the S' reference frame would record the elapsed time $c\tau'$. This corresponds to the spacetime diagram shown in Figure 15.

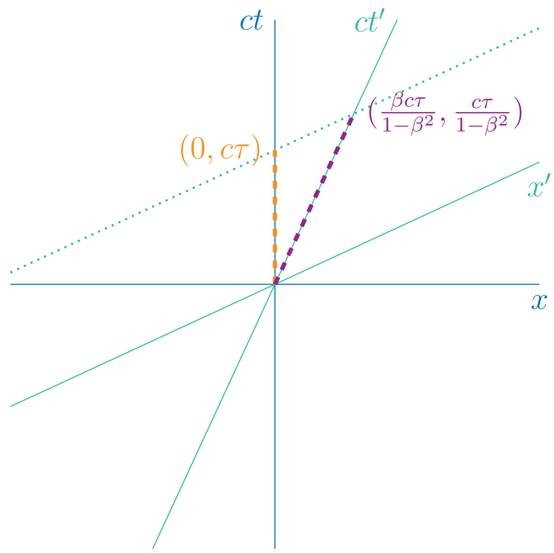

Figure 15: Spacetime diagram for a time interval $c\tau$ in the S frame. The length of this time interval as measured in the S' frame is the Minkowski length of the dashed segment along the ct'-axis.

Since Minkowski lengths are the same in both frames, we have

$$\begin{aligned}(c\tau')^2 &= -\left(\frac{\beta c\tau}{1-\beta^2}\right)^2 + \left(\frac{c\tau}{1-\beta^2}\right)^2 \\ &= (c\tau)^2\left(\frac{1-\beta^2}{(1-\beta^2)^2}\right) \\ &= (c\tau)^2\left(\frac{1}{1-\beta^2}\right),\end{aligned}$$

so,

$$c\tau' = \gamma c\tau. \tag{25}$$

Equation (25) is referred to as *time dilation.* In hyperbolic trigonometry terms, Equation (25) can be rewritten as

$$\mathrm{mcos}(\alpha') = \frac{c\tau'}{c\tau},$$

which can be deduced from Figure 15 and the definition of $\mathrm{mcos}(\alpha')$. As with length contraction, the direction of motion was entirely irrelevant to this calculation, so the roles of S and S' can be swapped in this analysis. The elapsed time a clock records in its own reference frame is the *proper time* of that reference frame.

3.4 Constant Acceleration

Thus far we have translated between reference frames of *inertial* observers – meaning non-accelerating observers. Our final topic of the course will be Special Relativity with constant acceleration. Since lengths and times are observer-dependent, the notion of constant acceleration is a bit subtle here. Some physics considerations can give us a hint at the "right" notion. While we have abandoned the idea of absolute velocities, having been convinced by Poincaré and Einstein that only relative velocities are meaningful, we actually do not need to abandon the idea of an absolute acceleration, if properly interpreted. We have Newton's second law of motion: $\mathbf{F} = m\mathbf{a}$. We can measure forces, and we'll use this to determine how we *should* define acceleration. From this perspective, we would measure acceleration from the point of view of an inertial observer instantaneously co-moving with the accelerated observer. That is, at the instant of measurement, the inertial observer and the accelerated observer would be at rest with respect to each other. Acceleration measured this way is called *proper acceleration.*

Let's determine the trajectory of a rocket with constant proper acceleration.[12] To do this, we'll consider a system with three different reference frames, two of which are inertial and one is the non-inertial reference frame of the rocket. We'll take all relative motion to occur in the (x, ct)-plane of each coordinate system, and we will ignore the y and z directions. Let's take our first inertial reference frame S to be the rest frame of the rocket immediately prior to the start of acceleration. We will express the trajectory in this frame. The next inertial reference frame S' will be the frame instantaneously co-moving with the rocket at some point after the rocket begins accelerating. We'll match up the origins of S and S' so the two frames are related by a Lorentz boost, with no need for a translation. Let's denote the rocket's velocity in S by v_R, the rocket's velocity in S' by v'_R, and the velocity of the S' frame in the S frame by w.

Notation 33. In the analysis that follows, we will want to associate β's, Lorentz factors γ, and Minkowski angles α to all three of these velocities. We'll use subscripts to keep track of which velocity we are considering in each instance. So, if the relevant velocity is v, we will write β_v, γ_v, and α_v.

When we say the rocket has proper acceleration a, this means $\frac{dv'_R}{dt'} = a$. We can write this in more convenient terms if we divide by c^2:

$$\frac{a}{c^2} = \frac{d}{d(ct')}\left(\frac{v'_R}{c}\right) = \frac{d\beta_{v'_R}}{d(ct')}. \tag{26}$$

Question 34. Why are we considering the rate of change of the velocity of the rocket in the S' frame rather than in the rockets own frame? In other words, why didn't we define the proper acceleration to be the rate of change of the rocket's proper velocity with respect to the rocket's proper time?

[12]What follows here is based on a nice answer of "Alfred Centauri" on *Physics StackExchange.*

We would like to express Equation (26) in the S reference frame in order to work out the rocket's trajectory. So, we will want to be able to express $\beta_{v'_R}$ in terms of β_{v_R} and β_w. The Lorentz boosts associated to these relative velocities compose by addition of Minkowski angles:

$$\alpha_{v_R} = \alpha_w + \alpha_{v'_R}.$$

Rearranging, we have

$$\alpha_{v'_R} = \alpha_{v_R} - \alpha_w.$$

Now recall that $\beta_v = i\,\mathrm{mtan}(\alpha_v)$, so

$$\beta_{v'_R} = i\,\mathrm{mtan}(\alpha_{v'_R}) = i\,\mathrm{mtan}(\alpha_{v_R} - \alpha_w).$$

But the exponential expression for the Minkowski tangent has exactly the same form as the exponential expression for the usual Euclidean tangent, so the angle addition formula for tangent applies to Minkowski tangent as well. Then

$$\begin{aligned} i\,\mathrm{mtan}(\alpha_{v_R} - \alpha_w) &= i\frac{\mathrm{mtan}(\alpha_{v_R}) - \mathrm{mtan}(\alpha_w)}{1 + \mathrm{mtan}(\alpha_{v_R})\,\mathrm{mtan}(\alpha_w)} \\ &= \frac{i\,\mathrm{mtan}(\alpha_{v_R}) - i\,\mathrm{mtan}(\alpha_w)}{1 - (i\,\mathrm{mtan}(\alpha_{v_R}))\,(i\,\mathrm{mtan}(\alpha_w))} \\ &= \frac{\beta_{v_R} - \beta_w}{1 - \beta_{v_R}\beta_w}, \end{aligned}$$

and we conclude that

$$\beta_{v'_R} = \frac{\beta_{v_R} - \beta_w}{1 - \beta_{v_R}\beta_w}. \tag{27}$$

Next, in Equation (26), we differentiate with respect to ct'. We want to express this as a derivative in the S frame. Using Equation (23), we have

$$ct' = \gamma_w\left(ct - \beta_w x\right).$$

For any point (x, ct) on the ct'-axis, we have $x = \beta_w ct$, so along this line we have

$$ct' = \gamma_w\left(ct - {\beta_w}^2 ct\right) = \frac{ct}{\gamma_w},$$

and

$$\frac{d\beta_{v'_R}}{d(ct')} = \gamma_w \frac{d\beta_{v'_R}}{d(ct)}. \tag{28}$$

Since both S and S' are inertial, w and in turn γ_w are constants. So, using Equations (27) and (28), Equation (26) becomes

$$\begin{aligned}
\frac{a}{c^2} = \frac{d\beta_{v'_R}}{d(ct')} &= \gamma_w \frac{d}{d(ct)} \left(\frac{\beta_{v_R} - \beta_w}{1 - \beta_{v_R}\beta_w} \right) \\
&= \gamma_w \left(\frac{(1 - \beta_{v_R}\beta_w) \frac{d}{d(ct)} (\beta_{v_R} - \beta_w) - (\beta_{v_R} - \beta_w) \frac{d}{d(ct)} (1 - \beta_{v_R}\beta_w)}{(1 - \beta_{v_R}\beta_w)^2} \right) \\
&= \gamma_w \left(\frac{(1 - \beta_{v_R}\beta_w) \frac{d\beta_{v_R}}{d(ct)} - (\beta_{v_R} - \beta_w)(-\beta_w) \frac{d\beta_{v_R}}{d(ct)}}{(1 - \beta_{v_R}\beta_w)^2} \right) \\
&= \gamma_w \left(\frac{1 - \beta_{v_R}\beta_w + \beta_{v_R}\beta_w - \beta_w{}^2}{(1 - \beta_{v_R}\beta_w)^2} \right) \frac{d\beta_{v_R}}{d(ct)} \\
&= \frac{1}{\gamma_w (1 - \beta_{v_R}\beta_w)^2} \frac{d\beta_{v_R}}{d(ct)}.
\end{aligned}$$

Now let's look at the particular instant in which the rocket and S' are co-moving. In this case, $v_R = w$, and we have

$$\begin{aligned}
\frac{a}{c^2} &= \frac{1}{\gamma_{v_R} \left(1 - \beta_{v_R}{}^2\right)^2} \frac{d\beta_{v_R}}{d(ct)} \\
&= \frac{\gamma_{v_R}{}^4}{\gamma_{v_R}} \frac{d\beta_{v_R}}{d(ct)} \\
&= \gamma_{v_R}{}^3 \frac{d\beta_{v_R}}{d(ct)}.
\end{aligned}$$

We are now ready to write down the differential equation whose solution will be our rocket's trajectory, as described in the S reference frame. Denote the rocket's position in the S frame by x_R. Then

$$\beta_{v_R} = \frac{dx_R}{d(ct)},$$

so

$$\frac{a}{c^2} = \gamma_{v_R}{}^3 \frac{d^2 x_R}{d(ct)^2}. \tag{29}$$

Question 35. Consider the function

$$x_R(ct) = \frac{c^2}{a} \left(\sqrt{1 + \left(\frac{a}{c^2}\right)^2 (ct)^2} - 1 \right).$$

Show that

$$\frac{d^2 x_R}{d(ct)^2} = \frac{a}{c^2} \left(1 + \left(\frac{a}{c^2}\right)^2 (ct)^2 \right)^{-\frac{3}{2}}$$

and

$$\gamma_{v_R}{}^3 = \left(1 + \left(\frac{a}{c^2}\right)^2 (ct)^2\right)^{\frac{3}{2}}.$$

Conclude that $x_R(ct)$ is a solution to Equation (29).

Your final project is to apply this to investigate a hypothetical road trip to Proxima Centauri b, with a constant proper acceleration of $1g$ for the first half of the voyage, followed by deceleration of $1g$ for the second half. There will be many possible approaches for analyzing this situation, but the easiest and most elegant option is to use what you've learned of hyperbolic trigonometry. Good luck and have fun!