Notes on Geometry and Spacetime

Version 2.7, November 2009

David B. Malament
Department of Logic and Philosophy of Science
University of California, Irvine
dmalamen@uci.edu

Contents

These notes have not yet been fully de-bugged. Please read them with caution. (Corrections will be much appreciated.)

1 Preface

The notes that follow bring together a somewhat unusual collection of topics.

In section 3, I discuss the foundations of "special relativity". I emphasize the invariant, "geometrical approach" to the theory, and spend a fair bit of time on one special topic: the status of the relative simultaneity relation within the theory. At issue is whether the standard relation, the one picked out by Einstein's "definition" of simultaneity, is conventional in character, or is rather in some significant sense forced on us.

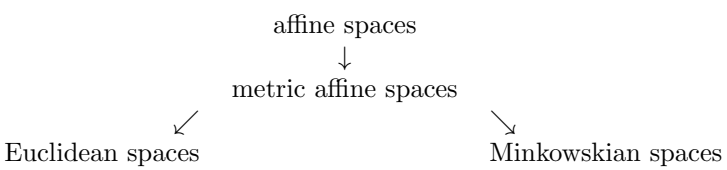
Section 2 is preparatory. When the time comes, I take "Minkowski spacetime" to be a four-dimensional affine space endowed with a Lorentzian inner product. So, to prepare the way, I first give a brief account of "metric affine spaces" that is sufficiently general to include the Minkowskian variety. There is more material in this section than is strictly necessary for what follows. But it is helpful to have it available for reference purposes.

Section 4 is an afterthought. It deals with non-Euclidean (i.e., hyperbolic) plane geometry. This geometry was, of course, first developed by Gauss, Lobatchevsky, Bolyai *et al.* in the 19th century. But it turns out that one of the nicest routes to it is via special relativity. More precisely: one gets a simple, easily visualized model of hyperbolic plane geometry (the so-called "hyperboloid model") if one starts with three-dimensional Minkowski spacetime, and then restricts attention to a particular surface in it. With Minkowskian geometry in hand, very little additional work is needed to develop this application; and it is very tempting indeed to do it!

2 Preliminary Mathematics

It is one of the assumptions behind this course that a relatively modest investment in abstract mathematics pays significant dividends for the understanding of the special theory of relativity and recent debates among philosophers concerning its foundations. It provides the resources with which to pose and answer precise questions that bear on those debates.

We begin the investment in this first part of the course. Here we review certain facts about: (i) affine spaces, (ii) metric affine spaces, and (iii) Euclidean spaces. This will help prepare the way for our consideration of Minkowskian spaces in the next. (It is helpful to think of Euclidean and Minkowskian spaces as but different species of metric affine space, and develop them in parallel.)



2.1 Vector Spaces

One has to start somewhere. In that follows, we take for granted familiarity with basic facts about vector spaces and linear maps (equivalent to, say, chapters 1 and 3 of Lang [8]). Here we give a quick summary to establish notation and terminology, and list a number of problems for review purposes.

A *vector space* (over $\mathbb{R}$) is a structure $(V, +, \mathbf{0}, \cdot)$ where V is a set (whose elements are called "vectors"), $\mathbf{0}$ is an element of V (the "zero element"), $+$ is a map from $V \times V$ to V ("vector addition") and $\cdot$ is a map from $\mathbb{R} \times V$ to V ("scalar multiplication") satisfying the following conditions.

(VS1) For all u, v in V, $u + v = v + u$.

(VS2) For all u, v, w in V, $(u + v) + w = u + (v + w)$.

(VS3) For all u in V, $u + \mathbf{0} = u$.

(VS4) For all u in V, there is a v in V such that $u + v = \mathbf{0}$.

(VS5) For all a in $\mathbb{R}$, and all u, v in V, $a \cdot (u + v) = a \cdot u + a \cdot v$.

(VS6) For all a, b in $\mathbb{R}$, and all u in V, $(a + b) \cdot u = a \cdot u + b \cdot u$.

(VS7) For all a, b in $\mathbb{R}$, and all u in V, $(a\, b) \cdot u = a \cdot (b \cdot u)$.

(VS8) For all u in V, $1 \cdot u = u$.

(Of course, one sometimes considers vector spaces defined over fields other than $\mathbb{R}$. But we have no need to do so. For us, a "vector space" will always be a vector space over the reals.) Sometimes we will abuse our notation and write $a\, u$ rather than $a \cdot u$. And sometimes, following standard practice, we will be a bit casual about the distinction between a vector space $\mathbf{V} = (V, +, \mathbf{0}, \cdot)$, and its underlying point set V. We will refer to "the vector space V", etc.

In what follows, let $(V, +, \mathbf{0}, \cdot)$ be a vector space.

Problem 2.1.1. *Prove that for all vectors u in V, there is a* unique *vector v in V such that $u + v = \mathbf{0}$. (We write the latter as $(-u)$; given vectors u and v, we sometimes write $(u - v)$ for $(u + (-v))$.)*

Problem 2.1.2. *Prove that for all vectors u in V, if $u + u = u$, then $u = \mathbf{0}$.*

Problem 2.1.3. *Prove that for all vectors u in V, and all real numbers a,*

(i) $0 \cdot u = \mathbf{0}$

(ii) $-u = (-1) \cdot u$

(iii) $a \cdot \mathbf{0} = \mathbf{0}$.

Example 2.1.1. *For every $n \geq 1$, the set $\mathbb{R}^n = \{(a_1, ..., a_n) : a_i \in \mathbb{R}$ for all $i\}$ has a natural vector space structure $(\mathbb{R}^n, +, \mathbf{0}, \cdot)$, where $\mathbf{0}$ is $(0, ..., 0)$, and the operations $+$ and $\cdot$ are defined by*

$$\begin{aligned} (a_1, ..., a_n) + (b_1, ..., b_n) &= (a_1 + b_1, ..., a_n + b_n) \\ a \cdot (b_1, ..., b_n) &= (ab_1, ..., ab_n). \end{aligned}$$

A nonempty subset W of V is said to be a *(linear) subspace* of V if it satisfies two conditions:

(SS1) For all u, v in V, if u and v both belong to W, then so does $(u + v)$.
(SS2) For all u in V, and all a in $\mathbb{R}$, if u belongs to W, then so does $(a\,u)$.

So, for example, the set of all vectors of the form $(a, 0, 0)$ is a subspace of $\mathbb{R}^3$, and so is the set of all vectors of the form $(a, 0, c)$. But the set of all vectors of the form $(a, 0, a + 1)$ is *not* a subspace of $\mathbb{R}^3$.

If W is a subspace of V, then it forms a vector space in its own right if we define vector addition and scalar multiplication as in V, and use the same zero element $\mathbf{0}$ as in V. (Question: How do we know that if W is a subspace of V, $\mathbf{0}$ belongs to W?) Conditions (SS1) and (SS2) are precisely the conditions needed to guarantee that these operations are well defined over W.

Problem 2.1.4. *Prove that the intersection of any non-empty set of subspaces of V is a subspace of V.*

Let S be a subset of V, and let $L(S)$ be the intersection of all subspaces of V that contain S (as a subset). Then $L(S)$ is itself a subspace of V. (This follows from the result in problem 2.1.4.) We call it the *(sub)space spanned* by S or the *linear span* of S. This definition makes sense for any subset S of V, empty or non-empty, finite or infinite. (The linear span of the empty set $\varnothing$ turns out to be the singleton set $\{\mathbf{0}\}$. This follows since every subspace of V contains $\varnothing$ as a subset, and the intersection of all subspaces is $\{\mathbf{0}\}$.) But if S is non-empty, an equivalent (and more familiar) characterization is available. In this case, we can then take $L(S)$ to be the set of all (finite) *linear combinations* of vectors in S, i.e., all vectors of form $a_1u_1 + ... + a_ku_k$, where $k \geq 1$ and $u_1, ..., u_k$ are elements of S. (Question: Why are we justified in writing the sum without parentheses?)

Problem 2.1.5. *Let S be a subset of V. Show that $L(S) = S$ iff S is a subspace of V.*

We say that S is *linearly dependent* if there exist (distinct) vectors $u_1, ..., u_k$ ($k \geq 1$) in S and coefficients $a_1, ...a_k$, not all 0, such that $a_1u_1 + ... + a_ku_k = \mathbf{0}$. (Otherwise, it is *linearly independent*.) For example, $\{(1, 0, 0), (0, 1, 0), (1, 1, 0)\}$ is a linearly dependent subset of $\mathbb{R}^3$, but $\{(1, 0, 0), (1, 1, 0), (1, 1, 1)\}$ is linearly independent. Note that if a subset S of V contains $\mathbf{0}$, then it is automatically linearly dependent. (This follows since $a\,\mathbf{0} = \mathbf{0}$ for all a.) Hence every subspace of V is linearly dependent. Note also that the empty set $\varnothing$ qualifies as linearly independent.

Problem 2.1.6. *Let S be a subset of V. Show that S is linearly dependent iff there is a vector u in S that belongs to the linear span of $(S - \{u\})$.*

Finally, we say that S is a *basis* of V if (i) S is linearly independent, and (ii) S spans V, i.e., $L(S) = V$. So, for example, the set $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ is a basis for $\mathbb{R}^3$ since it is linearly independent and every element (a_1, a_2, a_3) in $\mathbb{R}^3$ can be expressed in the form $a_1(1, 0, 0) + a_2(0, 1, 0) + a_3(0, 0, 1)$. Note, as well, that the empty set qualifies as a basis for a (trivial) vector space with only one element (the zero element $\mathbf{0}$). (We defined linear spans as we did, in part,

to secure this result. One would like to be able to assert that all vector spaces have bases. If we had taken the linear span of a set S to be the set of finite linear combinations of elements in S – even when S is the empty set – then the trivial vector space $\{\mathbf{0}\}$ would not have had one.)

In the next proposition we collect several important facts about bases. To simplify our formulation, we limit attention to the case where V is *finite dimensional*, i.e., the case where there exists a finite subset S of V with $L(S) = V$. (So, for example, the vector space $\mathbb{R}^n$ is finite dimensional for every $n \geq 1$.) Not all vector spaces are finite dimensional, but they are the only ones of concern to us in what follows.

Proposition 2.1.1. *Let V be finite dimensional. Then all the following hold.*

(i) *There exists a finite basis for V.*

(ii) *All bases for V have the same (finite) number of elements. (That number is called the* dimension *of V, and is denoted $dim(V)$.)*

(iii) *If $dim(V) = n$, every linearly independent subset of V with n elements is a basis for V.*

(iv) *If $dim(V) = n$, every subset of V with n elements that spans V is a basis for V.*

(v) *If W is a subspace of V, W is finite dimensional, and $dim(W) \leq dim(V)$.*

We skip the proof. It can be found in Lang [8] and almost any other basic text in linear algebra.

For all $n \geq 1$, the vector space $\mathbb{R}^n$ has dimension n. A trivial vector space with one element (the zero vector $\mathbf{0}$) has dimension 0 (since the empty set is a basis for the space). As we will observe in a moment, these are, "up to isomorphism", the only finite dimensional vector spaces.

If $\{u_1, ..., u_n\}$ is a basis for V ($n \geq 1$), every vector u in V can be expressed uniquely in the form $u = a_1u_1 + ... + a_nu_n$. That u can be expressed in this form at all follows from the fact that u belongs to the linear span of $\{u_1, ..., u_n\}$, namely V. That the expression is unique follows from the linear independence of $\{u_1, ..., u_n\}$. (For if we had two representations

$$a_1u_1 + ... + a_nu_n = u = b_1u_1 + ... + b_nu_n,$$

it would have to be the case that

$$(a_1 - b_1)u_1 + ... + (a_n - b_n)u_n = \mathbf{0}.$$

And so, by linear independence, it would have to be the case that $a_i = b_i$ for all i.)

Now consider two vectors spaces $(V, +, 0, \cdot)$ and $(V', +', 0', \cdot')$. A map $\Phi : V \to V'$ is *linear* if, for all u, v in V, and all a in $\mathbb{R}$,

$$\begin{aligned} \Phi(u + v) &= \Phi(u) +' \Phi(v) \\ \Phi(a \cdot u) &= a \cdot' \Phi(u). \end{aligned}$$

These two conditions imply that $\Phi(\mathbf{0}) = \mathbf{0}'$, and

$$\Phi(a_1 \cdot u_1 + ... + a_k \cdot u_k) = a_1 \cdot' \Phi(u_1) +' ... +' a_k \cdot' \Phi(u_k).$$

for all $k \geq 1$, all real numbers $a_1, a_2, ..., a_k$, and all vectors $u_1, u_2, ..., u_k$ in V. It follows that $\Phi[V]$ is a subspace of V', and that, for all subsets S of V, $\Phi[L(S)] = L(\Phi[S])$. (Two points about notation: (1) If T is a subset of V, $\Phi[T]$ is here understood to be the range of T under Φ, i.e, the set all vectors of form $\Phi(u)$ in V' where u is a vector in V. (2) When there is no danger of confusion, we will use a uniform notation for the vector space operations and zero elements in different vector spaces. That will allow us to write the equation above, more simply, in the form:

$$\Phi(a_1u_1 + ... + a_ku_k) = a_1\Phi(u_1) + ... + a_k\Phi(u_k).)$$

If $\Phi : V \to V'$ is a linear map, the *kernel* of Φ, $\ker(\Phi)$, is the set of all elements u in V that are mapped to $\mathbf{0}'$, i.e., $\ker(\Phi) = \{u \in V : \Phi(u) = \mathbf{0}'\}$. It follows easily that $\ker(\Phi)$ is a subspace of V, and that Φ is injective (i.e., one-to-one) iff $\ker(\Phi) = \{\mathbf{0}\}$. We say that the linear map $\Phi : V \to V'$ is an *isomorphism* if Φ is injective and $\Phi[V] = V'$. Of course, V and V' are *isomorphic* if there exists an isomorphism $\Phi : V \to V'$. (Remark: It might seem, at first glance, that our definition of "isomorphism" allows for an inappropriate asymmetry. We require that Φ be linear, but do not impose the requirement on the inverse map Φ^{-1}. But, in fact, it is easy to check that the inverse map of an isomorphism *must* be linear. There is no need to add a clause to the definition.)

Problem 2.1.7. *Show that two finite dimensional vector spaces are isomorphic iff they have the same dimension.*

2.2 Affine Spaces

Vector spaces have a distinguished $\mathbf{0}$ element. Thus they are not appropriate for representing homogeneous spacetime structure. An "affine space" can be thought of as a vector space with the $\mathbf{0}$ element washed out. More precisely, we have the following definition.

An *affine space* is a structure $(A, \mathbf{V}, +)$ where A is a non-empty set, $\mathbf{V}$ is a vector space $(V, +, \mathbf{0}, \cdot)$, and $+$ is a map from $A \times V$ to A satisfying the following conditions.

(AS1) For all p, q in A, there is a unique u in V such that $q = p + u$.

(AS2) For all p in A, and all u, v in V, $(p + u) + v = p + (u + v)$.

(Here our notation is imperfect because the symbol '+' is used for two different maps: the old map from $V \times V$ to V (addition within the vector space $\mathbf{V}$), and the new one from $A \times V$ to A. But no ambiguity arises in practice.) If $q = p + u$, we write u as $\overrightarrow{pq}$. Thus, $q = p + \overrightarrow{pq}$. We refer to the elements of A as "points", and refer to $\overrightarrow{pq}$ as the "vector that runs from p to q". Behind the formalism is an intuitive picture. We think of $\overrightarrow{pq}$ as an *arrow* with tail p and head q. (So

the assertion $q = p + \overrightarrow{pq}$ can be understood to mean that if one starts at p, and then follows the arrow $\overrightarrow{pq}$, one ends up at q.)

In what follows, let $(A, \mathbf{V}, +)$ be a finite dimensional affine space. (We understand the *dimension* of an affine space to be the dimension of its underlying vector space.)

Proposition 2.2.1. *For all points p, q, r in A,*

(i) $\overrightarrow{pp} = \mathbf{0}$ *(or, equivalently, $p + \mathbf{0} = p$)*

(ii) $\overrightarrow{pq} = \mathbf{0} \;\Rightarrow\; p = q$

(iii) $\overrightarrow{qp} = -\overrightarrow{pq}$

(iv) $\overrightarrow{pq} + \overrightarrow{qr} = \overrightarrow{pr}$

Proof. (i) By (AS1), there is a unique u such that $p + u = p$. Hence, using (AS2),

$$p + (u + u) = (p + u) + u = p + u = p.$$

So, by uniqueness, $u + u = u$ and, therefore (recall problem 2.1.2), $u = \mathbf{0}$. Thus $p + \mathbf{0} = p$, i.e., $\overrightarrow{pp} = \mathbf{0}$.
(ii) Assume $\overrightarrow{pq} = \mathbf{0}$. Then $q = p + \overrightarrow{pq} = p + \mathbf{0} = p$.
(iii) If $q = p + u$, then $u = \overrightarrow{pq}$, and $q + (-u) = (p + u) + (-u) = p + (u + (-u)) = p + \mathbf{0} = p$. So $\overrightarrow{qp} = -u = -\overrightarrow{pq}$.
(iv) If $q = p + u$ and $r = q + v$, then $u = \overrightarrow{pq}$, $v = \overrightarrow{qr}$, and $r = (p + u) + v = p + (u + v)$. So $\overrightarrow{pr} = u + v = \overrightarrow{pq} + \overrightarrow{qr}$. □

It is important that the points of an affine space (elements of A) are not themselves vectors. (Or, at least, they need not be. No restrictions have been placed on the set A other than those that follow from (AS1) and (AS2).) But given any p in A, the rule of association $q \mapsto \overrightarrow{pq}$ determines a one-to-one map of A onto V. (We can think of p as an arbitrarily chosen "origin".) To verify that it is one-to-one, assume that q and r are points in A that are assigned the same vector, i.e., $\overrightarrow{pr} = \overrightarrow{pq}$. Then, by clauses (iv) and (iii) of the proposition,

$$\overrightarrow{qr} = \overrightarrow{qp} + \overrightarrow{pr} = (-\overrightarrow{pq}) + \overrightarrow{pq} = \mathbf{0}.$$

Hence, $r = q$ by clause (ii). To verify that the rule of association maps A *onto* V, we must show that given any vector u in V, there is a point q in A such that $\overrightarrow{pq} = u$. But this condition is clearly satisfied by the point $q = p + u$.

Given any point p in A and any subspace W of V, the set

$$p + W = \{p + u \colon u \in W\}$$

is called the *affine subspace of A through p determined by W*. If W is one dimensional, we call the set a *line*; if it is two dimensional, we call it a *plane*; and so forth.

Problem 2.2.1. *Show that for all points p and q in A, and all subspaces W of V, the following conditions are equivalent.*

(i) *q belongs to $p+W$*

(ii) *p belongs to q+W*

(iii) $\overrightarrow{pq} \in W$

(iv) *p+W and q+W coincide (i.e., contain the same points)*

(v) *p+W and q+W intersect (i.e., have at least one point in common)*

Problem 2.2.2. *Let $p_1 + W_1$ and $p_2 + W_2$ be lines, and let u_1 and u_2 be non-zero vectors, respectively, in W_1 and W_2. Show that the lines intersect iff $\overrightarrow{p_1 p_2}$ is a linear combination of u_1 and u_2.*

We say that the lines $p_1 + W_1$ and $p_2 + W_2$ are *parallel* if $W_1 = W_2$. (We are allowing a line to count as parallel to itself.) An equivalent characterization is given in the following proposition. The equivalence should seem obvious, but a bit of work is necessary to give a complete proof.

Proposition 2.2.2. *Two lines are parallel iff either they coincide, or they are co-planar (i.e., subsets of some plane) and do not intersect.*

Proof. Let $p_1 + W_1$ and $p_2 + W_2$ be any two lines, and let u_1 and u_2 be non-zero vectors, respectively, in W_1 and W_2.

Assume first that the lines are parallel ($W_1 = W_2$), but do not coincide. Then, by problem 2.2.1, they do not intersect and $\overrightarrow{p_1 p_2} \notin W_1$. The latter assertion implies that the vectors u_1 and $\overrightarrow{p_1 p_2}$ are linearly independent and, so, span a two-dimensional subspace W of V. To complete the first half of the proof, it will suffice for us to show that the lines $p_1 + W_1$ and $p_2 + W_2$ are both subsets of the plane $p_1 + W$. Certainly $p_1 + W_1$ is a subset of $p_1 + W$, since W_1 is a subset of W. Similarly, $p_2 + W_2$ is a subset of $p_2 + W$. But $p_1 + W = p_2 + W$. (This follows from problem 2.2.1 again and the fact that $\overrightarrow{p_1 p_2} \in W$.) So, as claimed, both lines are subsets of $p_1 + W$.

For the converse, assume first that $p_1 + W_1 = p_2 + W_2$. Then p_2 belongs to $p_1 + W_1$ and, therefore, $p_2 = p_1 + k_1 u_1$ for some k_1. So $\overrightarrow{p_1 p_2} = k_1 u_1 \in W_1$. It follows (by problem 2.2.1) that $p_1 + W_1 = p_2 + W_1$. So $p_2 + W_1 = p_2 + W_2$. Hence (since $p_2 + u_1$ clearly belongs to $p_2 + W_1$), $p_2 + u_1$ belongs to $p_2 + W_2$. So there is a number k_2 such that $p_2 + u_1 = p_2 + k_2 u_2$. It follows that $p_2 = p_2 + (u_1 - k_2 u_2)$ and, therefore, $u_1 = k_2 u_2$. Thus the non-zero vectors u_1 and u_2 are linearly dependent. So $W_1 = W_2$, i.e., our lines are parallel.

Alternatively – we are still working on the converse – assume that the lines $p_1 + W_1$ and $p_2 + W_2$ do not intersect, and are both subsets of the plane $q + W$ (where W is some two-dimensional subspace of V). Since p_1 and p_2 both belong to $q + W$, it must be the case (problem 2.2.1) that $\overrightarrow{p_1 q} \in W$ and $\overrightarrow{p_2 q} \in W$. So, since W is a subspace, $\overrightarrow{p_1 p_2} \in W$. Furthermore, since $p_1 + W_1$ and $p_2 + W_2$ are subsets of $q + W$, it must be the case that u_1 and u_2 belong to W. (Consider the point $r = p_1 + u_1$ in $p_1 + W_1$. Since it belongs to $q + W$, $\overrightarrow{qr} \in W$. But $u_1 = \overrightarrow{p_1 r} = \overrightarrow{p_1 q} + \overrightarrow{q r}$. So, since both $\overrightarrow{p_1 q}$ and $\overrightarrow{q r}$ belong to W, u_1 does too. And similarly for u_2.) Since the three vectors u_1, u_2, and $\overrightarrow{p_1 p_2}$ all belong to a two-dimensional subspace (namely W), they cannot be linearly independent. So there are numbers a, b, c, not all 0, such that $a u_1 + b u_2 + c \overrightarrow{p_1 p_2} = \mathbf{0}$. But c must be 0. Otherwise, we could divide by c and express $\overrightarrow{p_1 p_2}$ as a linear

combination of u_1 and u_2. And this, by problem 2.2.2, would contradict our assumption that the lines $p_1 + W_1$ and $p_2 + W_2$ do not intersect. So u_1 and u_2 are linearly dependent. Thus, in this case too, $W_1 = W_2$, i.e, our lines are parallel. □

Let p and q be any two (distinct) points in A, and let W be the subspace of V spanned by the vector $\overrightarrow{pq}$. We take the *line determined by p and q* to be the set

$$L(p,q) = p + W = \{p + a\,\overrightarrow{pq}\colon\ a \in \mathbb{R}\}.$$

It is easy to verify (e.g., as a consequence of problem 2.2.1) that $L(q,p) = L(p,q)$, and that $L(p,q)$ is the *only* line that contains both p and q. We take the *line segment determined by p and q*, in contrast, to be the subset

$$LS(p,q) = \{p + a\,\overrightarrow{pq}\colon\ 0 \le a \le 1\}.$$

Again we have symmetry: $LS(q,p) = LS(p,q)$. Three points p, q, r are said to be *collinear*, of course, if there is a line to which they all belong. If they are distinct, this is equivalent to the requirement that $L(p,q) = L(q,r) = L(r,p)$. (If the points are not distinct, they are automatically collinear.)

Proposition 2.2.3. *Let p, q be distinct points in A, and let o be any point in A whatsoever (not necessarily distinct from p and q). Then, for every point r on $L(p,q)$, there is a unique number a such that $\overrightarrow{or} = a\,\overrightarrow{op} + (1-a)\,\overrightarrow{oq}$. Conversely, for every number a, the right side expression defines a point on $L(p,q)$.*

Proof. Let r be a point on $L(p,q)$. We can express it in the form $r = p + b\,\overrightarrow{pq}$ for some number b. Hence

$$\overrightarrow{or} = \overrightarrow{op} + \overrightarrow{pr} = \overrightarrow{op} + b\,\overrightarrow{pq} = \overrightarrow{op} + b\,(-\overrightarrow{op} + \overrightarrow{oq}) = (1-b)\,\overrightarrow{op} + b\,\overrightarrow{oq}.$$

So $\overrightarrow{or}$ assumes the desired form iff $a = (1-b)$. We can reverse the argument for the converse assertion. □

Problem 2.2.3. *Let p, q, r, s be any four (distinct) points in A. Show that the following conditions are equivalent.*

(i) $\overrightarrow{pr} = \overrightarrow{sq}$

(ii) $\overrightarrow{sp} = \overrightarrow{qr}$

(iii) *The midpoints of the line segments $LS(p,q)$ and $LS(s,r)$ coincide, i.e., $p + \frac{1}{2}\,\overrightarrow{pq} = s + \frac{1}{2}\,\overrightarrow{sr}$.*

Problem 2.2.4. *Let $p_1, ..., p_n$ ($n \ge 1$) be distinct points in A. Show that there is a point o in A such that $\overrightarrow{op}_1 + ... + \overrightarrow{op}_n = \mathbf{0}$. (If particles are present at the points $p_1, ..., p_n$, and all have the same mass, then o is the "center of mass" of the n particle system. Hint: Let q be any point at all, and take*

$$o = q + \frac{1}{n}(\overrightarrow{qp}_1 + ... + \overrightarrow{qp}_n).)$$

It is not our purpose to develop affine geometry systematically. But we will present one classic result, Desargues' theorem. It provides a nice example of the use of our algebraic methods. First we need a simple lemma.

Proposition 2.2.4. *(Collinearity Criterion) Let p, q, r be distinct points in A. They are collinear iff given any point o (a choice of "origin"), there exist numbers a, b, c, not all 0, such that $a + b + c = 0$ and $a\,\overrightarrow{op} + b\,\overrightarrow{oq} + c\,\overrightarrow{or} = \mathbf{0}$.*

Proof. Assume first the points are collinear. Let o be any point. Since r lies on $L(p, q)$, it follows from proposition 2.2.3 that there is a number k such that $\overrightarrow{or} = k\,\overrightarrow{op} + (1 - k)\,\overrightarrow{oq}$. The desired conditions will be satisfied if we take $a = k$, $b = (1 - k)$, and $c = -1$. The argument can also be reversed. Let o be any point and assume there exist numbers a, b, c, not all 0, satisfying the given equations. Without loss of generality (interchanging the roles of p, q, r, if necessary) we may assume that $c \neq 0$. If we take $k = -\dfrac{a}{c}$, then $(1-k) = -\dfrac{b}{c}$, and $\overrightarrow{or} = k\,\overrightarrow{op} + (1 - k)\,\overrightarrow{oq}$. So, by proposition 2.2.3 again, r belongs to $L(p, q)$. □

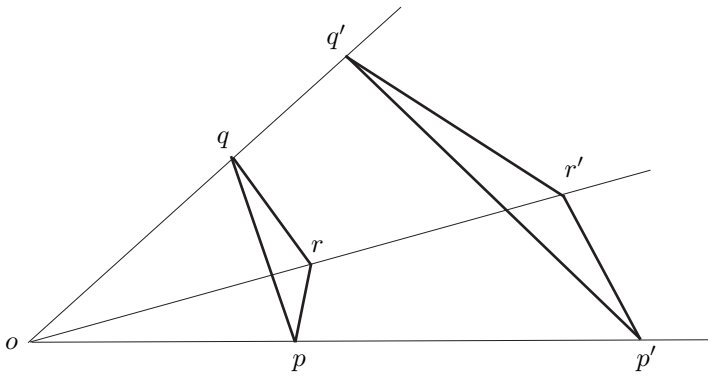


Figure 2.2.1: Triangles Perspective from a Point

A *triangle*, for us, is just a set of three (distinct) non-collinear points. Desargue's theorem deals with the "perspective" properties of triangles in affine spaces of dimension at least 2. (Assume for the moment that our background affine space satisfies this condition.) Let T and T' be disjoint triangles satisfying two conditions: (i) no point of one is equal to any point of the other, and (ii) no pair of points in one triangle determines the same line as any pair of points in the other. We say they are *perspective from a point o* if we can label their points so that $T = \{p, q, r\}$, $T' = \{p', q', r'\}$, and the lines $L(p, p')$, $L(q, q')$, and $L(r, r')$ all contain the point o. (See figure 2.2.1.) We say they are *perspective from a line L* if we can label them so the lines determined by corresponding sides intersect, and the intersection points $L(p, q) \cap L(p', q')$, $L(q, r) \cap L(q', r')$, and $L(p, r) \cap L(p', r')$ are all on L. (See figure 2.2.2.)

Proposition 2.2.5. *(Desargues' Theorem) Consider any two triangles satisfying conditions (i) and (ii). Assume they are perspective from a point o, and (with labels as above) the lines determined by their corresponding sides intersect in points*

$$\begin{aligned} x &= L(p,q) \cap L(p',q') \\ y &= L(q,r) \cap L(q',r') \\ z &= L(p,r) \cap L(p',r'). \end{aligned}$$

Then x, y, and z are collinear (and so the triangles are perspective from a line).

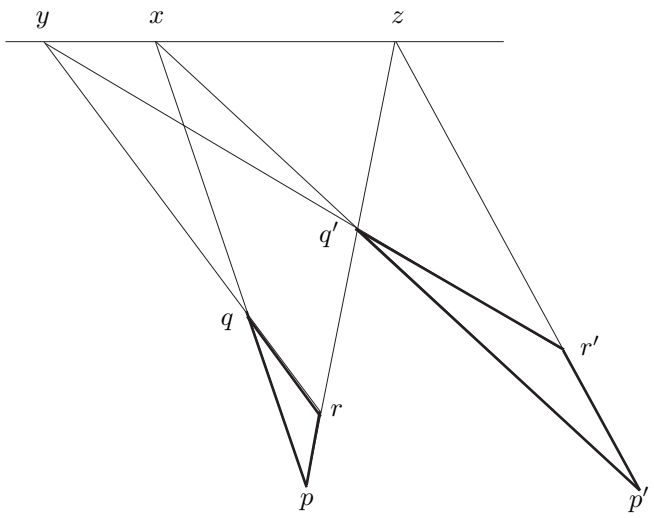


Figure 2.2.2: Triangles Perspective from a Line

Proof. (Roe [10]) We are assuming that the triples $\{o,p,p'\}$, $\{o,q,q'\}$, and $\{o,r,r'\}$ are all collinear. So there are numbers a, b, c such that $\overrightarrow{op}' = a\,\overrightarrow{op}$, $\overrightarrow{oq}' = b\,\overrightarrow{oq}$, and $\overrightarrow{or}' = c\,\overrightarrow{or}$. (Well, not quite. We will only be able to find the numbers if o is distinct from p, q and r. But if it is equal to one of them, then it must be distinct from p', q' and r'. So in that case we can run the argument with the roles of p, q, r and p', q', r' interchanged.) The numbers a, b, c must all be different from 1 (since $p \neq p'$, etc.).

Now since x lies on both $L(p,q)$ and $L(p',q')$, it follows from proposition 2.2.3 that there are numbers d and f such that

$$d\,\overrightarrow{op} + (1-d)\,\overrightarrow{oq} = \overrightarrow{ox} = f\,\overrightarrow{op}' + (1-f)\,\overrightarrow{oq}'.$$

Hence (substituting $a\,\overrightarrow{op}$ for $\overrightarrow{op}'$ and $b\,\overrightarrow{oq}$ for $\overrightarrow{oq}'$ in the expression on the right),

$$(d - af)\,\overrightarrow{op} + (1 - d - b + bf)\,\overrightarrow{oq} = \mathbf{0}.$$

But $\overrightarrow{op}$ and $\overrightarrow{oq}$ are linearly independent. (They could only be proportional if the points o, p, p', q, q' were collinear, violating our assumption that the lines $L(p,q)$

and $L(p', q')$ are distinct.) So $d = af$ and $(1 - d) = b(1 - f)$. Now it cannot be the case that $b = a$, since otherwise it would follow that $a = 1$, contradicting our remark above. So we can solve these equations for d in terms of a and b, and obtain

$$d = \frac{a(1-b)}{a-b}$$

and, hence,

$$\overrightarrow{ox} = \left(\frac{a(1-b)}{a-b}\right)\overrightarrow{op} - \left(\frac{b(1-a)}{a-b}\right)\overrightarrow{oq}.$$

Similarly, we have $b \neq c, a \neq c$,

$$\overrightarrow{oy} = \left(\frac{b(1-c)}{b-c}\right)\overrightarrow{oq} - \left(\frac{c(1-b)}{b-c}\right)\overrightarrow{or}.$$

and

$$\overrightarrow{oz} = \left(\frac{a(1-c)}{a-c}\right)\overrightarrow{op} - \left(\frac{c(1-a)}{a-c}\right)\overrightarrow{or}.$$

Therefore,

$$(a-b)(1-c)\,\overrightarrow{ox} \;+\; (b-c)(1-a)\,\overrightarrow{oy} \;-\; (a-c)(1-b)\,\overrightarrow{oz} \;=\; \mathbf{0}.$$

But we also have the simple algebraic identity

$$(a-b)(1-c) \;+\; (b-c)(1-a) \;-\; (a-c)(1-b) \;=\; 0.$$

And the factors $(a-b), (b-c), (a-c), (1-a), (1-b), (1-c)$ are all non-zero. So it follows by the collinearity criterion (proposition 2.2.4) that the points x, y, z are collinear. □

Note that our formulation of Desargues' theorem does not (quite) assert that if two triangles (satisfying conditions (i) and (ii)) are perspective from a point, then they are perspective from a line. We have to *assume* that the lines determined by corresponding sides of the triangles intersect. (Otherwise we would not have points x, y, z that are, at least, candidates for being collinear.) But there is a more general version of Desargues' theorem within "projective geometry" in which this assumption is not necessary. In this more general version, the collinear intersection points that figure in the conclusion of the proposition can be points "at infinity". (Further discussion of Desargues' theorem can be found in almost any book on projective geometry.)

We know that, up to isomorphism, there is only one n-dimensional vector space (for any $n \geq 0$). (Recall problem 2.1.7.) This result extends easily to affine spaces. The only slight subtlety is in the way one characterizes an "isomorphism" between affine spaces.

Consider first the canonical examples. We get a 0-dimensional affine space if we take A to be a singleton set $\{p\}$, take $\mathbf{V}$ to be a trivial vector space whose only element is the zero vector $\mathbf{0}$, and take $+$ to be the map that associates with p and $\mathbf{0}$ the element p (i.e., $p + \mathbf{0} = p$). Note that AS1 and AS2 are satisfied.

For $n \geq 1$, we get an n-dimensional affine space if we take A to be the set $\mathbb{R}^n$, take $\mathbf{V}$ to be the vector space $\mathbb{R}^n$, and take $+$ to be the operation that associates with a point $p = (a_1, ..., a_n)$ and a vector $v = (b_1, ..., b_n)$ the point $p + v = (a_1 + b_1, ..., a_n + b_n)$. We refer to it as the "affine space $\mathbb{R}^n$".

Now let $(A, \mathbf{V}, +)$ and $(A', \mathbf{V}', +')$ be affine spaces (not necessarily finite dimensional). We say that a bijection $\varphi : A \to A'$ between their underlying point sets is an *(affine space) isomorphism* if there is a (vector space) isomorphism $\Phi : V \to V'$ satisfying the following condition.

(I1) For all p and q in A, $\overrightarrow{\varphi(p)\,\varphi(q)} = \Phi(\overrightarrow{pq})$.

Of course, the two spaces are said to be *isomorphic* if there exists an isomorphism mapping one onto the other. (See figure 2.2.3.)

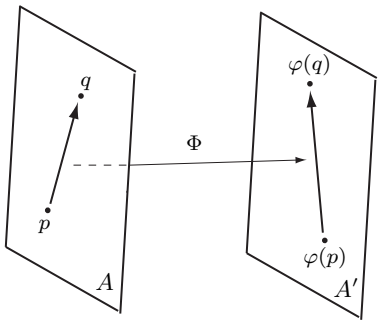


Figure 2.2.3: The affine space isomorphism φ takes points in A to points in A'. Its associated vector space isomorphism Φ takes vectors in V to vectors in V'.

This definition may seem less than perfectly clear. The idea is this. To qualify as an (affine space) isomorphism, the bijection φ must induce a map from V to V' (sending $\overrightarrow{pq}$ to $\overrightarrow{\varphi(p)\,\varphi(q)}$) that itself qualifies as a (vector space) isomorphism. Notice that (I1) can also be formulated, equivalently, as follows:

(I2) For all p in A and u in V, $\varphi(p + u) = \varphi(p) + \Phi(u)$.

For suppose that (I1) holds. Given p and u, let $q = p + u$. Then $u = \overrightarrow{pq}$ and we have, by (I1),

$$\overrightarrow{\varphi(p)\,\varphi(p+u)} = \overrightarrow{\varphi(p)\,\varphi(q)} = \Phi(\overrightarrow{pq}) = \Phi(u).$$

It follows that $\varphi(p + u) = \varphi(p) + \Phi(u)$. So we have (I2). Conversely, suppose (I2) holds. Then for all p and q,

$$\varphi(q) = \varphi(p + \overrightarrow{pq}) = \varphi(p) + \Phi(\overrightarrow{pq}).$$

So $\overrightarrow{\varphi(p)\,\varphi(q)} = \Phi(\overrightarrow{pq})$. This gives us (I1).

Associated with every affine space isomorphism $\varphi : A \to A'$ is a unique vector space isomorphism $\Phi : V \to V'$ satisfying condition (I1). (It is unique because if Φ_1 and Φ_2 both satisfy the condition, then (using the (I2) formulation)

$$\Phi_1(u) = \overrightarrow{\varphi(p)\,\varphi(p+u)} = \Phi_2(u)$$

for all p in A and all u in V. So $\Phi_1 = \Phi_2$.) The association is not invertible. One cannot recover φ from Φ. (There will always be infinitely many bijections $\varphi : A \to A'$ that, together with a given Φ, satisfy (I1).) But as the next proposition indicates, the recovery is possible once one adds as a side constraint the requirement that φ take some particular point o in A to some particular point o' in A'. (We can think of o and o' as "origins" for A and A'.) We will use the proposition repeatedly in what follows.

Proposition 2.2.6. *Let $(A, \boldsymbol{V}, +)$ and $(A', \boldsymbol{V}', +')$ be affine spaces. Further, let $\Phi : V \to V'$ be an isomorphism, let o and o' be points, respectively, in A and A', and let $\varphi : A \to A'$ be defined by setting $\varphi(p) = o' + \Phi(\overrightarrow{op})$ for all points p in A. Then*

(i) $\varphi(o) = o'$

(ii) $\overrightarrow{\varphi(p)\,\varphi(q)} = \Phi(\overrightarrow{pq})$ *for all p and q in A*

(iii) φ *is a bijection*

(iv) φ *is the only bijection between A and A' satisfying conditions (i) and (ii).*

Proof. Clause (i) holds because we have

$$\varphi(o) = o' + \Phi(\overrightarrow{oo}) = o' + \Phi(\mathbf{0}) = o' + \mathbf{0}' = o'.$$

(Here and in what follows we use proposition 2.2.1 and the fact that Φ is an isomorphism.) For (ii), notice that

$$\begin{aligned} \varphi(q) = o' + \Phi(\overrightarrow{oq}) &= o' + \Phi(\overrightarrow{op}) - \Phi(\overrightarrow{op}) + \Phi(\overrightarrow{oq}) = \varphi(p) + [-\Phi(\overrightarrow{op}) + \Phi(\overrightarrow{oq})] \\ &= \varphi(p) + \Phi(-\overrightarrow{op} + \overrightarrow{oq}) = \varphi(p) + \Phi(\overrightarrow{pq}) \end{aligned}$$

for all p and q in A. For (iii) we show, in order, that φ is injective and that it maps A onto A'. Assume first there are points p and q in A such that $\varphi(p) = \varphi(q)$. Then $\Phi(\overrightarrow{pq}) = \overrightarrow{\varphi(p)\,\varphi(q)} = \overrightarrow{\varphi(p)\,\varphi(p)} = \mathbf{0}'$. Since Φ is injective, it follows that $\overrightarrow{pq} = \mathbf{0}$. Therefore, $p = q$. Thus, φ is injective. Next, let p' be any point in A'. Consider the point $p = o + \Phi^{-1}(\overrightarrow{o'p'})$. Clearly, $\overrightarrow{op} = \Phi^{-1}(\overrightarrow{o'p'})$. Hence, $\overrightarrow{o'p'} = \Phi(\overrightarrow{op})$ and, therefore, $p' = o' + \Phi(\overrightarrow{op}) = \varphi(p)$. Since the arbitrary point p' has a preimage under φ, we see that φ maps A onto A'. So φ is a bijection as claimed and we have (iii). Finally, for (iv), assume $\psi : A \to A'$ is a bijection such that $\psi(o) = o'$ and $\overrightarrow{\psi(p)\,\psi(q)} = \Phi(\overrightarrow{pq})$ for all p and q in A. Then it follows, in particular, that

$$\overrightarrow{o'\,\psi(p)} = \overrightarrow{\psi(o)\,\psi(p)} = \Phi(\overrightarrow{op})$$

for all p in A. Hence, $\psi(p) = o' + \Phi(\overrightarrow{op}) = \varphi(p)$ for all p in A. So $\psi = \varphi$, and we have (iv). □

Proposition 2.2.7. *Finite dimensional affine spaces are isomorphic iff they have the same dimension.*

Proof. Let $\mathbf{A} = (A, \mathbf{V}, +)$ and $\mathbf{A}' = (A', \mathbf{V}', +')$ be finite dimensional affine spaces. Assume first that there exists an isomorphism $\varphi : A \to A'$ with corresponding map $\Phi : V \to V'$. Since Φ is a (vector space) isomorphism, $\mathbf{V}$ and $\mathbf{V}'$ have the same dimension. (Recall problem 2.1.7.) So, by the way we have characterized the dimension of affine spaces, $\mathbf{A}$ and $\mathbf{A}'$ have the same dimension.

Conversely, assume $\mathbf{V}$ and $\mathbf{V}'$ have the same dimension. Then, by problem 2.1.7 again, there exists a (vector space) isomorphism $\Phi : V \to V'$. Let o be a point in A, let o' be a point in A', and let $\varphi : A \to A$ be defined by setting $\varphi(p) = o' + \Phi(\overrightarrow{op})$ for all p. We know from proposition 2.2.6 that φ is an isomorphism with associated map Φ. So $\mathbf{A}$ and $\mathbf{A}'$ are isomorphic. □

Problem 2.2.5. *Let $(V, \boldsymbol{A}, +)$ be a two-dimensional affine space. Let $\{p_1, q_1, r_1\}$ and $\{p_2, q_2, r_2\}$ be two sets of non-collinear points in A. Show that there is a unique isomorphism $\varphi : A \to A$ such that $\varphi(p_1) = p_2$, $\varphi(q_1) = q_2$, and $\varphi(r_1) = r_2$. Hint: Use proposition 2.2.6 and the fact that a vector space isomorphism is uniquely determined by its action on the elements of any basis.*

2.3 Metric Affine Spaces

A (generalized) *inner-product* on a vector space $(V, +, \mathbf{0}, \cdot)$ is a map $\langle\ ,\ \rangle$ from $V \times V$ to $\mathbb{R}$ satisfying the following conditions:

(IP1) For all u, v in V, $\langle u, v\rangle = \langle v, u\rangle$.

(IP2) For all u, v, w in V, $\langle u, v + w\rangle = \langle u, v\rangle + \langle u, w\rangle$.

(IP3) For all r in $\mathbb{R}$ and all u, v in V, $\langle u, r\,v\rangle = r\langle u, v\rangle$.

(IP4) For all non-zero u in V, there is a v in V such that $\langle u, v\rangle \neq 0$.

(We say "generalized" because we do not assume that $\langle u, u\rangle > 0$ for all $u \neq \mathbf{0}$.)

In what follows, let $(V, +, \mathbf{0}, \cdot)$ be a finite dimensional vector space, and let $\langle\ ,\ \rangle$ be a generalized inner product on V.

Problem 2.3.1. *(Polarization Identity) Prove that for all vectors v and w in V,*

$$\langle v, w\rangle = \tfrac{1}{2}(\langle v, v\rangle + \langle w, w\rangle - \langle v - w, v - w\rangle).$$

We say that two vectors u and v are *orthogonal* (written $u \perp v$) if $\langle u, v\rangle = 0$. By extension, we say u is *orthogonal* to a subspace W (written $u \perp W$) if u is orthogonal to every vector in W. So, for example, the zero vector $\mathbf{0}$ is orthogonal to V. (By (IP3), $\langle v, \mathbf{0}\rangle = \langle v, 0 \cdot \mathbf{0}\rangle = 0\,\langle v, \mathbf{0}\rangle = 0$, for all v in V.) Moreover, by (IP4), $\mathbf{0}$ is the only vector with this property. Finally, we say that W is *orthogonal* to another subspace W' (written $W \perp W'$) if every vector in W is orthogonal to W'.

Given a vector u, the set $u^\perp$ consisting of all vectors orthogonal to u is a subspace. More generally, the set $W^\perp$ of all vectors orthogonal to a subspace W is a subspace. We call these the *orthogonal or perp(endicular) subspaces* of u and W.

Proposition 2.3.1. *(Projection Theorem) Assume V has dimension $n \geq 1$, and assume u is a vector in V such that $\langle u, u\rangle \neq 0$. Then all the following hold.*

(i) *Every vector v in V has a unique decomposition of the form $v = a\,u + w$, where $w \in u^{\perp}$.*

(ii) *If S is a basis for $u^{\perp}$, $S \cup \{u\}$ is a basis for V.*

(iii) *$u^{\perp}$ has dimension $(n-1)$.*

Proof. Since $\langle u, u\rangle \neq 0$, u cannot be the zero-vector. We will use this fact repeatedly.

(i) Let v be any vector in V. Since $\langle u, u\rangle \neq 0$, there is a real number a such that $a\langle u, u\rangle = \langle u, v\rangle$. Hence $\langle u, v - a\,u\rangle = 0$. Take $w = v - a\,u$. For uniqueness, note that if $w_1 = v - a_1 u$ and $w_2 = v - a_2 u$ are both orthogonal to u, then so is the difference vector

$$w_1 - w_2 = (v - a_1 u) - (v - a_2 u) = (a_2 - a_1)u.$$

So $0 = \langle u, (a_2 - a_1)u\rangle = (a_2 - a_1)\langle u, u\rangle$. Since $\langle u, u\rangle \neq 0$, it follows that $a_2 = a_1$, and therefore $w_1 = w_2$.

(ii) Let S be a basis for $u^{\perp}$, and let $S' = S \cup \{u\}$. Suppose first that the dimension of $u^{\perp}$ is 0. In this case, $u^{\perp} = \{\mathbf{0}\}$, S is the empty set, and $S' = \{u\}$. So to show that S' is a basis for V, we must establish that every vector v in V can be expressed in the form $v = a \cdot u$, for some real number a. But this follows from part (i). For we can certainly express every such v in the form $v = a \cdot u + w$, where $w \in u^{\perp}$. And if $w \in u^{\perp}$, then $w = \mathbf{0}$. So we have $v = a \cdot u + \mathbf{0} = a \cdot u$.

Suppose next that the dimension of $u^{\perp}$ is at least 1, and S is non-empty. We claim first that S' is linearly independent. To see this, suppose to the contrary there exist vectors $w_1, ..., w_k$ in S and numbers $a, b_1, ..., b_k\,(k \geq 1)$, not all 0, such that

$$a\,u + b_1 w_1 + ... + b_k w_k = \mathbf{0}.$$

Then, taking the inner product of u with each side,

$$a\,\langle u, u\rangle + \langle u,\, b_1 w_1 + ... + b_k w_k\rangle = 0.$$

Now $b_1 w_1 + ... + b_k w_k$ is orthogonal to u (since it is a linear combination of vectors in $u^{\perp}$). So the second term in this sum must be 0. Hence, $a\langle u, u\rangle = 0$. But $\langle u, u\rangle \neq 0$. So $a = 0$. It follows that $b_1 w_1 + ... + b_k w_k = \mathbf{0}$, and not all the numbers $b_1, ...b_k$ are 0. But this is impossible. S is linearly independent (since it is a basis for $u^{\perp}$). Thus, as claimed, S' is linearly independent. Next we claim that S' spans V. This follows from part (i). Every vector v in V can be expressed in the form $v = a\,u + w$ where $w \in u^{\perp}$. And since S is a basis for $u^{\perp}$, w can be expressed as a linear combination $w = b_1 w_1 + ... + b_k w_k$ of vectors in S $(k \geq 1)$. Hence v itself can be expressed as a linear combination of the form $v = au + b_1 w_1 + ... + b_k w_k$. Thus, S' spans V as claimed, and we may conclude that S' is a basis of V.

(iii) Let S be a basis for $u^{\perp}$. Then, by (ii), $S \cup \{u\}$ is a basis for V, and must contain n vectors. But u itself cannot be a member of S. (All vectors in

S are orthogonal to u, and we have assumed that u is not orthogonal to itself.) So S contains $(n-1)$ vectors. Since S is a basis for $u^\perp$, it follows that $u^\perp$ has dimension $(n-1)$. □

The projection theorem has an important corollary. Let S be a basis for V. We say that it is *orthonormal* (with respect to the inner product $\langle\ ,\ \rangle$) if, for all u, v in S,

(i) $u \neq v \Rightarrow \langle u, v\rangle = 0$

(ii) $\langle u, u\rangle^2 = 1$.

(It is important that we are not insisting that $\langle u, u\rangle = 1$ for all u in S. We are allowing for the possibility that, for at least some u, the inner product is -1.)

Proposition 2.3.2. *V has an orthonormal basis.*

Proof. The empty set qualifies as an orthonormal basis for any vector space of dimension 0. So we may assume that $\dim(V) \geq 1$. We claim, first, that there exists a vector u in V such that $\langle u, u\rangle \neq 0$. Suppose not. Then by the polarization identity (problem 2.3.1), it follows that $\langle v, w\rangle = 0$ for all v and w. But this is impossible. Since $\dim(V) \geq 1$, there exists a non-zero vector v in V, and so, by (IP4), there is a vector w in V (corresponding to v) such that $\langle v, w\rangle \neq 0$. Thus, as claimed, there is a vector u in V such that $\langle u, u\rangle \neq 0$.

Now we proceed by induction on $n = \dim(V)$. Assume, first, that $n = 1$, and consider the vector

$$u' = \frac{u}{|\langle u, u\rangle|^{\frac{1}{2}}}.$$

Clearly, $\langle u', u'\rangle$ is either 1 or -1 (depending on whether $\langle u, u\rangle$ is positive or negative). Either way, $\{u'\}$ qualifies as an orthonormal basis for V.

Next, assume that $n \geq 2$, and that the proposition holds for vector spaces of dimension $(n-1)$. By the projection theorem, $u^\perp$ has dimension $(n-1)$. We claim that the induction hypothesis is, therefore, applicable to $u^\perp$ (together with the restriction of $\langle\ ,\ \rangle$ to $u^\perp$). But there is something here that must be checked. We need to know that the restriction of $\langle\ ,\ \rangle$ is a generalized inner product, i.e., satisfies conditions (IP1) – (IP3) and (IP4). The first three are automatically inherited under restriction. What we need to check is that the fourth does so as well, i.e., that for all nonzero vectors v in $u^\perp$, there is a w in $u^\perp$ (not just in V) such that $\langle v, w\rangle \neq 0$. But it is not hard to do so. Assume to the contrary that v is a nonzero vector in $u^\perp$ that is orthogonal to all vectors in $u^\perp$. Then v must be orthogonal to all vectors v' in V. (Why? Consider any such vector v'. It can be expressed in the form $v' = a\,u + w$ where $w \in u^\perp$. By our assumption, v is orthogonal to w (since w is in $u^\perp$). And it is also orthogonal to u (since v is in $u^\perp$). So $\langle v, v'\rangle = \langle v, a\,u + w\rangle = a\langle v, u\rangle + \langle v, w\rangle = 0$.) But this is impossible. By (IP4) again, there is no non-zero vector orthogonal to all vectors in V.

Thus, as claimed, our induction hypothesis is applicable to the $(n-1)$ dimensional space $u^\perp$ (together with the induced inner product on it). So it must

be the case that $u^\perp$ has an orthonormal basis S. And, therefore, by the projection theorem, $S' = S \cup \{u'\}$ qualifies as an orthonormal basis for V. (Here u' is, again, just the normalized version of u considered above.) Thus the proposition holds in the case where the dimension of V is n. It follows, by the principle of induction, that it holds no matter what the dimension of V. □

Given an orthonormal basis S of V (with respect to the generalized inner product $\langle\ ,\ \rangle$), there is some number of vectors u in S (possibly 0) such that $\langle u, u\rangle = 1$, and some number of vectors u in S (possibly 0) such that $\langle u, u\rangle = -1$. We next show that these two numbers are the same in all bases. We do so by giving the two numbers an invariant, i.e., basis independent, characterization.

Let us say that a subspace W of V is (with respect to $\langle\ ,\ \rangle$) *positive definite* if, for all w in W, $w \neq \mathbf{0} \Rightarrow \langle w, w\rangle > 0$; *negative definite* if, for all w in W, $w \neq \mathbf{0} \Rightarrow \langle w, w\rangle < 0$; and *definite* if it is one or the other.

Problem 2.3.2. *Let W be a subspace of V. Show that the following conditions are equivalent.*

(i) *W is definite.*

(ii) *There does not exist a non-zero vector w in W with $\langle w, w\rangle = 0$.*

(Hint: To show that (ii) implies (i), assume that W is neither positive definite nor negative definite. Then there exist non-zero vectors u and v in W such that $\langle u, u\rangle \leq 0$ and $\langle v, v\rangle \geq 0$. Consider the function $f : [0, 1] \to \mathbb{R}$ defined by $f(x) = \langle x\, u + (1 - x)v,\ x\, u + (1 - x)v\rangle$. It is continuous. (Why?) So ... .)

The *signature* of $\langle\ ,\ \rangle$ is a pair of non-negative integers (n^+, n^-) where

$n^+ =$ the maximal possible dimension for a positive definite subspace
$n^- =$ the maximal possible dimension for a negative definite subspace.

(This definition make sense. Positive and negative definite subspaces are not, in general, unique. But among all such, there must be ones with maximal dimension (since V itself is of finite dimension).)

Proposition 2.3.3. *Let $\langle\ ,\ \rangle$ have signature (n^+, n^-), and let S be an orthonormal basis for V. Then there are n^+ vectors u in S such that $\langle u, u\rangle = 1$, and n^- vectors u in S such that $\langle u, u\rangle = -1$. (And, therefore, $n^+ + n^- = dim(V)$.)*

Proof. We give the proof for n^+. (The proof for n^- is the same except for obvious modifications.) Let $n = \dim(V)$, and let m be the number of vectors u in S such that $\langle u, u\rangle > 0$. We must show that $m = n^+$. If $n = 0$, the assertion is trivial. (For then S is the empty set, and $n^+ = 0 = m$.) So we may assume that $n \geq 1$. Let S be $\{u_1, ..., u_n\}$. We may also assume that $m \geq 1$. For suppose $m = 0$, i.e., $\langle u_i, u_i\rangle = -1$ for all i. Then given any non-zero vector u in U, we can express it as a linear combination $u = a_1u_1 + ... + a_nu_n$ with at least one non-zero coefficient, and it follows that

$$\begin{aligned}\langle u, u\rangle &= \langle a_1u_1 + ... + a_nu_n, a_1u_1 + ... + a_nu_n\rangle \\ &= \sum_{i,j=1}^{n} a_ia_j\langle u_i, u_j\rangle = \sum_{i=1}^{n} a_i^2\langle u_i, u_i\rangle = -a_1^2 - ... - a_n^2 < 0.\end{aligned}$$

Thus, if $m = 0$, there are no non-zero vectors u in V such that $\langle u, u\rangle > 0$ and, therefore, $n^+ = 0 = m$, as required.

So we may assume that $n \geq 1$, and $m \geq 1$. Reordering the elements in S if necessary, we may also assume that, for all i,

$$\begin{array}{rcrcl} 1 \leq i \leq m & \Longrightarrow & \langle u_i, u_i\rangle & = & +1 \\ m < i \leq n & \Longrightarrow & \langle u_i, u_i\rangle & = & -1. \end{array}$$

Let W be a positive definite subspace of V with dimension n^+, and let U be the subspace of V spanned by the set $\{u_i : 1 \leq i \leq m\}$. We claim that U is, itself, positive definite. (To see this, let u be a non-zero vector in U. We can express it as a linear combination $u = a_1u_1 + ... + a_mu_m$ with at least one non-zero coefficient and, therefore,

$$\begin{array}{rcl} \langle u, u\rangle & = & \langle a_1u_1 + ... + a_mu_m, a_1u_1 + ... + a_mu_m\rangle \\ & = & \displaystyle\sum_{i,j=1}^{m} a_ia_j\langle u_i, u_j\rangle = a_1^2 + ... + a_m^2 > 0.) \end{array}$$

It follows by the maximality of W that $m \leq n^+$.

Now we define a map $\varphi : W \to U$ as follows. Given a vector w in W, if it has expansion $w = b_1u_1 + ... + b_nu_n$, we set $\varphi(w) = b_1u_1 + ... + b_mu_m$. (We get from w to $\varphi(w)$ by truncating the expansion for w after the m^{th} term.) It is easily checked that φ is linear. Furthermore, we claim, φ is injective, or, equivalently, that $\ker(\varphi) = \{\mathbf{0}\}$. To see this, suppose $w = b_1u_1 + ... + b_nu_n$ and $\varphi(w) = b_1u_1 + ... + b_mu_m = \mathbf{0}$. Then $b_1 = ... = b_m = 0$, since $\{u_1, ..., u_m\}$ is linearly independent. (It is a subset of S and S is a basis for V.) So $w = b_{m+1}u_{m+1} + ... + b_nu_n$. Hence,

$$\begin{array}{rcl} \langle w, w\rangle & = & \langle b_{m+1}u_{m+1} + ... + b_nu_n, b_{m+1}u_{m+1} + ... + b_nu_n\rangle \\ & = & -(b_{m+1})^2 - ... - (b_n)^2 \leq 0. \end{array}$$

But w belongs to W, and W is positive definite. So it must be the case that $w = \mathbf{0}$. Thus, φ is injective, as claimed, and φ is an isomorphism between W and $\varphi[W]$. It follows that $n^+ = \dim(W) = \dim(\varphi[W]) \leq \dim(U) = m$. (The second equality follows from problem 2.1.7. The inequality follows from clause (v) of proposition 2.1.1 and the fact that $\varphi[W]$ is a subspace of U.) So it must be the case that $m = n^+$. □

Now we turn to metric affine geometry proper. We take a finite dimensional *metric affine space* to be a finite dimensional affine space $(A, \mathbf{V}, +)$ together with a generalized inner product $\langle\ ,\ \rangle$ on $\mathbf{V}$. We take the dimension of the space to be that of $\mathbf{V}$, and the signature to be that of $\langle\ ,\ \rangle$.

It is clear that, for every $n \geq 0$, and every pair of non-negative integers (n^+, n^-) with $n^+ + n^- = n$, there is an n dimensional metric affine space $\langle \mathbf{A}, \langle\ ,\ \rangle\rangle$ whose signature is (n^+, n^-). Indeed, if $n \geq 1$, it suffices to let $\mathbf{A}$ be

the affine space $\mathbb{R}^n$, and let $\langle\ ,\ \rangle$ be the generalized inner product that assigns to vectors $u = (a_1, ..., a_n)$ and $v = (b_1, ..., b_n)$ the number

$$\langle u, v\rangle = a_1b_1 + ... + a_{n^+}b_{n^+} - ... - a_nb_n.$$

(One gets a 0-dimensional metric affine space with signature $(0, 0)$ if one takes $\mathbf{A}$ to be the 0-dimensional affine space $(\{p\}, \{\mathbf{0}\}, +)$ discussed in section 2.2, and takes $\langle\ ,\ \rangle$ to be the trivial inner product that makes the assignment $\langle \mathbf{0}, \mathbf{0}\rangle = 0$.) These examples are, in an appropriate sense, the only ones. To make that sense precise, we need a definition.

Let $(\mathbf{A}, \langle\ ,\ \rangle)$ and $(\mathbf{A}', \langle\ ,\ \rangle')$ be metric affine spaces (with corresponding underlying vector spaces $\mathbf{V}$ and $\mathbf{V}'$). Recall that an (affine space) isomorphism between $\mathbf{A}$ and $\mathbf{A}'$ is a bijection $\varphi: A \to A'$ satisfying the requirement that there exist a (vector space) isomorphism $\Phi: V \to V'$ such that, for all p and q in A, $\overrightarrow{\varphi(p)\,\varphi(q)} = \Phi(\overrightarrow{pq})$. (As we observed in section 2.2, if one exists, it is certainly unique.) We say that φ is an *isometry* (between $(\mathbf{A}, \langle\ ,\ \rangle)$ and $(\mathbf{A}', \langle\ ,\ \rangle')$) if, in addition, $\langle u, v\rangle = \langle \Phi(u), \Phi(v)\rangle'$ for all vectors u and v in V. (So, to qualify as an isometry, φ must respect the linear structure of $\mathbf{V}$ and the metric structure of $\langle\ ,\ \rangle$.) $(\mathbf{A}, \langle\ ,\ \rangle)$ and $(\mathbf{A}', \langle\ ,\ \rangle')$ are said to be *isometric*, of course, if there exists an isometry mapping one onto the other.

Proposition 2.3.4. *Two finite dimensional metric affine spaces are isometric iff they have the same dimension and signature.*

Proof. Let $(\mathbf{A}, \langle\ ,\ \rangle)$ and $(\mathbf{A}', \langle\ ,\ \rangle')$ be finite dimensional metric affine spaces (with corresponding underlying vector spaces $\mathbf{V}$ and $\mathbf{V}'$). Assume there exists an isometry $\varphi : A \to A'$ with corresponding map $\Phi : V \to V'$. Since φ is an (affine space) isomorphism, it follows from proposition 2.2.7 that $\mathbf{A}$ and $\mathbf{A}'$ have the same dimension. Since Φ preserves inner products, it takes positive and negative subspaces in V onto subspaces of the same type in V'. So $\langle\ ,\ \rangle$ and $\langle\ ,\ \rangle'$ necessarily have the same signature.

Conversely, suppose that the two metric affine spaces have the same dimension $n \geq 0$ and same signature (n^+, n^-). It follows, we claim, that there exists an isomorphism $\Phi : V \to V'$ between their underlying vector spaces that preserves inner products, i.e., such that $\langle u, v\rangle = \langle \Phi(u), \Phi(v)\rangle'$ for all vectors u and v in V. If $n = 0$, the claim is trivial (since then the map Φ taking $\mathbf{0}$ in V to $\mathbf{0}'$ in V' qualifies as an inner product preserving isomorphism). If $n \geq 1$, we can generate Φ by considering orthonormal bases for V and V'. Suppose $\{u_1, ..., u_n\}$ and $\{u'_1, ..., u'_n\}$ are two such, and suppose the two are ordered so that, for all i,

$$\begin{array}{lll} 1 \leq i \leq n^+ & \Longrightarrow & \langle u_i, u_i\rangle = +1 = \langle u'_i, u'_i\rangle \\ n^+ < i \leq n & \Longrightarrow & \langle u_i, u_i\rangle = -1 = \langle u'_i, u'_i\rangle. \end{array}$$

(Existence is guaranteed by proposition 2.3.3.) We define $\Phi: V \to V'$ by setting

$$\Phi(a_1u_1 + ... + a_nu_n) = a_1u'_1 + ... + a_nu'_n$$

for all numbers $a_1, ..., a_n$. It follows easily that Φ is an isomorphism and preserves inner products. The latter holds since, for all vectors u and v in V, if $u = a_1u_1 + ... + a_nu_n$ and $v = b_1u_1 + ... + b_nu_n$, then

$$\begin{aligned}\langle \Phi(u), \Phi(v)\rangle' &= \langle a_1u'_1 + ... + a_nu'_n,\, b_1u'_1 + ... + b_nu'_n\rangle' \\ &= a_1b_1 + ... + a_{n^+}b_{n^+} - ... - a_nb_n = \langle u, v\rangle.\end{aligned}$$

To finish the argument, we make use of proposition 2.2.6. Let o be a point in A, let o' be a point in A', and let $\varphi: A \to A'$ be defined by setting $\varphi(p) = o' + \Phi(\overrightarrow{op})$ for all p in A. Then φ is an affine space isomorphism whose associated map Φ preserves inner products, i.e., φ is an isometry. □

Two signatures are of special interest to us: $(n, 0)$ and $(1, n-1)$. At least when $n \geq 2$, the first characterizes *Euclidean geometry*, and the second *Minkowskian geometry*. We consider the former in the next section, and the latter in section 3.1.

2.4 Euclidean Geometry

Let $\mathbf{V} = (V, +, \mathbf{0}, \cdot)$ be an n dimensional vector space ($n \geq 0$), and let $\langle\ ,\ \rangle$ be a generalized inner product on $\mathbf{V}$. Note that the following conditions are equivalent.

(E1) The signature of $\langle\ ,\ \rangle$ is $(n, 0)$.

(E2) V is positive definite, i.e., for all u in V, $u \neq \mathbf{0} \Rightarrow \langle u, u\rangle > 0$.

To see this, suppose first that (E2) holds. Then there exists an n-dimensional subspace of V (namely V itself) that is positive definite. So n is the maximal possible dimension for a positive definite subspace of V. (No subspace, positive definite or not, has dimension greater than n.) Thus, $n^+ = n$ and $n^- = (n - n^+) = 0$. So (E1) holds. Conversely, suppose (E1) holds. Then there is an n-dimensional subspace of V that is positive definite. But the only n-dimensional subspace of V is V itself. (This follows, for example, by clause (iii) of proposition 2.1.1.) So (E2) holds.

Sometimes the term "inner product" is reserved for maps $V \times V$ to $\mathbb{R}$ satisfying conditions (IP1) – (IP3), and (E2). (In this case, there is no need to add (IP4) since it follows from (E2).) We have used the expression "generalized inner product" to head off possible confusion. Whatever terminology one employs, condition (E2) greatly simplifies the situation. Instead of dealing with three types of non-zero vector – classified according to whether the inner product of the vector with itself is negative, zero, or positive – we have only one type. $\langle u, u\rangle$ is positive for all nonzero vectors u.

When (E2) holds, we say that $\langle\ ,\ \rangle$, itself, is *positive definite*. We will assume it is so in this section.

Given a vector u in V, we take its *norm* $\|u\|$ to be the number $\langle u, u\rangle^{\frac{1}{2}}$. (Of course, this norm would not be well defined (in general) if we were not assuming

that $\langle\ ,\ \rangle$ is positive definite.) It follows that

$$\begin{aligned} \|u\| &\geq 0 \\ \|u\| &= 0 \iff u = \mathbf{0} \\ \|a\,u\| &= |a|\,\|u\| \end{aligned}$$

for all real numbers a. If we think of u as an arrow, we can think of $\|u\|$ as its length.

The polarization identity (problem 2.3.1) can (in the present context) be cast in the form:

$$\langle v, w\rangle = \frac{1}{2}(\|v\|^2 + \|w\|^2 - \|v - w\|^2).$$

This tells us that we can reverse the order of definition, and recover the inner product $\langle\ ,\ \rangle$ from its associated norm. Another useful identity is formulated in the following problem. It is called the "parallelogram identity". (Can you explain where the name comes from?)

Problem 2.4.1. *Prove that for all vectors u and v in V,*

$$\|u + v\|^2 + \|u - v\|^2 = 2(\|u\|^2 + \|v\|^2).$$

Proposition 2.4.1. *(Schwarz Inequality) For all vectors u and v in V,*

$$|\langle u, v\rangle| \leq \|u\|\,\|v\|,$$

with equality iff one of the two vectors is a multiple of the other.

Proof. Note first that if $u = \mathbf{0}$, then the inequality holds (since $|\langle u, v\rangle| = 0 = \|u\|\,\|v\|$), and so does the biconditional (since in this case we have $u = 0\,v$). So we may assume that $u \neq \mathbf{0}$ and, hence, that $\langle u, u\rangle > 0$. Let $w = v - \dfrac{\langle u, v\rangle}{\langle u, u\rangle}\,u$. Then

$$\begin{aligned} \langle w, w\rangle &= \langle v, v\rangle - 2\langle v, u\rangle\frac{\langle u, v\rangle}{\langle u, u\rangle} + \langle u, u\rangle\left(\frac{\langle u, v\rangle}{\langle u, u\rangle}\right)^2 \\ &= \langle v, v\rangle - 2\frac{\langle u, v\rangle^2}{\langle u, u\rangle} + \frac{\langle u, v\rangle^2}{\langle u, u\rangle} \\ &= \langle v, v\rangle - \frac{\langle u, v\rangle^2}{\langle u, u\rangle}. \end{aligned}$$

Now $\langle w, w\rangle \geq 0$ (since this inequality holds for all vectors, whether zero or nonzero). So

$$\langle v, v\rangle - \frac{\langle u, v\rangle^2}{\langle u, u\rangle} \geq 0,$$

or, equivalently,

$$\langle u, v\rangle^2 \leq \langle u, u\rangle\langle v, v\rangle.$$

Taking positive square roots of both sides yields the desired inequality.

Now assume that $|\langle u, v\rangle| = \|u\|\,\|v\|$, i.e., $\langle u, v\rangle^2 = \langle u, u\rangle\langle v, v\rangle$. Then (by the string of equations above), $\langle w, w\rangle = 0$. So, by (E2), $w = \mathbf{0}$. Hence,

$$v = \frac{\langle u, v\rangle}{\langle u, u\rangle}\, u.$$

Thus the two vectors are proportional. Conversely, assume that one of the two vectors is a multiple of the other. Since u is non-zero, we have $v = a\,u$ for some a. Then

$$\begin{aligned} \|u\|^2\|v\|^2 &= \langle u, u\rangle\langle v, v\rangle = \langle u, u\rangle\langle a\,u, au\rangle = a\langle u, u\rangle\langle u, a\,u\rangle \\ &= \langle u, a\,u\rangle\langle u, a\,u\rangle = \langle u, v\rangle^2. \end{aligned}$$

So, $|\langle u, v\rangle| = \|u\|\,\|v\|$. □

Problem 2.4.2. *Use the following hint to give a second proof of proposition 2.4.1 Given vectors u and v in V, consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by*

$$f(x) = \langle xu + v, xu + v\rangle = \langle u, u\rangle x^2 + 2\langle u, v\rangle x + \langle v, v\rangle.$$

It follows from (E2) that $f(x) \geq 0$ for all x, and $f(x) = 0$ iff $xu + v = \mathbf{0}$. These conditions impose constraints on the coefficients of f.

Proposition 2.4.2. *(Triangle Inequality) For all vectors u and v in V,*

$$\|u + v\| \leq \|u\| + \|v\|,$$

with equality iff one of the two vectors is a non-negative multiple of the other.

Proof. By the Schwarz inequality, we have

$$\begin{aligned} \|u + v\|^2 &= \langle u + v, u + v\rangle = \langle u, u\rangle + 2\langle u, v\rangle + \langle v, v\rangle \\ &= \|u\|^2 + 2\langle u, v\rangle + \|v\|^2 \leq \|u\|^2 + 2|\langle u, v\rangle| + \|v\|^2 \\ &\leq \|u\|^2 + 2\|u\|\,\|v\| + \|v\|^2 = (\|u\| + \|v\|)^2. \end{aligned}$$

So $\|u + v\| \leq \|u\| + \|v\|$. Furthermore, equality will hold iff all the sums in the sequence are equal. Thus, it will hold iff $|\langle u, v\rangle| = \langle u, v\rangle$ and $|\langle u, v\rangle| = \|u\|\,\|v\|$. But, by the second half of proposition 2.4.1, these two conditions will both hold iff one of the two vectors is a multiple of the other and the proportionality factor is nonnegative. □

Now we come to Euclidean geometry proper. We take an n-dimensional *Euclidean space* to be a metric affine space with dimension n and signature $(n, 0)$. As noted above, "up to isometry", there is only one such.

In what follows, let $(\mathbf{A}, \langle\ ,\ \rangle)$ be an n-dimensional Euclidean space with $n \geq 2$.

All the usual notions and theorems of Euclidean geometry can be recovered within our framework. First, $\langle\ ,\ \rangle$ induces a distance function on A. Given points p and q in A, we take the distance between them to be

$$d(p, q) = \|\overrightarrow{pq}\| = \langle \overrightarrow{pq}, \overrightarrow{pq}\rangle^{\frac{1}{2}}.$$

It follows easily that, for all points p, q, and r in A,

$$\begin{aligned} d(p,p) &= 0 \\ d(p,q) &= d(q,p) \\ d(p,r) &\leq d(p,q) + d(q,r). \end{aligned}$$

Second, we have a notion of orthogonality. Given points p, q, r, s, with $p \neq q$ and $r \neq s$, we say that the lines $L(p,q)$ and $L(r,s)$ (and the line segments $LS(p,q)$ and $LS(r,s)$) are *orthogonal* if the vectors $\overrightarrow{pq}$ and $\overrightarrow{rs}$ are so.

Third, we have a notion of angular measure. Given points o, p, q, with o distinct from p and q, the vectors $\overrightarrow{op}$ and $\overrightarrow{oq}$ form a (possibly degenerate) angle. We take its angular measure to be the number θ such that

$$\cos\theta = \frac{\langle \overrightarrow{op}, \overrightarrow{oq} \rangle}{\|\overrightarrow{op}\| \, \|\overrightarrow{oq}\|}$$

and $0 \leq \theta \leq \pi$. (We write this θ as $\measuredangle(p,o,q)$.) Notice that the definition is well posed since, by the Schwarz inequality, the expression on the right side is between -1 and 1 (and for any number x in that interval, there is a unique number θ in the interval $[0,\pi]$ such that $\cos\theta = x$). Notice too that $\measuredangle(p,o,q) = \measuredangle(q,o,p)$. Finally, notice that the number $\measuredangle(p,o,q)$ does not depend on the length of the segments $LS(o,p)$ and $LS(o,q)$, but only on their relative orientation. More precisely, if p' and q' are such that $\overrightarrow{op'} = a\,\overrightarrow{op}$ and $\overrightarrow{oq'} = b\,\overrightarrow{oq}$, with $a, b > 0$, then $\measuredangle(p',o,q') = \measuredangle(p,o,q)$.

Problem 2.4.3. *(The measure of a straight angle is π.) Let p, q, r be (distinct) collinear points, and suppose that q is between p and r (i.e., $\overrightarrow{pq} = a\,\overrightarrow{pr}$ with $0 < a < 1$). Show that $\measuredangle(p,q,r) = \pi$.*

Problem 2.4.4. *(Law of Cosines) Let p, q, r be points, with q distinct from p and r. Show that*

$$\|\overrightarrow{pr}\|^2 = \|\overrightarrow{qp}\|^2 + \|\overrightarrow{qr}\|^2 - 2\|\overrightarrow{qp}\| \, \|\overrightarrow{qr}\| \cos\measuredangle(p,q,r).$$

(Hint: Use the polarization identity (problem 2.3.1).)

Problem 2.4.5. *(Right Angle in a Semicircle Theorem) Let p, q, r, o be (distinct) points such that (i) p, o, r are collinear, and (ii) $\|\overrightarrow{op}\| = \|\overrightarrow{oq}\| = \|\overrightarrow{or}\|$. (So q lies on a semicircle with diameter $LS(p,r)$ and center o.) Show that $\overrightarrow{qp} \perp \overrightarrow{qr}$, and so $\measuredangle(p,q,r) = \dfrac{\pi}{2}$. (Hint: First show that $\overrightarrow{op} = -\overrightarrow{or}$, $\overrightarrow{qp} = -\overrightarrow{oq} + \overrightarrow{op}$, and $\overrightarrow{qr} = -\overrightarrow{oq} - \overrightarrow{op}$. Then expand the inner product $\langle \overrightarrow{qp}, \overrightarrow{qr} \rangle$ using the latter two equalities.)*

Problem 2.4.6. *(Stewart's Theorem) Let p, q, r, s be points (not necessarily distinct) with s (collinear with and) between q and r. (So $\overrightarrow{qs} = a\,\overrightarrow{qr}$ for some $a \in [0,1]$.) Show that*

$$\|\overrightarrow{pq}\|^2 \|\overrightarrow{sr}\| + \|\overrightarrow{pr}\|^2 \|\overrightarrow{qs}\| - \|\overrightarrow{ps}\|^2 \|\overrightarrow{qr}\| = \|\overrightarrow{qr}\| \|\overrightarrow{qs}\| \|\overrightarrow{sr}\|.$$

The next proposition shows that our choice for the angular measure function (involving the inverse cosine function) satisfies an additivity condition that one would expect any natural notion of angular measure to satisfy. It turns out that it is the *only* candidate, up to a constant, that satisfies the additivity condition as well as certain other modest (invariance and continuity) conditions. This goes some way to explaining where our choice "comes from". We will not prove the uniqueness theorem here, but *will* prove a corresponding theorem for Minkowskian angular measure in section 3.3. And we will leave it as an exercise there (problem 3.3.1) to rework the proof so as to apply to the present (Euclidean) case.

Given co-planar points o, p, q, r, with o distinct from p, q, and r, we say the vector $\overrightarrow{oq}$ is *between* $\overrightarrow{op}$ and $\overrightarrow{or}$ if there exist $a, b \geq 0$ such that $\overrightarrow{oq} = a\,\overrightarrow{op} + b\,\overrightarrow{or}$. (This notion of "betweenness" is a bit delicate. It need not be the case (even when o, p, q, and r are co-planar) that one of the three vectors $\overrightarrow{op}, \overrightarrow{oq}, \overrightarrow{or}$ qualifies as being between the other two. It could be the case, for example, that the points are co-planar and the three angles $\measuredangle(p, o, q), \measuredangle(q, o, r)$, and $\measuredangle(r, o, p)$ are all equal.)

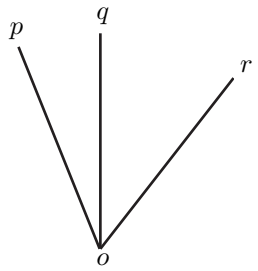


Figure 2.4.1: $\measuredangle(p, o, q) + \measuredangle(q, o, r) = \measuredangle(p, o, r)$

Proposition 2.4.3. *Let o, p, q, r be co-planar points with o distinct from p, q, and r. If $\overrightarrow{oq}$ is between $\overrightarrow{op}$ and $\overrightarrow{or}$, then $\measuredangle(p, o, q) + \measuredangle(q, o, r) = \measuredangle(p, o, r)$. (See figure 2.4.1.)*

Proof. Since the measure of an angle does not depend on the length of its "sides", we may assume that $\|\overrightarrow{op}\| = \|\overrightarrow{oq}\| = \|\overrightarrow{or}\| = 1$. Let $\theta_1 = \measuredangle(p, o, q)$, $\theta_2 = \measuredangle(q, o, r)$, and let $a, b \geq 0$ be such that $\overrightarrow{oq} = a\,\overrightarrow{op} + b\,\overrightarrow{or}$. Then we have

$$\begin{aligned} \cos\theta_1 &= \langle \overrightarrow{op}, \overrightarrow{oq} \rangle = a + b\,\langle \overrightarrow{op}, \overrightarrow{or} \rangle \\ \cos\theta_2 &= \langle \overrightarrow{oq}, \overrightarrow{or} \rangle = a\,\langle \overrightarrow{op}, \overrightarrow{or} \rangle + b, \end{aligned}$$

and hence,

$$\cos^2\theta_1 = a^2 + 2ab\,\langle \overrightarrow{op}, \overrightarrow{or} \rangle + b^2 \langle \overrightarrow{op}, \overrightarrow{or} \rangle^2 \tag{2.4.1}$$

$$\cos^2\theta_2 = a^2 \langle \overrightarrow{op}, \overrightarrow{or} \rangle^2 + 2ab\,\langle \overrightarrow{op}, \overrightarrow{or} \rangle + b^2 \tag{2.4.2}$$

$$\cos\theta_1 \cos\theta_2 = (a^2 + b^2)\langle \overrightarrow{op}, \overrightarrow{or} \rangle + ab\,\langle \overrightarrow{op}, \overrightarrow{or} \rangle^2 + ab. \tag{2.4.3}$$

Taking the norm of $\overrightarrow{oq}$ we also have

$$1 = \langle \overrightarrow{oq}, \overrightarrow{oq} \rangle = a^2 + b^2 + 2ab\,\langle \overrightarrow{op}, \overrightarrow{or} \rangle. \tag{2.4.4}$$

Subtracting first (2.4.1) from (2.4.4), and then (2.4.2) from (2.4.4), we arrive at

$$\begin{aligned}\sin^2\theta_1 &= b^2\,(1-\langle\overrightarrow{op},\overrightarrow{or}\rangle^2)\\ \sin^2\theta_2 &= a^2\,(1-\langle\overrightarrow{op},\overrightarrow{or}\rangle^2).\end{aligned}$$

Hence, since $a, b \geq 0$ and $|\langle\overrightarrow{op},\overrightarrow{or}\rangle| \leq \|\overrightarrow{op}\|\|\overrightarrow{or}\| = 1$,

$$\sin\theta_1 \sin\theta_2 = ab\,(1-\langle\overrightarrow{op},\overrightarrow{or}\rangle^2). \tag{2.4.5}$$

Combining (2.4.3), (2.4.4), and (2.4.5) yields

$$\cos\theta_1\cos\theta_2 = \langle\overrightarrow{op},\overrightarrow{or}\rangle + ab\,(1-\langle\overrightarrow{op},\overrightarrow{or}\rangle^2) = \langle\overrightarrow{op},\overrightarrow{or}\rangle + \sin\theta_1\sin\theta_2.$$

So

$$\begin{aligned}\cos\measuredangle(p,o,r) &= \langle\overrightarrow{op},\overrightarrow{or}\rangle = \cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2 = \cos(\theta_1+\theta_2)\\ &= \cos(\measuredangle(p,o,q)+\measuredangle(q,o,r)).\end{aligned}$$

Since cos is injective over the domain $[0,\pi]$, $\measuredangle(p,o,q)+\measuredangle(q,o,r) = \measuredangle(p,o,r)$. □

The proposition has as a corollary the following classic result.

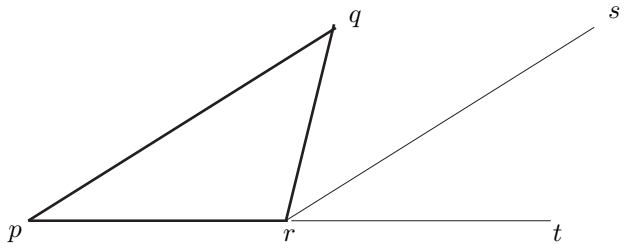


Figure 2.4.2: $\measuredangle(p,r,q)+\measuredangle(r,q,p)+\measuredangle(q,p,r) = \pi$

Proposition 2.4.4. *(Angle Sum of a Triangle Theorem) Let p, q, and r be any three (distinct) points. Then $\measuredangle(p,r,q)+\measuredangle(r,q,p)+\measuredangle(q,p,r) = \pi$.*

Proof. Let $s = r+\overrightarrow{pq}$, and $t = r+\overrightarrow{pr}$. (See figure 2.4.2.) Then $\overrightarrow{rs} = \overrightarrow{pq}$, $\overrightarrow{rt} = \overrightarrow{pr}$, and $\overrightarrow{pr} = \frac{1}{2}\overrightarrow{pt}$ (since $\overrightarrow{pt} = \overrightarrow{pr}+\overrightarrow{rt} = 2\overrightarrow{pr}$). Hence

$$\measuredangle(q,p,r) = \cos^{-1}\frac{\langle\overrightarrow{pq},\overrightarrow{pr}\rangle}{\|\overrightarrow{pq}\|\|\overrightarrow{pr}\|} = \cos^{-1}\frac{\langle\overrightarrow{rs},\overrightarrow{rt}\rangle}{\|\overrightarrow{rs}\|\|\overrightarrow{rt}\|} = \measuredangle(s,r,t).$$

Furthermore, since $\langle\overrightarrow{qr},-\overrightarrow{rs}\rangle = \langle-\overrightarrow{qr},\overrightarrow{rs}\rangle = \langle\overrightarrow{rq},\overrightarrow{rs}\rangle$,

$$\begin{aligned}\measuredangle(r,q,p) &= \cos^{-1}\frac{\langle\overrightarrow{qr},\overrightarrow{qp}\rangle}{\|\overrightarrow{qr}\|\|\overrightarrow{qp}\|} = \cos^{-1}\frac{\langle\overrightarrow{qr},-\overrightarrow{rs}\rangle}{\|\overrightarrow{qr}\|\|-\overrightarrow{rs}\|}\\ &= \cos^{-1}\frac{\langle\overrightarrow{rq},\overrightarrow{rs}\rangle}{\|\overrightarrow{rq}\|\|\overrightarrow{rs}\|} = \measuredangle(q,r,s).\end{aligned}$$

Hence

$$\measuredangle(p,r,q) + \measuredangle(r,q,p) + \measuredangle(q,p,r) = \measuredangle(p,r,q) + \measuredangle(q,r,s) + \measuredangle(s,r,t).$$

But by the additivity of angular measure, the sum on the right side is equal to the straight angle $\measuredangle(p,r,t)$. And the latter, by problem 2.4.3, is equal to π. So we are done. $\square$

3 Minkowskian Geometry and Its Physical Significance

We take an n-dimensional Minkowskian space (with $n \geq 2$) to be a metric affine space of dimension n and signature $(1, n-1)$. We know (by proposition 2.3.4) that, up to isometry, there is only one such space. In section 3.1, we proceed formally and develop certain elements of Minkowskian geometry in parallel to our development of Euclidean geometry in 2.4. Then, in section 3.2, we turn to the physical significance of Minkowskian geometry (in the case where n is 4). There we take the underlying set of points to represent the totality of all point-event locations in spacetime, and relate the geometry to physical processes in the world (involving point particles, light rays, clocks, and so forth). In section 3.3, we show that the standard Minkowskian angular measure function is the only candidate that satisfies certain natural conditions. Finally, in section 3.4 we consider the relative simultaneity relation in special relativity, and prove a cluster of related uniqueness results for it as well. The latter are motivated by a longstanding debate over the status of the relative simultaneity relation in special relativity. (Is it "conventional" in character, or rather, in some significant sense, forced on us?)

3.1 Minkowskian Geometry – Formal Development

In what follows, let $(\mathbf{A}, \langle\ ,\ \rangle)$ be an n-dimensional Minkowskian space ($n \geq 2$), with underlying point set A and vector space $\mathbf{V}$.

The inner product generates a classification of vectors in V. We say u is

timelike	if $\langle u, u\rangle > 0$
null (or *lightlike*)	if $\langle u, u\rangle = 0$
causal	if $\langle u, u\rangle \geq 0$
spacelike	if $\langle u, u\rangle < 0$.

We also say that a subspace W of V is *spacelike* if it is negative definite, i.e., if all its non-zero vectors are spacelike. (There are no subspaces of dimension higher than 1 all of whose non-zero vectors are causal. See problem 3.1.1 below.)

Proposition 3.1.1. *Let u be a timelike vector in V. Then the following all hold.*

(i) *Every vector v in V has a unique decomposition of form $v = au + w$ where $a \in \mathbb{R}$ and w is in $u^\perp$.*

(ii) *$u^\perp$ has dimension $(n-1)$.*

(iii) *$u^\perp$ is spacelike.*

(iv) *Given any $(n-1)$-dimensional subspace W that is spacelike, there is a timelike vector v such that $v^\perp = W$.*

Proof. Clauses (i) and (ii) follow as special cases of the projection theorem (proposition 2.3.1). We proved the latter in a more general setting, without restriction on the signature of the inner product.

(iii) $u^\perp$ has an orthonormal basis $\{u_2, u_3, ..., u_n\}$ (with respect to the inner product $\langle\ ,\ \rangle$, as restricted to $u^\perp$). This follows from proposition 2.3.2. (We saw in the proof of that proposition that the restriction of $\langle\ ,\ \rangle$ to $u^\perp$ qualifies as a generalized inner product on that subspace.) Let u_1 be the normalized vector $\frac{u}{\langle u,u\rangle^{\frac{1}{2}}}$. Then the expanded set $\{u_1, ..., u_n\}$ is an orthonormal basis for V (by proposition 2.3.1). Since the signature of $\langle\ ,\ \rangle$ is $(1, n-1)$, and $\langle u_1, u_1\rangle = 1$, it must be the case that $\langle u_i, u_i\rangle = -1$, for $i = 2, ..., n$. It follows that $u^\perp$ is a negative definite subspace. For given any vector $v = a_2u_2 + ... + a_nu_n$ in $u^\perp$,

$$\begin{aligned}\langle v, v\rangle &= \langle a_2u_2 + ... + a_nu_n, a_2u_2 + ... + a_nu_n\rangle \\ &= a_2^2\langle u_2, u_2\rangle + ... + a_n^2\langle u_n, u_n\rangle = -a_2^2 - ... - a_n^2.\end{aligned}$$

So $\langle v, v\rangle < 0$ unless $a_2 = ... = a_n = 0$, i.e., unless $v = \mathbf{0}$.

(iv) Let W be an $(n-1)$-dimensional subspace that is spacelike, i.e., the restriction of $\langle\ ,\ \rangle$ to W is negative definite. By proposition 2.3.2 again, there exists a basis for W consisting of orthogonal vectors $w_1, ..., w_{n-1}$ where $\langle w_i, w_i\rangle = -1$ for each i. Now let

$$v = u + \langle u, w_1\rangle w_1 + ... + \langle u, w_{n-1}\rangle w_{n-1}.$$

Clearly, v is orthogonal to each of the w_i, and hence to W. Moreover, v is timelike since

$$\langle v, v\rangle \ = \ \langle u, u\rangle + \langle u, w_1\rangle^2 + ... + \langle u, w_{n-1}\rangle^2 \ \geq \ \langle u, u\rangle \ > \ 0.$$

Since v is orthogonal to W, W is a subspace of $v^\perp$. Hence, since W and $v^\perp$ have the same dimension (namely $(n-1)$), it follows (e.g., from clause (iii) of proposition 2.1.1) that $W = v^\perp$. □

One immediate consequence of clause (iii) should be emphasized because it will come up so often in proofs: it is *not* possible for a timelike vector to be orthogonal to any other non-zero causal vector. That raises the question whether it is possible for two non-zero null vectors to be orthogonal. The answer is given in the following proposition.

Proposition 3.1.2. *Null vectors are orthogonal if and only if they are proportional (i.e., one is a scalar multiple of the other).*

Proof. Let v and w be null vectors. If $w = a\,v$ for some a, it follows trivially that v and w are orthogonal. (For in this case, $\langle v, w\rangle = a\,\langle v, v\rangle = 0$.)

Assume conversely that $\langle v, w\rangle = 0$. Let u be a timelike vector. By proposition 3.1.1, every non-zero vector in $u^\perp$ is spacelike. So (since w is not spacelike), either $w = \mathbf{0}$ or $\langle u, w\rangle \neq 0$. In the first case, v and w are trivially proportional. So we may assume that $\langle u, w\rangle \neq 0$. In this case, there is a number a such that $a\,\langle u, w\rangle = \langle u, v\rangle$. Hence, $(v - a\,w) \in u^\perp$. But $(v - a\,w)$ is not spacelike. (The right side of

$$\langle v - aw,\, v - aw\rangle = \langle v, v\rangle - 2a\,\langle v, w\rangle + a^2\,\langle w, w\rangle$$

is 0 since v and w are null, and $v \perp w$.) So, by proposition 3.1.1 again, $v - a\,w = \mathbf{0}$, i.e., v and w are proportional. □

Problem 3.1.1. *Show that there are no subspaces of dimension higher than 1 all of whose vectors are causal.*

In the present context, we take the (generalized) *norm* $\|u\|$ of a vector u to be the number $|\langle u, u\rangle|^{\frac{1}{2}}$.

Proposition 3.1.3. *Let u and v belong to a spacelike subspace W. Then*

(i) $|\langle u, v\rangle| \leq \|u\|\|v\|$, *with equality iff u and v are proportional.*

(ii) $\|u + v\| \leq \|u\| + \|v\|$, *with equality iff u and v are proportional with a non-negative proportionality factor.*

Proof. Though $\langle\ ,\ \rangle$ is not definite, its restriction to a spacelike subspace is negative definite. So the inner product $\langle\ ,\ \rangle'$ on W defined by $\langle u, v\rangle' = -\langle u, v\rangle$ is positive definite. Application of the Schwarz inequality and the triangle inequality to $\langle\ ,\ \rangle'$ yields clauses (i) and (ii). □

Intuitively, the set of timelike vectors forms a double cone. We can capture this idea formally as follows. Let us say that two timelike vectors u and v are *co-oriented* (or *have the same orientation*) if $\langle u, v\rangle > 0$.

Proposition 3.1.4. *Co-orientation is an equivalence relation on the set of timelike vectors in V.*

Proof. Reflexivity and symmetry are immediate. For transitivity, assume u, v, w are timelike vectors in V, and the pairs u, v and v, w are co-oriented. We must show that u, w are co-oriented.

Since $\langle u, v\rangle > 0$ and $\langle v, w\rangle > 0$, there is a real number $k > 0$ such that $\langle u, v\rangle = k\,\langle w, v\rangle$. Hence, $\langle u - k\,w, v\rangle = 0$. Since v is timelike, it follows either that $u - k\,w$ is the zero vector or it is spacelike. In the first case, $u = k\,w$, and so the pair u, w is certainly co-oriented ($\langle u, w\rangle = k\,\langle w, w\rangle > 0$). So we may assume that $u - k\,w$ is spacelike. But then

$$\langle u, u\rangle - 2k\,\langle u, w\rangle + k^2\langle w, w\rangle = \langle u - k\,w, u - k\,w\rangle < 0.$$

Since $\langle u, u\rangle, \langle w, w\rangle$, and k are all positive, it follows that $\langle u, w\rangle$ is positive as well. So, again, we are led to the conclusion that the pair u, w is co-oriented. □

We call the equivalence classes of timelike vectors under this relation *temporal lobes*. There must be at least two lobes since, for any timelike vector u in V, u and $-u$ are not equivalent. There cannot be more than two since, for all timelike vectors u, v in V, either $\langle u, v\rangle > 0$ or $\langle -u, v\rangle > 0$. (Remember, two timelike vectors cannot be orthogonal.) Hence there are exactly two lobes. It is easy to check that each lobe is convex, i.e., if u, v are co-oriented timelike vectors and $a, b > 0$, then $(au + bv)$ is a timelike vector co-oriented with u and v.

The relation of co-orientation can easily be extended to the larger set of non-zero causal (i.e., timelike or null) vectors. Given any two such vectors u and v, we take them to be *co-oriented* if *either* $\langle u, v\rangle > 0$ or $v = au$ with $a > 0$. (The second possibility must be allowed since we want a null vector to count as co-oriented with itself.) Once again, co-orientation turns out to be an equivalence relation with two equivalence classes that we call *causal lobes*. These lobes, too, are convex. (Only minor changes in the proof of proposition 3.1.4 are required to establish that the extended co-orientation relation is transitive.)

Problem 3.1.2. *One might be tempted to formulate the extended definition this way: two causal vectors are "co-oriented" if* $\langle u, v\rangle \geq 0$. *But this will not work. Explain why.*

We take a *temporal orientation* of Minkowski spacetime to be a specification of one temporal (or causal) lobe as the "future lobe". In the presence of such an orientation, we can speak of "future directed" and "past directed" timelike (and null) vectors.

The next proposition formulates Minkowskian counterparts to propositions 2.4.1 and 2.4.2 (concerning Euclidean spaces).

Proposition 3.1.5. *Let u and v be causal vectors in V.*

(i) *("Wrong way Schwarz inequality"):* $|\langle u, v\rangle| \geq \|u\|\|v\|$, *with equality iff u and v are proportional.*

(ii) *("Wrong way triangle inequality"): If u and v are co-oriented,*

$$\|u + v\| \geq \|u\| + \|v\|,$$

with equality iff u and v are proportional.

Proof. (i) If both u and v are null, the assertion follows immediately from proposition 3.1.2. So we may assume that one of the vectors, say u, is timelike. We can express v in the form $v = au + w$ where $w \in u^\perp$. Hence, $\langle u, v\rangle = a\langle u, u\rangle$ and $\langle v, v\rangle = a^2\langle u, u\rangle + \langle w, w\rangle$. Since w belongs to $u^\perp$, it must be spacelike or the zero vector. In either case, $\langle w, w\rangle \leq 0$. So, since u and v are causal, it follows that

$$\langle u, v\rangle^2 = a^2\langle u, u\rangle^2 = (\langle v, v\rangle - \langle w, w\rangle)\langle u, u\rangle \geq \langle v, v\rangle\langle u, u\rangle = \|u\|^2\|v\|^2.$$

Equality holds iff $\langle w, w\rangle = 0$. But, as noted above, w is either spacelike or the zero vector. So, equality holds iff $w = \mathbf{0}$, i.e, $v = a\,u$.

(ii) Assume u, v are co-oriented. Then either $\langle u, v\rangle > 0$, or both vectors are null and $v = au$ for some number $a > 0$. In the latter case, $\|u + v\| = \|u\| = \|v\| = 0$, and the assertion follows trivially. So we may assume that $\langle u, v\rangle > 0$. Hence (by clause (i)), $\langle u, v\rangle \geq \|u\|\|v\|$ Therefore,

$$\begin{aligned}(\|u\| + \|v\|)^2 &= \|u\|^2 + 2\|u\|\|v\| + \|v\|^2 \;\leq\; \langle u, u\rangle + 2\langle u, v\rangle + \langle v, v\rangle \\ &= \langle u + v, u + v\rangle = \|u + v\|^2.\end{aligned}$$

(For the last equality we need the fact that, since u, v are co-oriented, $u + v$ is causal.) Equality holds here iff $\langle u, v\rangle = \|u\|\|v\|$. But, by clause (i) again, this condition holds iff u and v are proportional. □

So far we have formulated our remarks in terms of vectors in V. Now we switch and formulate them directly in terms of points in A. Our terminology carries over naturally. Given points p, q in A, we say they are timelike (null, etc.) related if $\overrightarrow{pq}$ is timelike (null, etc.). If p and q are both causally related to a third point o, we say that p and q are in the same causal lobe of o if $\overrightarrow{op}$ and $\overrightarrow{oq}$ are co-oriented. And so forth.

Problem 3.1.3. *Let o, p, q be three points in A such that p is spacelike related to o, and q is timelike related to o. Show that any two of the following conditions imply the third.*

(i) *$\overrightarrow{pq}$ is null.*
(ii) $\overrightarrow{op} \perp \overrightarrow{oq}$
(iii) $\|\overrightarrow{op}\| = \|\overrightarrow{oq}\|$

Given points o, p, q in A, with o distinct from p, q, the vectors $\overrightarrow{op}$ and $\overrightarrow{oq}$ form a (possbily degenerate) angle. There are two special cases in which we can associate a natural angular measure $\measuredangle(p, o, q)$ with it. If $\overrightarrow{op}$ and $\overrightarrow{oq}$ are both spacelike, we can proceed much as in the case of Euclidean geometry. We can take $\measuredangle(p, o, q)$ to be the unique number θ in the interval $[0, \pi]$ such that $\langle \overrightarrow{op}, \overrightarrow{oq} \rangle = -\|\overrightarrow{op}\| \|\overrightarrow{oq}\| \cos\theta$. (We need to insert the minus sign because the restriction of the inner product $\langle\ ,\ \rangle$ to the subspace spanned by two spacelike vectors is negative definite.)

Of greater interest to us in what follows is the second case. If $\overrightarrow{op}$ and $\overrightarrow{oq}$ are timelike and co-oriented, we take $\measuredangle(p, o, q)$ to be the unique number $\theta \geq 0$ such that $\langle \overrightarrow{op}, \overrightarrow{oq} \rangle = \|\overrightarrow{op}\| \|\overrightarrow{oq}\| \cosh\theta$. (Note that by the wrong way Schwarz inequality, $\dfrac{\langle \overrightarrow{op}, \overrightarrow{oq} \rangle}{\|\overrightarrow{op}\| \|\overrightarrow{oq}\|} \geq 1$. So existence and uniqueness follow from the fact that the hyperbolic cosine function maps $[0, \infty)$ onto $[1, \infty)$ injectively. Basic facts about the hyperbolic functions are summarized at the end of the section.)

In both special cases, it follows immediately that $\measuredangle(p, o, q) = \measuredangle(q, o, p)$, and that $\measuredangle(p, o, q)$ does not depend on the length of the vectors $\overrightarrow{op}$ and $\overrightarrow{oq}$. More precisely, if $\overrightarrow{op}' = a\,\overrightarrow{op}$ and $\overrightarrow{oq}' = b\,\overrightarrow{oq}$, with $a, b > 0$, then $\measuredangle(p', o, q') = \measuredangle(p, o, q)$.

These angular measure functions satisfy additivity conditions much like the one considered in proposition 2.4.3 (in the context of Euclidean geometry). For the case of "spacelike angles", the proof is essentially the same. For "timelike angles", a few systematic changes are necessary. One arrives at the new version of the proof by systematically substituting the hyperbolic functions cosh and sinh for the trigonometric functions sin and cos.

Proposition 3.1.6. *. Let o, p, q, r be co-planar points with o distinct from p, q, r. Suppose that (i) the vectors $\overrightarrow{op}$, $\overrightarrow{oq}$, $\overrightarrow{or}$ are timelike and co-oriented, and (ii) $\overrightarrow{oq}$ is between $\overrightarrow{op}$ and $\overrightarrow{or}$. Then $\measuredangle(p, o, q) + \measuredangle(q, o, r) = \measuredangle(p, o, r)$.*

Proof. We may assume $\|\overrightarrow{op}\| = \|\overrightarrow{oq}\| = \|\overrightarrow{or}\| = 1$. Let $\theta_1 = \measuredangle(p, o, q)$, $\theta_2 = \measuredangle(q, o, r)$, and let $a, b \geq 0$ be such that $\overrightarrow{oq} = a\,\overrightarrow{op} + b\,\overrightarrow{or}$. Then we have

$$\begin{aligned} \cosh\theta_1 &= \langle \overrightarrow{op}, \overrightarrow{oq} \rangle = a + b\,\langle \overrightarrow{op}, \overrightarrow{or} \rangle \\ \cosh\theta_2 &= \langle \overrightarrow{oq}, \overrightarrow{or} \rangle = a\langle \overrightarrow{op}, \overrightarrow{or} \rangle + b, \end{aligned}$$

and hence,

$$\begin{aligned}
\cosh^2\theta_1 &= a^2 + 2ab\,\langle\overrightarrow{op},\overrightarrow{or}\rangle + b^2\langle\overrightarrow{op},\overrightarrow{or}\rangle^2 && (3.1.1)\\
\cosh^2\theta_2 &= a^2\langle\overrightarrow{op},\overrightarrow{or}\rangle^2 + 2ab\,\langle\overrightarrow{op},\overrightarrow{or}\rangle + b^2 && (3.1.2)\\
\cosh\theta_1\cosh\theta_2 &= (a^2+b^2)\langle\overrightarrow{op},\overrightarrow{or}\rangle + ab\,\langle\overrightarrow{op},\overrightarrow{or}\rangle^2 + ab. && (3.1.3)
\end{aligned}$$

Taking the norm of $\overrightarrow{oq}$ we also have

$$1 = \langle\overrightarrow{oq},\overrightarrow{oq}\rangle = a^2 + b^2 + 2\,ab\,\langle\overrightarrow{op},\overrightarrow{or}\rangle. \qquad (3.1.4)$$

Subtracting first (3.1.4) from (3.1.1), and then (3.1.4) from (3.1.2), we arrive at

$$\begin{aligned}
\sinh^2\theta_1 &= b^2(\langle\overrightarrow{op},\overrightarrow{or}\rangle^2 - 1)\\
\sinh^2\theta_2 &= a^2(\langle\overrightarrow{op},\overrightarrow{or}\rangle^2 - 1).
\end{aligned}$$

Hence, since $a, b \geq 0$ and $\|\langle\overrightarrow{op},\overrightarrow{or}\rangle\| \geq \|\overrightarrow{op}\|\|\overrightarrow{or}\| = 1$,

$$\sinh\theta_1\sinh\theta_2 \;=\; ab\,(\langle\overrightarrow{op},\overrightarrow{or}\rangle^2 - 1). \qquad (3.1.5)$$

Combining (3.1.3), (3.1.4), (3.1.5) yields

$$\cosh\theta_1\cosh\theta_2 \;=\; \langle\overrightarrow{op},\overrightarrow{or}\rangle - ab\,(\langle\overrightarrow{op},\overrightarrow{or}\rangle^2 - 1) \;=\; \langle\overrightarrow{op},\overrightarrow{or}\rangle - \sinh\theta_1\sinh\theta_2.$$

So

$$\begin{aligned}
\cosh\measuredangle(p,o,r) &= \langle\overrightarrow{op},\overrightarrow{or}\rangle \;=\; \cosh\theta_1\cosh\theta_2 + \sinh\theta_1\sinh\theta_2\\
&= \cosh(\theta_1+\theta_2) \;=\; \cosh(\measuredangle(p,o,q) \,+\, \measuredangle(q,o,r)).
\end{aligned}$$

Since cosh is injective over the domain $[0,\infty)$, $\measuredangle(p,o,q) + \measuredangle(q,o,r) = \measuredangle(p,o,r)$. □

In section 3.3 we will show that the angular measure function we have introduced for timelike angles is (up to a constant) the *only* candidate that satisfies both the additivity condition above and certain natural continuity and invariance conditions.

The next proposition gives a Minkowskian analogue of the Pythagorean theorem and the standard projection formulas of Euclidean trigonometry.

Proposition 3.1.7. *Let p and q be points timelike related to o, falling in the same temporal lobe of o, such that $\overrightarrow{op} \perp \overrightarrow{pq}$ (see figure 3.1.1.) Then*

(i) $\|\overrightarrow{oq}\|^2 = \|\overrightarrow{op}\|^2 - \|\overrightarrow{pq}\|^2$.

(ii) $\|\overrightarrow{op}\| = \|\overrightarrow{oq}\|\cosh\theta$ and $\|\overrightarrow{pq}\| = \|\overrightarrow{oq}\|\sinh\theta$, where $\theta = \measuredangle(poq)$.

Proof. (i) Since $\overrightarrow{oq} = \overrightarrow{op} + \overrightarrow{pq}$ and $\overrightarrow{op} \perp \overrightarrow{pq}$, $\langle\overrightarrow{oq},\overrightarrow{oq}\rangle = \langle\overrightarrow{op},\overrightarrow{op}\rangle + \langle\overrightarrow{pq},\overrightarrow{pq}\rangle$. Our result now follows because $\overrightarrow{oq}$ and $\overrightarrow{op}$ are timelike, and $\overrightarrow{pq}$ is spacelike or the zero vector (by clause (iii) of proposition 3.1.1).

(ii) We have $\langle\overrightarrow{op},\overrightarrow{oq}\rangle = \|\overrightarrow{op}\|\|\overrightarrow{oq}\|\cosh\theta$ and

$$\langle\overrightarrow{op},\overrightarrow{oq}\rangle = \langle\overrightarrow{op},\overrightarrow{op}+\overrightarrow{pq}\rangle = \langle\overrightarrow{op},\overrightarrow{op}\rangle = \|\overrightarrow{op}\|^2.$$

Hence, $\|\overrightarrow{op}\| = \|\overrightarrow{oq}\|\cosh\theta$. This with (i) yields

$$\|\overrightarrow{pq}\|^2 = \|\overrightarrow{op}\|^2 - \|\overrightarrow{oq}\|^2 = \|\overrightarrow{oq}\|^2(\cosh^2\theta - 1) = \|\overrightarrow{oq}\|^2\sinh^2\theta.$$

Since $\theta \geq 0, \sinh\theta \geq 0$. So $\|\overrightarrow{pq}\| = \|\overrightarrow{oq}\|\sinh\theta$. □

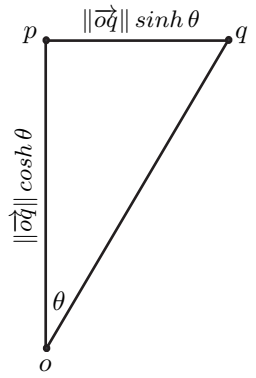


Figure 3.1.1: The "Minkowskian Pythagorean Theorem"

Problem 3.1.4. *Let p, q, r, s be distinct points in A such that (see figure 3.1.2)*

(i) *r, q, s lie on a timelike line with q between r and s, i.e., $\overrightarrow{rs}$ is timelike and $\overrightarrow{rq} = a\,\overrightarrow{rs}$ where $0 < a < 1$;*

(ii) *$\overrightarrow{rp}$ and $\overrightarrow{ps}$ are null.*

Show that $\overrightarrow{qp}$ is spacelike, and $\|\overrightarrow{qp}\|^2 = \|\overrightarrow{rq}\|\,\|\overrightarrow{qs}\|$.

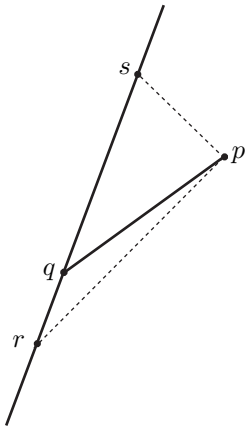


Figure 3.1.2: The "distance" between p and q can be expressed in terms of distances between points on a timelike line through one of them.

Problem 3.1.5. *Let L be a timelike line and let p be any point in A. Show the following.*

(i) *There is a unique point q on L such that $\overrightarrow{pq} \perp L$.*

(ii) *If $p \notin L$, there are exactly two points on L that are null related to p. (If $p \in L$, there is exactly one such point, namely p itself.)*

Our final topic in this section concerns the length of "timelike curves" in Minkowskian spaces. To prepare the way, we need to say something about limits and differentiability.

Let $\{u_i\}$ be a sequence of vectors in V. Given any vector u in V, we need to know what it means to say that $\{u_i\}$ *converges* to u or, equivalently, that u is the *limit* of $\{u_i\}$. (All other notions of concern to us can be defined in terms of this one.) We will take it to mean that, for all vectors w in V, the sequence $\langle u_i, w\rangle$ converges to $\langle u, w\rangle$. (The latter assertion makes sense because the elements involved are real numbers. Here we are back in the domain of basic calculus.) It should be noted that we cannot take the statement to mean that $\|u - u_i\|$ converges to 0. This characterization is available if the inner product with which one is working is positive (or negative) definite. But it is not available here – essentially because the Minkowski norm of a vector can be 0 without the vector being $\mathbf{0}$. For example, let u be a non-zero null vector, and let $u_i = 10u$ for all i. Then $\|u - u_i\| = \|9u\| = 0$ for all i, but we don't want to say that the sequence $\{u_i\}$ converges to u.

Next, consider a map $u : I \to V$ where I is an interval of the form (a, b), $[a, b)$, $(a, b]$ or $[a, b]$, with a and b elements of the extended real line $\mathbb{R}\cup\{\infty\}$. We say it is *continuous* at s in I, naturally, if, for any sequence $\{s_i\}$ in I converging to s, the sequence $u(s_i)$ converges to $u(s)$. And we say it is *differentiable* at s if, for any sequence $\{s_i\}$ in I converging to s (with $s_i \neq s$), the sequence $\dfrac{u(s) - u(s_i)}{(s - s_i)}$ converges to some vector in V. When the condition obtains, we take that vector to be the *derivative* of u at s and use the notation $\dfrac{du}{ds}(s)$ or $u'(s)$ for it.

Various facts about derivatives can now be established much as they would in a standard course in calculus. Suppose $u : I \to V$ and $v : I \to V$ are differentiable and so is the real valued function $f : I \to \mathbb{R}$. Then, for example, both the following hold for all s in I:

$$\frac{d}{ds}(u + v) = \frac{du}{ds} + \frac{dv}{ds} \tag{3.1.6}$$

$$\frac{d}{ds}(fu) = f\frac{du}{ds} + \frac{df}{ds}u. \tag{3.1.7}$$

And if w_0 is any individual vector, then

$$\frac{d}{ds}\langle u, w_0\rangle = \langle \frac{du}{ds}, w_0 \rangle. \tag{3.1.8}$$

Now, finally, consider a *curve* in A, i.e., a map $\gamma : I \to A$, where I is an interval as above. Given any point p in A, we can represent γ in the form $\gamma(s) = p + u(s)$, where $u : I \to V$ is defined by setting $u(s) = \overrightarrow{p\,\gamma(s)}$. We say that γ is *differentiable* at s in I if u is, and, when the condition is satisfied, take its *derivative* or *tangent vector* there to be $u'(s)$. (We use the notation $\dfrac{d\gamma}{ds}(s)$ or $\gamma'(s)$ for this derivative.) Of course, we need to verify that our definition of $\gamma'(s)$ is well-posed, i.e., does not depend on the initial choice of "base point" p. But this is easy. Let q and $w : I \to V$ be such that $\gamma(s) = p + u(s) = q + w(s)$, for all s in L. Then $w(s) = \overrightarrow{qp} + u(s)$ and, therefore,

$$w(s) - w(s_i) = (\overrightarrow{qp} + u(s)) - (\overrightarrow{qp} + u(s_i)) = u(s) - u(s_i)$$

for all s in I.

We know now what it means to say that a curve in A is differentiable. Other basic notions from calculus (e.g., second (and higher order) differentiability, piecewise differentiability, and so forth) can be handled similarly.

Let $\gamma : I \to A$ be differentiable. We say it is *timelike* (respectively *null, causal, spacelike*) if its tangent vector $\gamma'(s)$ is timelike (respectively null, causal, spacelike) at all points s in I. We can picture timelike curves, for example, as ones that thread the null cones of all the points through which they pass. And we can picture null curves as ones whose tangent vectors at every point are tangent to the null cone based at that point.

Note that null curves need not be straight. (A curve $\gamma : I \to A$ is *straight* if it can be represented in the form $\gamma(s) = p + f(s)\, v$, where v is a vector in V, and f is a real-valued function on I, i.e., if its tangent vectors at different points are all proportional to one another). For example, one can have a null curve in the shape of a helix. (All that is required is that the helix have exactly the right pitch.) Let's verify this explicitly, just for the practice. Let p be a point in A, let u be a unit timelike vector in V, and let v and w be unit spacelike vectors orthogonal to each other and to u. Consider the curve $\gamma : \mathbb{R} \to A$ defined by setting $\gamma(s) = p + su + (\cos s)v + (\sin s)\, w$. We have $\gamma'(s) = u - (\sin s)v + (\cos s)w$ and, so,

$$\begin{aligned} \langle \gamma'(s), \gamma'(s) \rangle &= \langle u, u \rangle + (\sin^2 s)\, \langle v, v \rangle + (\cos^2 s) \langle w, w \rangle \\ &= 1 - (\sin^2 s) - (\cos^2 s) = 0 \end{aligned}$$

for all s in $\mathbb{R}$. Thus γ is, indeed, null.

Given any differentiable curve $\gamma : I \to A$, we can associate with it a *length*:

$$\|\gamma\| = \int_I \|\gamma'(s)\|\, ds.$$

(So, for example, if γ is null, then $\|\gamma'(s)\| = 0$ for all s, and so $\|\gamma\| = 0$.) And we can take the length of a piecewise ("jointed") differentiable curve to be the sum of the lengths of its pieces. But this notion of length is not of much interest in general. One reason is the following. Given any two points in A, one can connect them with a piecewise differentiable curve whose length is 0. (It suffices to consider a zig-zag (piecewise differentiable) curve all of whose segments are null.) One can also connect them with differentiable curves of arbitrarily large length. Thus, in Minkowskian spaces, straight lines can neither be characterized as (images of) differentiable curves that *minimize* length between pairs of points, nor as ones that *maximize* length between them.

The situation does not get much better if one restricts attention to spacelike curves. For given any two spacelike related points in A, and any $\epsilon > 0$, one can connect them with a differentiable spacelike curve whose length is less than ϵ (as well as with differentiable spacelike curves of arbitrarily large length). One can arrive at the former by first connecting the points in question with a two segment piecewise differentiable null curve (with length 0), and then approximating it sufficiently closely with a differentiable spacelike curve. Continuity

considerations guarantee that the approximating curve will have length close to 0.

But one *does* get an interesting theory of length if one restricts attention to timelike curves. Given any two timelike related points in A, and any $\epsilon > 0$, one can connect them with a differentiable timelike curve whose length is less than ϵ. (The argument here is much the same as in the case of spacelike curves. One looks to timelike curves that closely approximate a jointed null curve.) But – here is the asymmetry – there exists a *longest* timelike curve connecting them (unique up to reparametrization), namely a straight curve. (See figure 3.1.3.) This follows as a consequence of the wrong way triangle inequality. We capture the claim in the next proposition. It will be important later in connection with our discussion of the (so-called) "clock paradox".

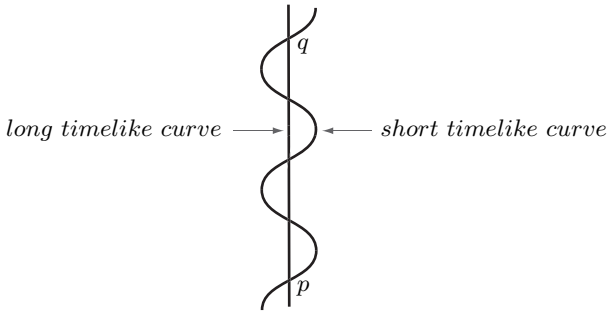


Figure 3.1.3: A long timelike curve from p to q and a very short one that swings back-and-forth, and approximates a broken null curve.

Proposition 3.1.8. *Let p and q be points in A that are timelike related, and let $\gamma : [s_1, s_2] \to A$ be a differentiable timelike curve with $\gamma(s_1) = p$ and $\gamma(s_2) = q$. Then $\|\gamma\| \leq \|\overrightarrow{pq}\|$, and equality holds iff γ is straight.*

Proof. We can express γ in the form $\gamma(s) = p + u(s)$ where $u : [s_1, s_2] \to V$ is differentiable. Since $\overrightarrow{pq}$ is timelike, it follows from proposition 3.1.1 that we can express $u(s)$ in the form $u(s) = f(s)\,\overrightarrow{pq} + w(s)$ where, for all s, $w(s)$ is orthogonal to $\overrightarrow{pq}$. Here $f : [s_1, s_2] \to \mathbb{R}$ and $w : [s_1, s_2] \to V$ are maps that, as is easy to check, make the assignments

$$f(s) = \frac{\langle u(s),\, \overrightarrow{pq}\rangle}{\langle \overrightarrow{pq},\, \overrightarrow{pq}\rangle} \quad \text{and} \quad w(s) = u(s) - f(s)\,\overrightarrow{pq}. \tag{3.1.9}$$

Note that since u is differentiable, so are f and w. Moreover,

$$\gamma'(s) \;=\; u'(s) \;=\; f'(s)\,\overrightarrow{pq} + w'(s) \tag{3.1.10}$$

for all s. These claims all follow from general facts about differentiability noted above. (Recall (3.1.6), (3.1.7), and (3.1.8).) It also follows that $w'(s) \perp \overrightarrow{pq}$ for

all s, since, by (3.1.8),

$$\langle w'(s), \overrightarrow{pq} \rangle \;=\; \frac{d}{ds} \langle w(s), \overrightarrow{pq} \rangle \;=\; 0.$$

We claim next that

$$\begin{array}{lllllll} u(s_1) & = & \mathbf{0} & \qquad f(s_1) & = & 0 & \qquad w(s_1) \;=\; \mathbf{0} \\ u(s_2) & = & \overrightarrow{pq} & \qquad f(s_2) & = & 1 & \qquad w(s_2) \;=\; \mathbf{0}. \end{array}$$

To see this, note first that $p = \gamma(s_1) = p + u(s_1)$ and $q = \gamma(s_2) = p + u(s_2)$. This gives us the equations in the first column. The equations in the next two columns then follow from (3.1.9).

Now recall (3.1.10). Since $w'(s)$ is orthogonal to the timelike vector $\overrightarrow{pq}$, it must be spacelike or the zero vector. Hence

$$\|\gamma'(s)\|^2 = (f'(s))^2 \|\overrightarrow{pq}\|^2 - \|w'(s)\|^2 \leq (f'(s))^2 \|\overrightarrow{pq}\|^2$$

for all s. But $f'(s) > 0$ for all s. (Why? $\gamma'(s)$, but not $w'(s)$, is timelike for all s. So, by (3.1.10) again, $f'(s) \neq 0$ for all s. Since $f(s_1) < f(s_2)$, it must be the case that f is everywhere increasing. So $f'(s) > 0$ for all s.) Thus, $\|\gamma'(s)\| \leq f'(s)\, \|\overrightarrow{pq}\|$, with equality iff $w'(s) = \mathbf{0}$ for all s. It follows that

$$\|\gamma\| = \int_{s_1}^{s_2} \|\gamma'(s)\|\, ds \leq \int_{s_1}^{s_2} f'(s)\, \|\overrightarrow{pq}\|\, ds = \|\overrightarrow{pq}\|\, (f(s_2) - f(s_1)) = \|\overrightarrow{pq}\|,$$

with equality iff $w'(s) = \mathbf{0}$ for all s. But the latter condition (the vanishing of $w'(s)$ for all s) holds iff $w(s) = \mathbf{0}$ for all s (since $w(s_1) = w(s_2) = \mathbf{0}$). So $\|\gamma\| = \|\overrightarrow{pq}\|$ precisely when $\gamma(s) = p + f(s)\, \overrightarrow{pq}$, i.e., when γ is a straight line segment connecting p and q. □

Appendix: Hyperbolic Functions

The functions $\cosh, \sinh$, and $\tanh$ are defined by

$$\begin{array}{rcl} \sinh\theta & = & \dfrac{(e^{\theta} - e^{-\theta})}{2} \\ \cosh\theta & = & \dfrac{(e^{\theta} + e^{-\theta})}{2} \\ \tanh\theta & = & \dfrac{\sinh\theta}{\cosh\theta}. \end{array}$$

They satisfy the following relations

$$\begin{aligned}
\cosh^2\theta - \sinh^2\theta &= 1 \\
\sinh(\theta_1+\theta_2) &= \sinh\theta_1\cosh\theta_2 + \cosh\theta_1\sinh\theta_2 \\
\sinh(\theta_1-\theta_2) &= \sinh\theta_1\cosh\theta_2 - \cosh\theta_1\sinh\theta_2 \\
\cosh(\theta_1+\theta_2) &= \cosh\theta_1\cosh\theta_2 + \sinh\theta_1\sinh\theta_2 \\
\cosh(\theta_1-\theta_2) &= \cosh\theta_1\cosh\theta_2 - \sinh\theta_1\sinh\theta_2 \\
\tanh(\theta_1+\theta_2) &= \frac{\tanh\theta_1+\tanh\theta_2}{1+\tanh\theta_1\tanh\theta_2} \\
\tanh(\theta_1-\theta_2) &= \frac{\tanh\theta_1-\tanh\theta_2}{1-\tanh\theta_1\tanh\theta_2}.
\end{aligned}$$

3.2 Minkowskian Geometry – Physical Interpretation

Relativity theory determines a class of geometrical models for the spacetime structure of our universe. Each represents a possible world (compatible with the constraints of the theory). Four-dimensional Minkowski space is the simplest of these models. (In the present context, it is customary to refer to it as Minkowski *spacetime*.) We can think of it as representing spacetime structure in the limiting case in which all gravitational effects vanish (or, equivalently, in which all spacetime curvature vanishes). The physical significance of Minkowski spacetime structure can be explained, at least partially, in terms of a cluster of interrelated physical principles that coordinate spacetime structure with physical objects and processes. Here we list a few important examples involving particles, light rays, clocks, and measuring rods. (The list could easily be extended.)

In what follows, let $(\mathbf{A}, \langle\ ,\ \rangle)$ be a four-dimensional Minkowskian space (with underlying point set A and vector space $\mathbf{V}$).

It is important that the principles, collectively, have an implicit counterfactual character. The claim is that (assuming relativity theory is correct), *if* all gravitational effects vanished, *then* we could set up a correspondence between points in A and "point-event-locations" in the world in such a way that the principles hold. (To avoid cumbersome formulations in what follows, we will sometimes collapse the distinction between the two and, for example, talk about particles "traveling from one point in A to another".)

Let's agree that "curves" in A are understood to be differentiable (unless there is indication to the contrary). But they certainly need not be straight. Indeed, the distinction between arbitrary, differentiable *curves* and straight *lines* is crucial in our formulation of the physical principles.

Group 1 (concerning point particles and light rays)

(CP 1) Timelike curves represent the spacetime trajectories of massive point particles, i.e., point particles with non-zero mass.

(CP 2) Timelike lines represent the spacetime trajectories of *free* massive point particles, i.e., massive point particles that are not subject to any forces.

(CP 3) Null lines represent the spacetime trajectories of light rays (traveling in a vacuum).

Group 2 (concerning clocks)

(CP 4) If p and q are timelike related points in A, then $\|\overrightarrow{pq}\|$ is the elapsed time between them as recorded by a freely falling natural clock whose spacetime trajectory contains these points.

(CP 4′) More generally, if p and q are timelike related points in A, and γ is a timelike curve that connects them, then the length $\|\gamma\|$ of γ is the elapsed time recorded by a natural clock with spacetime trajectory γ.

Group 3 (concerning measuring rods)

(CP5) If p and q are spacelike related points in A, then $\|\overrightarrow{pq}\|$ is the "spatial distance" between them as measured by a freely falling measuring rod, moving in such a way that the spacetime trajectories of its points are timelike lines orthogonal to $\overrightarrow{pq}$. (See figure 3.2.2 below.)

Several comments and qualifications are called for.

(1) In (CP 4) we take $\|\overrightarrow{pq}\|$ to be an interval of elapsed time. But $\|\overrightarrow{pq}\|$ is just a number (like 17). It should be understood that some choice of units for temporal distance, e.g., seconds, stands in the background. The same remark applies to (CP 4′). No additional choice of units for spatial distance is required to make sense of (CP5) because we have, in effect, built-in a connection between temporal and spatial distance. (It will turn out, for example, that the speed of light is 1.) If we measure temporal distance in seconds, then we measure spatial distance in "light seconds".

(2) We have made reference to the notion of particle "mass" in (CP1) and (CP2). We have in mind what is sometimes called "rest mass" or "inertial mass". Here it should be understood, simply, as a primitive attribute of particles. It is a basic fact of relativistic life that particles can be classified according to whether their mass is 0 or strictly positive. Conditions (CP1) and (CP2) concern particles of the latter sort.

(3) For certain purposes, even within classical (i.e., non-quantum mechanical) relativity theory, it is useful to think of light as constituted by streams of "photons" (a particular type of mass 0 particle), and take (CP3) to be the assertion that *null lines represent the spacetime trajectories of photons*. The latter formulation makes (CP3) look more like (CP1) and (CP2), and draws attention to the fact that the distinction between massive particles and mass 0 particles (like photons) has direct significance in terms of spacetime structure.

(4) We are here working within the framework of relativity as traditionally understood, and ignoring speculations about the possibility of particles (so-called "tachyons") that travel faster than light. (Their spacetime trajectories would be represented by spacelike curves.)

(5) We have built in the requirement that "curves" be smooth. So, depending on how one models collisions of point particles, one might want to restrict attention here to particles that do not experience collisions.

(6) The principles all involve complex idealizations. For example, (CP 4) and (CP 4′) take for granted that "natural clocks" can be represented as timelike curves. But real clocks exhibit some spatial extension, and so they are properly represented, not as timelike curves, but as "world tubes". What is true is that, in some circumstances, it is convenient and harmless to speak of point-sized "clocks". If pressed, one might try to cash this out in terms of a sequence of smaller and smaller clocks whose respective worldtubes converge to a single timelike curve.

(7) Without further qualification, (CP 4′) is really not even close to being true. It is formulated in terms of arbitrary natural clocks traversing arbitrary timelike curves (not just freely falling clocks traversing straight ones). But no clock does very well when subjected to extreme acceleration. Try smashing your wristwatch against the wall. A more careful formulation of (CP 4′) would have to be qualified along the following lines: natural clocks measure the Minkowskian length of the worldlines they traverse so long as they are not subjected to accelerations (or tidal forces) exceeding certain characteristic limits (that depend on the type of clock involved).

The issues raised here, particularly the role of idealization in the formulation of physical theory, are interesting and important. But they do not have much to do with relativity theory as such. They would arise in much the same way if we undertook to describe spacetime structure in classical Newtonian physics, and formulated a corresponding set of interpretive principles appropriate to that setting.

It would take us too far afield to properly discuss particle dynamics in relativity theory. But certain parts of the story are already to be found in the first two principles. (CP 2) captures a relativistic version of Newton's first law of motion. It asserts that *free* massive particles travel with constant (subluminal) velocity. And it follows from (CP 1) that no matter what forces we impress on a massive particle in a particle accelerator, we will never succeed in pushing it to the speed of light.

(CP4′) gives the whole story of relativistic clock behavior (modulo the concerns mentioned above). In particular, it implies the "path dependence" of clock readings. If two clocks start at a spacetime point p, and travel along different trajectories to a spacetime point q, then, in general, they will record different elapsed times for the trip. (E.g., one records an elapsed time of 365 seconds, the other 28 seconds.) This is true no matter how similar the clocks. (We may stipulate that they came off the same assembly line.) This is the case because, as (CP 4′) asserts, the elapsed time recorded by each of the clocks is just the length of the timelike curve that the clock traversed in getting from p to q and, in general, those lengths will be different.

In particular, if one clock (A) gets from p to q along a free fall trajectory (i.e., if it traverses a straight timelike line), and if the other (B) undergoes acceleration at some point during the trip (and so has a trajectory that is not straight), then

A will record a greater elapsed time than B. This follows because, as we know from proposition 3.1.8, the length of A's trajectory will be $\|\overrightarrow{pq}\|$ and the length of B's will be some number smaller than $\|\overrightarrow{pq}\|$. In Minkowskian geometry, again, of all timelike curves connecting p and q, the straight line connecting them is the longest, not the shortest. (Here is one way to remember how things work. To get a clock to read a smaller elapsed time between p to q than the maximal value, one will have to accelerate the clock. Now acceleration requires fuel, and fuel is not free. So we have the principle that *saving time costs money*!)

The situation described here was once thought paradoxical because it was believed that, "according to relativity theory", we are equally well entitled to think of clock B as being in a state of free fall and clock A as being the one that undergoes acceleration. And hence, by parity of reasoning, it should be clock B that records the greater elapsed time. The resolution, if one can call it that, is that relativity theory makes no such claim. The situations of A and B are *not* symmetric. B accelerates; A does not. The distinction between accelerated motion and free fall makes every bit as much (observer independent) sense in relativity theory as it does in classical physics.

Though in this course we are only dealing with so-called "special relativity", that special, limiting case of relativity in which all spacetime curvature is assumed to vanish, it is worth making one remark here about the general situation. If one considers only Minkowski spacetime, one might imagine that the distinction between free fall and accelerated motion plays a more important role in the determination of clock behavior than it does in fact. It is true *there* (in Minkowski spacetime) that given any two timelike related points p and q, there is a unique straight timelike line connecting them, and it is longer than any other (non-straight) timelike curve connecting the points. But relativity theory admits spacetime models (with curvature) in which the situation is qualitatively different. It can be the case that there is more than one timelike "geodesic" connecting two points, and these geodesics can have (and, in general, will have) different lengths. So it can be the case that clocks passing between two points record different elapsed times even though *both* are in a state of free fall. Furthermore – this follows from the preceding claim by continuity considerations alone – it can be the case that of two clocks passing between the points, the one that undergoes acceleration during the trip records a greater elapsed time than the one that remains in a state of free fall. (What *does* remain true in *all* relativistic spacetime models is a local version of the situation in Minkowksi spacetime. If one restricts attention to sufficiently small (and properly shaped) neighborhoods, then, given any two timelike related points in the neighborhood, there is a unique geodesic (in the neighborhood) connecting them, and its length is greater than that of any other timelike curve (in the neighborhood) connecting them.)

(CP 4′) has many interesting consequences. Here is one more. Given any two timelike related points p and q in A, and any $\epsilon > 0$, it is possible to travel from one to the other in such a way that the elapsed time of the trip (as recorded by the stopwatch one carries) is less than ϵ. This follows immediately since, as we saw in section 3.1, there is a timelike curve connecting p and q whose

length is less than ϵ. (Once again, there a zig-zig null curve connecting them, and it can be approximated arbitrarily closely with a timelike curve. See figure 3.1.3.) It suffices to follow that curve. (Here we pass over the possibility that the accelerations required for the trip would tear us and our stopwatch apart! We have encountered the principle that saving time costs money. Now we see that with enough money – enough to pay for all the fuel needed to zig and zag – *one can save as much time as one wants*!) This example shows the power of thinking about special relativity in geometric terms. The claim made is a striking one. But it is *obvious* if one has a good intuitive understanding of Minkowskian geometry (and keeps (CP 4′) in mind).

And speaking of intuitions, we should mention that though the path dependent behavior of clocks in relativity theory may seem startling at first, it does not take long to become perfectly comfortable with it. It helps to keep the analogy with automobile odometers in mind. If two cars are driven from NY to LA along different routes, their odometers will, in general, record different elapsed distances. Why? Because their routes have different lengths and odometers record route length. That is what the latter *do*. Similarly clocks record the length of their (four-dimensional) routes through Minkowski spacetime. That is what *they* do. The one phenomenon is no more puzzling than the other.

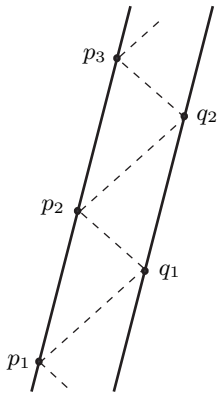


Figure 3.2.1: Bouncing Light Ray Clock

We described the different coordinating principles as "interrelated". Here is an example of the sense in which they are. We have considered clocks as primitive objects so far. But one can build a simple clock, of sorts, by bouncing a light ray between parallel mirrors in a joint state of free fall (see figure 3.2.1), and then use (CP 3) to prove that the clock measures elapsed spacetime distance along the worldlines of the mirrors. So one establishes a special case of (CP 4). More precisely, suppose light arrives at one of the mirrors at points $p_1, p_2, p_3, ...$ What one can prove is that if (i) the spacetime trajectories of the mirrors are parallel timelike lines, and (ii) the spacetime trajectory of the bouncing light

ray is a (zig-zag) null line, then the successive intervals $\|\overrightarrow{p_1p_2}\|, \|\overrightarrow{p_2p_3}\|, \ldots$ are all equal. So the number of light arrivals – "ticks" of the clock – is a measure of aggregate elapsed distance along the mirror's worldline. (The computation is given in the following proposition.)

Proposition 3.2.1. *Let p_1, p_2, p_3, q_1, q_2 be points satisfying the following conditions. (See figure 3.2.1.)*

(i) $\overrightarrow{p_1p_2}$ *is timelike*

(ii) $\overrightarrow{q_1q_2} = a\,\overrightarrow{p_1p_2}$ *for some* $a > 0$

(iii) $\overrightarrow{p_2p_3} = b\,\overrightarrow{p_1p_2}$ *for some* $b > 0$

(iv) *The vectors* $\overrightarrow{p_1q_1}$, $\overrightarrow{q_1p_2}$, $\overrightarrow{p_2q_2}$, $\overrightarrow{q_2p_3}$ *are all null.*

(v) *The vectors in (iv) all have the same orientation as* $\overrightarrow{p_1p_2}$. *(This assumption is actually redundant. It follows from the others.)*

Then $\overrightarrow{p_1p_2} = \overrightarrow{p_2p_3}$.

Proof. We will prove $\overrightarrow{p_1p_2} = \overrightarrow{q_1q_2}$. The very same argument can be used, symmetrically, to establish $\overrightarrow{q_1q_2} = \overrightarrow{p_2p_3}$. So our claim will follow.

Note first that, by (ii), $\overrightarrow{p_1p_2} + \overrightarrow{p_2q_2} = \overrightarrow{p_1q_1} + \overrightarrow{q_1q_2} = \overrightarrow{p_1q_1} + a\,\overrightarrow{p_1p_2}$, and so

$$(1-a)\,\overrightarrow{p_1p_2} = \overrightarrow{p_1q_1} - \overrightarrow{p_2q_2}.$$

Taking the inner product of each side with itself, and using (i) and (iv), we arrive at

$$-2\,\langle\overrightarrow{p_1q_1}, \overrightarrow{p_2q_2}\rangle = (1-a)^2\,\|\overrightarrow{p_1p_2}\|^2 \geq 0.$$

Hence, $\langle\overrightarrow{p_1q_1}, \overrightarrow{p_2q_2}\rangle \leq 0$. But, by (v), $\overrightarrow{p_1q_1}$ and $\overrightarrow{p_2q_2}$ are co-oriented. So it must be the case that $\langle\overrightarrow{p_1q_1}, \overrightarrow{p_2q_2}\rangle = 0$ and, therefore, $a = 1$. Thus $\overrightarrow{p_1p_2} = \overrightarrow{q_1q_2}$, as claimed. □

Let us now, finally, consider measuring rods and (CP5). It should be said immediately that, from the vantage point of relativity theory, measuring rods are extremely complicated objects. One can, for certain purposes, represent clocks by timelike curves. Correspondingly, one can (for certain purposes) represent measuring rods by two-dimensional timelike surfaces. (Let us agree that a surface is *timelike* if, at every point, there is a tangent vector to the surface that is timelike.) But the latter are much more complicated geometrically than timelike curves. In fact, we will consider only the special case of rods in a state of non-rotational free fall so that the timelike surfaces we have to work with will be (fragments of) two-dimensional planes.

Figure 3.2.2 shows one such fragment of a plane. The indicated parallel timelike lines are supposed to represent the worldlines of distance markers on our (non-rotating, free falling) measuring rod. Now suppose p and q are spacelike related points in the plane so situated that $\overrightarrow{pq}$ is orthogonal to the indicated timelike lines. (This amounts to saying, as we shall soon see, that p and q are simultaneous relative to an observer co-moving with the ruler.) To keep things simple, imagine that p and q both fall on marker lines (as in the figure). Our

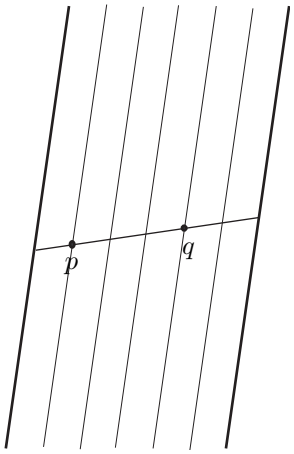


Figure 3.2.2: Measuring Rod

observer co-moving with the rod will naturally take the spatial distance between p and q to be the number of marker spaces between those lines. (CP5) tells us that this distance is (relative to our choice of units) precisely the magnitude $\|\overrightarrow{pq}\|$. Thus this coordinating principle too correlates geometrical magnitudes (determined by the inner product $\langle\ ,\ \rangle$) with (idealized) measurement procedures.

This completes our very brief discussion of principles (CP1) through (CP5). We now turn to the relation of "relative simultaneity" and the magnitudes derived from it – relative speed, relative temporal distance, and relative spatial distance.

Given a timelike vector u (or a timelike line L generated by u), we say that two points p and q are *simultaneous relative to* u (or relative to L) if $\overrightarrow{pq}$ is orthogonal to u. In section 3.4 we will consider to what extent this standard notion of relative simultaneity is arbitrary and to what extent it is forced. But for the moment, we put such concerns aside, and simply work out the consequences of adopting this notion.

Given any timelike vector u, the relation of simultaneity relative to u is an equivalence relation. Its associated equivalence classes may be called *simultaneity slices relative to* u. Each is a three-dimensional affine subspace generated by $u^\perp$. In particular, given any point p, the slice containing p is just $p+u^\perp$. (Recall from our discussion of affine subspaces in section 2.2 that a point q belongs to this set iff it is of form $q = p+v$ where v is in $u^\perp$. The latter condition holds iff $\overrightarrow{pq}$ is orthogonal to u. So q belongs $p + u^\perp$ iff q is simultaneous with p relative to u. Recall too that, by proposition 3.1.1, $u^\perp$ is a three-dimensional, spacelike subspace of V). One speaks of a "foliation" of A by (relative) simultaneity slices. (See figure 3.2.3.)

The standard textbook expressions for relative speed, relative temporal dis-

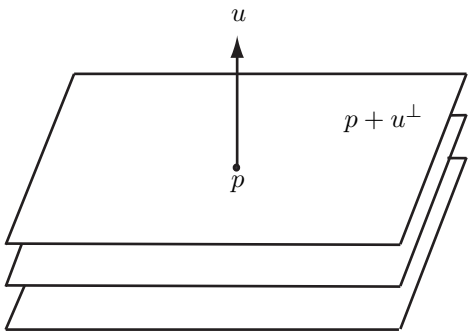


Figure 3.2.3: A Foliation of Minkowski Spacetime by Simultaneity Slices Relative to u.

tance, relative spatial distance and the like can all be derived using a simple formalism for projecting vectors. Let u be a unit timelike vector in V, which we may as well take to be future-directed. Let $P_u : V \to V$ and $P_{u^\perp} : V \to V$ be the associated linear maps that project a vector w onto its components, respectively, proportional to, and orthogonal to, u. Thus,

$$w \;=\; P_u(w) \,+\, P_{u^\perp}(w) \;=\; \underbrace{\langle u, w\rangle\, u}_{\textit{proportional to } u} \;+\; \underbrace{(w - \langle u, w\rangle\, u)}_{\textit{orthogonal to } u}. \quad (3.2.1)$$

Let L be a timelike line representing a free falling observer O, let o be a point on L, and let u be the (unique) future directed unit timelike vector that generates L. (So every point p on L can be uniquely expressed in the form $p = o + k\,u$, for some k.) Further, let L' be a second timelike line that contains o, and represents a free falling observer O'. Finally, let q be a point on L' distinct from o, and so positioned that $\overrightarrow{oq}$ is future-directed. Then there is a unique point p on L such that $\overrightarrow{op} \perp \overrightarrow{pq}$. It is given by $p = o + P_u(\overrightarrow{oq}) = o + \langle u, \overrightarrow{oq}\rangle\, u$. (See figure 3.2.4.)

Now consider the speed O attributes to O'. He takes the points p and q to be simultaneous. So, as he sees the matter, in moving from o to q, O' travels spatial distance $\|\overrightarrow{pq}\| = \|P_{u^\perp}(\overrightarrow{oq})\|$, in elapsed time $\|\overrightarrow{op}\| = \|P_u(\overrightarrow{oq})\|$. So

$$\begin{aligned} v_{OO'} &= \text{the speed that } O \text{ attributes to } O' \\ &= \frac{\|\overrightarrow{pq}\|}{\|\overrightarrow{op}\|} = \frac{\|\overrightarrow{oq}\|\sinh\theta}{\|\overrightarrow{oq}\|\cosh\theta} = \tanh\theta, \end{aligned} \quad (3.2.2)$$

where $\theta = \measuredangle(p, o, q)$. (Here we use clause (ii) of proposition 3.1.7.) Since the speed $v_{OO'}$ that O attributes to O' depends only on the angle between L and L', it follows immediately that it is equal to the speed $v_{O'O}$ that O' attributes to O. Attributions of speed here are perfectly symmetric, and we can formulate the basic result this way.

> *The relative speed between two freely falling individuals is* $\tanh\theta$*, where* θ *is the hyperbolic angle between their worldlines.*

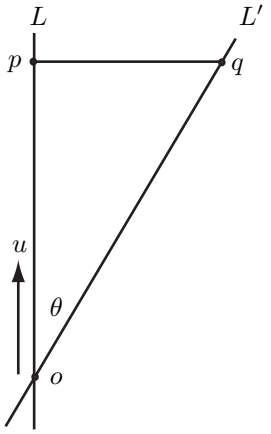


Figure 3.2.4: The relative speed of two observers is given by $\tanh\theta$, where θ is the hyperbolic angle between their worldlines.

It also follows immediately from (3.2.2) that *the relative speed v between freely falling individuals (we may safely drop the subscript) satisfies* $0 \leq v < 1$. Note for future reference that, since

$$v = \tanh\theta = \frac{\sinh\theta}{\cosh\theta} = \frac{(\cosh^2\theta - 1)^{\frac{1}{2}}}{\cosh\theta},$$

we have

$$\cosh\theta = \frac{1}{(1-v^2)^{\frac{1}{2}}}. \tag{3.2.3}$$

Now let L_1, L_2, L_3 be three timelike lines representing three freely falling individuals O_1, O_2, O_3 that pass each other at point o. Further let it be the case that the three lines are co-planar, with L_2 between the other two. Let θ_{ij} be the hyperbolic angle between lines L_i and L_j, and let v_{ij} be the relative speed between individuals O_i and O_j. By proposition 3.1.6, $\theta_{13} = \theta_{12} + \theta_{23}$. Hence

$$\begin{aligned} v_{13} &= \tanh\theta_{13} = \tanh(\theta_{12} + \theta_{23}) \\ &= \frac{\tanh\theta_{12} + \tanh\theta_{23}}{1 + \tanh\theta_{12}\,\tanh\theta_{23}} \\ &= \frac{v_{12} + v_{23}}{1 + v_{12}\,v_{23}}. \end{aligned} \tag{3.2.4}$$

This gives the "relativistic addition of velocities formula". The classical counterpart is $v_{13} = v_{12} + v_{23}$. The formula "explains", in a sense, why one cannot generate relative speeds greater than 1 by compounding them piggy-back style. If O_2 is on a train moving with speed $\frac{3}{4}$ relative to O_1, and shoots a bullet (in the direction of his motion relative to O_1) with speed $\frac{3}{4}$, then the bullet's speed relative to O_1 will be $\frac{24}{25}$ (not $\frac{3}{2}$).

Problem 3.2.1. *Let o, p, q, r, s be distinct points where (see figure 3.2.5)*

(i) *o, p, q lie on a timelike line L with p between o and q;*

(ii) *o, r, s lie on a timelike line L' with r between o and s;*

(iii) *$\overrightarrow{pr}$ and $\overrightarrow{qs}$ are null;*

(iv) *$\overrightarrow{oq}, \overrightarrow{pr}$, and $\overrightarrow{qs}$ are co-oriented.*

Show that

$$\frac{\|\overrightarrow{rs}\|}{\|\overrightarrow{pq}\|} = \left[\frac{1+v}{1-v}\right]^{\frac{1}{2}},$$

where v is the speed that the individual with worldline L attributes to the individual with worldline L' . This formula arises in discussions of the "relativistic Doppler effect". (Hint: First show that $\frac{\|\overrightarrow{rs}\|}{\|\overrightarrow{pq}\|} = \frac{\|\overrightarrow{os}\|}{\|\overrightarrow{oq}\|}$. Then show that if X is the ratio $\frac{\|\overrightarrow{os}\|}{\|\overrightarrow{oq}\|}$, $X^2 - 2X(1-v^2)^{-\frac{1}{2}} + 1 = 0$. Finally, show that $X > 1$. It follows that the equation has a unique solution: $X = \left[\frac{1+v}{1-v}\right]^{\frac{1}{2}}$.)

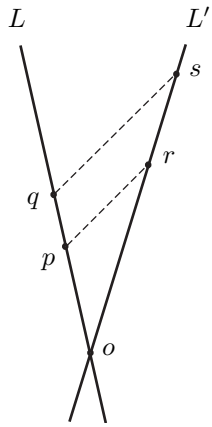


Figure 3.2.5: Figure for problem 3.2.1.

Problem 3.2.2. *Give a second derivation of the "relativistic addition of velocities formula" (3.2.4) using the result of problem 3.2.1, but not invoking proposition 3.1.6. (Hint: Apply the result to the pairs $\{L_1, L_2\}$, $\{L_2, L_3\}$, and $\{L_1, L_3\}$.)*

In our discussion of relative speed so far, we have considered two (or more) timelike lines meeting in a point o. Consider now the case where a timelike line L intersects with a null line L' at o. We can think of L as the worldline of a free falling observer O, and L' as the worldline of a light ray. (See figure 3.2.6.) Once again, let q be a point on L' to the future of o, and let p be the unique

point on L such that $\overrightarrow{op} \perp \overrightarrow{pq}$. In this case, arguing much as before, and using, for example, the result of problem 3.1.3, we reach the conclusion that

$$\text{the speed that O attributes to the light ray} = \frac{\|\overrightarrow{pq}\|}{\|\overrightarrow{op}\|} = 1. \tag{3.2.5}$$

Thus, the speed of light turns out to be constant in this strong sense: *it is the same, for all observers, in all directions.* In particular, if a light ray travels from o to q, and is then reflected back to L (after encountering a mirror), then, as judged by O, the speed of the light ray is same on its outgoing trip as it is on its return trip. (In both cases, the speed is 1.) This follows as an immediate consequence of the way we have construed the relation of relative simultaneity. We will return to consider the relation between (relative) light speed and (relative) simultaneity in section 3.4.

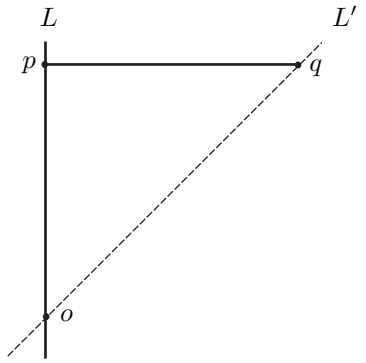


Figure 3.2.6: The speed of light (in vacuum), as determined relative to any observer, is 1.

Next we consider the notion of "relative elapsed time". Consider again the situation represented in figure 3.2.4. Timelike lines L and L' represent two freely falling individuals O and O' that pass each other at point o. O' says that the elapsed time between o and q is $\|\overrightarrow{oq}\|$, but O says that the elapsed time is $\|\overrightarrow{op}\|$, since he takes p and q to be simultaneous. But, by proposition 3.1.7 and equation (3.2.3), if v is the speed of O' relative to O,

$$\|\overrightarrow{op}\| = \|\overrightarrow{oq}\| \cosh\theta = \frac{\|\overrightarrow{oq}\|}{(1-v^2)^{\frac{1}{2}}} > \|\overrightarrow{oq}\|.$$

Thus,

> *if v is the relative speed between two freely falling individuals O and O', and if ΔT is the elapsed interval of time between two events on O''s worldline as determined by O', then O will determine the time interval to be* $\dfrac{\Delta T}{(1-v^2)^{\frac{1}{2}}}$.

O determines the elapsed interval to be greater than does O'. ("O says that O''s clock is running too slowly.") Again, our formulation is perfectly symmetric; the assertion remains true if one interchanges the roles of O and O'. ("O' also says that O's clock is running too slowly.") Though one does understand the statements in quotation marks, they are misleading and, in fact, have led to considerable misunderstanding. The relativistic "time dilation effect" should not be understood as resulting from some kind of clock "disturbance".

We close this section, finally, with a brief account of "relative spatial length". Consider a measuring rod in a state of free fall. Let its front and back ends be represented by timelike lines L' and L''. (See figure 3.2.7.) Further, let L be a timelike line, co-planar with L' and L'', that intersects the former at o. We take L to represent a free-falling observer O. Let θ be the hyperbolic angle between L and L', and let v be their relative speed. Finally, let p and r be points on L', and q a point on L'', such that (see figure 3.2.7), (i) $\overrightarrow{oq}$ is orthogonal to L', (ii) $\overrightarrow{pq}$ is orthogonal to L, and (iii) $\overrightarrow{qr}$ is co-aligned with L. It follows that $\measuredangle(o,r,q) = \measuredangle(p,r,q) = \theta$.

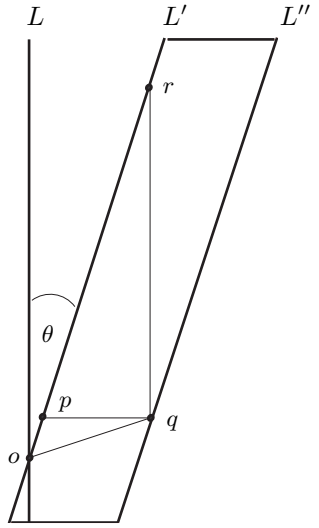


Figure 3.2.7: An observer with worldline L takes the line $L(p,q)$ to be a simultaneity slice of the rod. In contrast, an observer with worldline L' takes $L(o,q)$ to be one.

Now, by (CP5), O takes the length of the rod to be $\|\overrightarrow{pq}\|$. But an observer O' whose worldline is represented by L', i.e., one co-moving with the rod, takes it to be $\|\overrightarrow{oq}\|$. We are interested in expressing the first in terms of the second, and the relative speed v. To do so, we apply proposition 3.1.7 twice. First, we apply it to triangle p,q,r (with right angle $\measuredangle(p,q,r)$), and get $\|\overrightarrow{pq}\| = \|\overrightarrow{qr}\| \tanh\theta$. Next, we apply it to triangle r,o,q (with right angle $\measuredangle(r,o,q)$), and get $\|\overrightarrow{oq}\| = \|\overrightarrow{qr}\| \sinh\theta$.

It follows immediately that

$$\|\overrightarrow{pq}\| = \frac{\|\overrightarrow{oq}\|}{\cosh\theta} = \|\overrightarrow{oq}\|(1-v^2)^{\frac{1}{2}} < \|\overrightarrow{oq}\|.$$

Thus,

> *if v is the relative speed between two freely falling individuals O and O', and if ΔL is the length of a measuring rod co-moving with O' (aligned with the relative motion of O and O') as determined by O', then O will determine the length of the rod to be $\Delta L(1-v^2)^{\frac{1}{2}}$.*

Here we have the famous "length contraction" effect. The length O assigns to the rod is less than that assigned by a co-moving observer.

3.3 Uniqueness Result for Minkowskian Angular Measure

In this section, we show that the angular measure function we introduced for timelike angles in Minkowski spacetime is, up to a constant, the *only* candidate that satisfies both the additivity condition in proposition 3.1.6 and certain natural continuity and invariance conditions. We do so because the result is of some interest in its own right and, more important, because it will be useful to have a uniqueness result of this *type* in view before discussing the relation of relative simultaneity in special relativity. We will want to claim that the latter is uniquely distinguished by the structure of Minkowski spacetime in much the same sense that angular measure is.

In what follows, once again, let $(\mathbf{A}, \langle\ ,\ \rangle)$ be an n-dimensional Minkowskian space ($n \geq 2$), with underlying point set A and vector space $\mathbf{V}$. Further assume that a temporal orientation has been specified.

Proposition 3.3.1. *Let o be a point in A, and let H_o^+ be the set of all points p in A such that $\overrightarrow{op}$ is a future-directed unit timelike vector. (So H_o^+ is the "future half" of a two-sheeted hyperboloid.) Further, let $f : H_o^+ \times H_o^+ \to \mathbb{R}$ be a continuous map satisfying the following two conditions.*

(i) *(Additivity): For all points p, q, r in H_o^+ co-planar with o, if $\overrightarrow{oq}$ is between $\overrightarrow{op}$ and $\overrightarrow{or}$,*

$$f(p,r) = f(p,q) + f(q,r).$$

(ii) *(Invariance): If $\varphi : A \to A$ is an isometry of $(\boldsymbol{A}, \langle\ ,\ \rangle)$ that keeps o fixed and preserves temporal orientation, then, for all p and q in H_o^+,*

$$f(\varphi(p), \varphi(q)) = f(p,q).$$

Then there is a constant K such that, for all p and q in H_o^+, $f(p,q) = K \measuredangle(p,o,q)$. (Recall that $\measuredangle(p,o,q)$ is defined by the requirement that $\langle \overrightarrow{op}, \overrightarrow{oq} \rangle = \cosh \measuredangle(p,o,q)$.)

Note that we have the resources in hand for understanding the requirement that $f: H_o^+ \times H_o^+ \to \mathbb{R}$ be "continuous". This comes out as the condition that, for all p and q in H_o^+, and all sequences $\{p_i\}$ and $\{q_i\}$ in H_o^+, if $\{p_i\}$ converges to p and $\{q_i\}$ converges to q, then $f(p_i, q_i)$ converges to $f(p, q)$. (And the condition that $\{p_i\}$ converges to p can be understood to mean that the vector sequence $\{\overrightarrow{p_i\, p}\}$ converges to the zero vector $\mathbf{0}$.)

Note also that the invariance condition is well formulated. For if $\varphi: A \to A$ is an isometry of $(\mathbf{A}, \langle\ ,\ \rangle)$ that keeps o fixed and preserves temporal orientation, then $\varphi(p)$ and $\varphi(q)$ are both points on H_o^+ (and so $(\varphi(p), \varphi(q))$ is in the domain of f). $\varphi(p)$ belongs to H_o^+ since

$$\|\overrightarrow{o\,\varphi(p)}\| = \|\overrightarrow{\varphi(o)\,\varphi(p)}\| = \|\Phi(\overrightarrow{op})\| = \|\overrightarrow{op}\| = 1$$

and $\overrightarrow{o\,\varphi(p)}$ is future-directed. And similarly for $\varphi(q)$. (Here Φ is the vector space isomorphism associated with ϕ. Recall the discussion preceding proposition 2.2.6.)

Proof. Given any four points p_1, q_1, p_2, q_2 in H_o^+ with $\langle \overrightarrow{op}_1, \overrightarrow{oq}_1 \rangle = \langle \overrightarrow{op}_2, \overrightarrow{oq}_2 \rangle$, there is a temporal orientation preserving isometry $\varphi: A \to A$ such that $\varphi(o) = o$, $\varphi(p_1) = p_2$, and $\varphi(q_1) = q_2$. (We prove this after completing the main part of the argument.) It follows from the invariance condition that $f(p_1, q_1) = f(p_2, q_2)$. Thus we see that the number $f(p, q)$ depends only on the inner product $\langle \overrightarrow{op}, \overrightarrow{oq} \rangle$, i.e., there is a map $g: [1, \infty) \to \mathbb{R}$ such that

$$f(p, q) = g(\langle \overrightarrow{op}, \overrightarrow{oq} \rangle),$$

for all p and q in H_o^+. Since f is continuous, so must g be.

Next we use the fact that f satisfies the additivity condition to extract information about g. Let θ_1 and θ_2 be any two non-negative real numbers. We claim that

$$g(\cosh(\theta_1 + \theta_2)) = g(\cosh\,\theta_1) + g(\cosh\,\theta_2). \tag{3.3.1}$$

To see this, let p be any point in H_o^+, and let s be any point in A such that $\overrightarrow{os}$ is a unit (spacelike) vector orthogonal to $\overrightarrow{op}$. (Certainly such points exist. It suffices to start with any unit vector u in $\overrightarrow{op}^{\perp}$, and take $s = o + u$.) Further, let points q and r be defined by:

$$\overrightarrow{oq} = (\cosh\,\theta_2)\,\overrightarrow{op} + (\sinh\,\theta_2)\,\overrightarrow{os} \tag{3.3.2}$$

$$\overrightarrow{or} = \cosh(\theta_1 + \theta_2)\,\overrightarrow{op} + \sinh(\theta_1 + \theta_2)\,\overrightarrow{os}. \tag{3.3.3}$$

Clearly, q and r belong to H_o^+ (since $\cosh^2\,\theta - \sinh^2\,\theta = 1$ for all θ). Multiplying the first of these equations by $\sinh\,(\theta_1 + \theta_2)$, the second by $\sinh\,\theta_2$, and then subtracting the second from the first, yields

$$\begin{aligned}
&\sinh(\theta_1 + \theta_2)\overrightarrow{oq} - (\sinh\,\theta_2)\,\overrightarrow{or} \\
&\quad = \ [\sinh(\theta_1 + \theta_2)\,\cosh\,\theta_2\, -\, \cosh(\theta_1 + \theta_2)\,\sinh\,\theta_2]\,\overrightarrow{op} \\
&\quad = \ [\sinh((\theta_1 + \theta_2)\, -\, \theta_2)]\,\overrightarrow{op}\ =\ (\sinh\,\theta_1)\,\overrightarrow{op}.
\end{aligned}$$

It follows that $\overrightarrow{oq}$ is between $\overrightarrow{op}$ and $\overrightarrow{or}$. (If $\theta_1 = 0 = \theta_2$, then $q = p = r$ and the three vectors are identical. Alternatively, if either $\theta_1 > 0$ or $\theta_2 > 0$, we can express $\overrightarrow{oq}$ in the form $\overrightarrow{oq} = a\,\overrightarrow{op} + b\,\overrightarrow{or}$, with non-negative coefficients

$$\begin{aligned} a &= \frac{\sinh\,\theta_1}{\sinh(\theta_1+\theta_2)} \\ b &= \frac{\sinh\,\theta_2}{\sinh(\theta_1+\theta_2)}.) \end{aligned}$$

So, by the additivity assumption,

$$g(\langle\overrightarrow{op},\overrightarrow{or}\rangle) \;=\; f(p,r) \;= f(p,q) + f(q,r) \;=\; g(\langle\overrightarrow{op},\overrightarrow{oq}\rangle) + g(\langle\overrightarrow{oq},\overrightarrow{or}\rangle). \quad (3.3.4)$$

But equations (3.3.2) and (3.3.3) (and the orthogonality of $\overrightarrow{op}$ and $\overrightarrow{os}$) imply that:

$$\begin{aligned} \langle\overrightarrow{op},\overrightarrow{or}\rangle &= \cosh(\theta_1+\theta_2) \\ \langle\overrightarrow{op},\overrightarrow{oq}\rangle &= \cosh\,\theta_2 \\ \langle\overrightarrow{oq},\overrightarrow{or}\rangle &= \cosh(\theta_1+\theta_2)\,\cosh\,\theta_2 \,-\, \sinh(\theta_1+\theta_2)\,\sinh\,\theta_2 \\ &= \cosh((\theta_1+\theta_2) \,-\, \theta_2) \;=\; \cosh\,\theta_1. \end{aligned}$$

Substituting these values into (3.3.4) yields our claim (3.3.1).

Our argument to this point has established that the composite map

$$g \circ \cosh : [0,\infty) \to \mathbb{R}$$

is additive. It follows by the continuity of g (and cosh) that there is a number K such that $g(\cosh(x)) = Kx$, for all x in $[0,\infty)$. (See proposition 3.3.3 below.) And now we are done. For given any points p and q in H^+_o, we need only substitute for x the number $\measuredangle(p,o,q)$ to reach the conclusion:

$$f(p,q) = g(\langle\overrightarrow{op},\overrightarrow{oq}\rangle) = g(\cosh\,\measuredangle(p,o,q)) = K\,\measuredangle(p,o,q).$$

Here we use the fact that $\overrightarrow{op}$ and $\overrightarrow{oq}$ are unit vectors and, so, $\cosh\,\measuredangle(p,o,q) = \langle\overrightarrow{op},\overrightarrow{oq}\rangle$. □

Now we turn to two lemmas that we need to complete the proof.

Proposition 3.3.2. *Let o and H^+_o be as in proposition 3.3.1. Given any four points p_1, q_1, p_2, q_2 in H^+_o with $\langle\overrightarrow{op}_1,\overrightarrow{oq}_1\rangle = \langle\overrightarrow{op}_2,\overrightarrow{oq}_2\rangle$, there is a temporal orientation preserving isometry $\varphi : A \to A$ of $(\boldsymbol{A}, \langle\ ,\ \rangle)$ such that $\varphi(o) = o$, $\varphi(p_1) = p_2$, and $\varphi(q_1) = q_2$.*

Proof. It will suffice for us to show that there is a vector space isomorphism $\Phi : V \to V$ preserving the Minkowski inner product such that

$$\begin{aligned} \Phi(\overrightarrow{op}_1) &= \overrightarrow{op}_2 \\ \Phi(\overrightarrow{oq}_1) &= \overrightarrow{oq}_2. \end{aligned}$$

For then the corresponding map $\varphi : A \to A$ defined by setting $\varphi(p) = o + \Phi(\overrightarrow{op})$ will be an isometry of $(\mathbf{A}, \langle , \rangle)$ that makes the correct assignments to o, p_1, and q_1:

$$\begin{array}{rclcrcl} \varphi(o) & = & o + \Phi(\overrightarrow{oo}) & = & o + \Phi(\mathbf{0}) & = & o + \mathbf{0} = o \\ \varphi(p_1) & = & o + \Phi(\overrightarrow{op}_1) & = & o + \overrightarrow{op}_2 & = & p_2 \\ \varphi(q_1) & = & o + \Phi(\overrightarrow{oq}_1) & = & o + \overrightarrow{oq}_2 & = & q_2. \end{array}$$

(Once again, recall the discussion preceding proposition 2.2.6.) Moreover, φ will preserve temporal orientation, i.e., for all p and q in A, if $\overrightarrow{pq}$ is timelike, then $\overrightarrow{\varphi(p)\,\varphi(q)}$ is co-oriented with $\overrightarrow{pq}$. (Why? Assume without loss of generality that $\overrightarrow{pq}$ is future-directed. (If not, we can work with $\overrightarrow{qp}$ instead.) So $\langle \overrightarrow{pq}, \overrightarrow{op}_1 \rangle > 0$ and, hence,

$$\langle \overrightarrow{\varphi(p)\,\varphi(q)}, \overrightarrow{op}_2 \rangle = \langle \Phi(\overrightarrow{pq}), \Phi(\overrightarrow{op}_1) \rangle = \langle \overrightarrow{pq}, \overrightarrow{op}_1 \rangle > 0.$$

Thus, $\overrightarrow{\varphi(p)\,\varphi(q)}$ is co-oriented with the future-directed vector $\overrightarrow{op}_2$. So $\overrightarrow{\varphi(p)\,\varphi(q)}$ is future-directed itself.)

We will realize Φ as a composition of two (Minkowski inner product preserving) vector space isomorphisms. The first will be a "boost" (or "timelike rotation") $\Phi_1 : V \to V$ that takes $\overrightarrow{op}_1$ to $\overrightarrow{op}_2$. The second will be a spatial rotation $\Phi_2 : V \to V$ that leaves $\overrightarrow{op}_2$ fixed, and takes $\Phi_1(\overrightarrow{oq}_1)$ to $\overrightarrow{oq}_2$. (Clearly, if these conditions are satisfied, then $(\Phi_2 \circ \Phi_1)(\overrightarrow{op}_1) = \overrightarrow{op}_2$ and $(\Phi_2 \circ \Phi_1)(\overrightarrow{oq}_1) = \overrightarrow{oq}_2$.) We consider Φ_1 and Φ_2 in turn.

If $p_1 = p_2$, we can take Φ_1 to be the identity map. Otherwise, the vectors $\overrightarrow{op}_1$ and $\overrightarrow{op}_2$ span a two-dimensional subspace W of V. In this case, we define Φ_1 by setting

$$\begin{array}{rcl} \Phi_1(\overrightarrow{op}_1) & = & \overrightarrow{op}_2 \\ \Phi_1(\overrightarrow{op}_2) & = & -\overrightarrow{op}_1 \,+\, 2\,\langle \overrightarrow{op}_1, \overrightarrow{op}_2 \rangle\, \overrightarrow{op}_2 \\ \Phi_1(w) & = & w \quad \text{for all } w \text{ in } W^{\perp}. \end{array}$$

(A linear map is uniquely determined by its action on the elements of a basis.) Thus, Φ_1 reduces to the identity on $W^{\perp}$, takes W to itself, and (within W) takes $\overrightarrow{op}_1$ to $\overrightarrow{op}_2$. Moreover, it preserves the Minkowski inner product. (Notice, in particular, that

$$\begin{array}{rcl} \langle \Phi_1(\overrightarrow{op}_1), \Phi_1(\overrightarrow{op}_2) \rangle & = & \langle \overrightarrow{op}_2, -\overrightarrow{op}_1 \,+\, 2\,\langle \overrightarrow{op}_1, \overrightarrow{op}_2 \rangle\, \overrightarrow{op}_2 \rangle \\ & = & -\langle \overrightarrow{op}_1, \overrightarrow{op}_2 \rangle + 2 \langle \overrightarrow{op}_1, \overrightarrow{op}_2 \rangle\, \langle \overrightarrow{op}_2, \overrightarrow{op}_2 \rangle = \langle \overrightarrow{op}_1, \overrightarrow{op}_2 \rangle, \end{array}$$

since $\langle \overrightarrow{op}_2, \overrightarrow{op}_2 \rangle = 1$.)

Next we turn to Φ_2. Since $\overrightarrow{op}_2$ is a unit timelike vector, it follow from proposition 3.1.1 that we can express $\Phi_1(\overrightarrow{oq}_1)$ and $\overrightarrow{oq}_2$ in the form

$$\Phi_1(\overrightarrow{oq}_1) = a\,\overrightarrow{op}_2 \,+\, u \tag{3.3.5}$$

$$\overrightarrow{oq}_2 = b\,\overrightarrow{op}_2 \,+\, v, \tag{3.3.6}$$

where u and v are spacelike vectors orthogonal to $\overrightarrow{op}_2$. Now we must have $a = b$ since, by our initial assumption that $\langle \overrightarrow{op}_1, \overrightarrow{oq}_1 \rangle = \langle \overrightarrow{op}_2, \overrightarrow{oq}_2 \rangle$,

$$a = \langle \overrightarrow{op}_2,\, \Phi_1(\overrightarrow{oq}_1) \rangle = \langle \Phi_1(\overrightarrow{op}_1),\, \Phi_1(\overrightarrow{oq}_1) \rangle = \langle \overrightarrow{op}_1,\, \overrightarrow{oq}_1 \rangle = \langle \overrightarrow{op}_2,\, \overrightarrow{oq}_2 \rangle = b.$$

Moreover, since $\Phi_1(\overrightarrow{oq}_1)$ and $\overrightarrow{oq}_2$ are both unit timelike vectors, it follows from (3.3.5) and (3.3.6) that

$$a^2 - \langle u,\, u \rangle \;=\; 1 \;=\; b^2 - \langle v,\, v \rangle.$$

So $\|u\| = \|v\|$.

Now the restriction of the Minkowski inner product to the three-dimensional subspace $(\overrightarrow{op}_2)^{\perp}$ is (negative) definite. So $(\overrightarrow{op}_2)^{\perp}$ together with that inner product is, essentially, just three-dimensional Euclidean space (conceived as an affine metric space). But given any two vectors in Euclidean space of the same length, there is a rotation that takes one to the other. Thus we can find a vector space isomorphism of $(\overrightarrow{op}_2)^{\perp}$ onto itself that preserves the induced inner product, and takes u to v. We can extend it to a vector space isomorphism $\Phi_2 : V \to V$ that preserves the Minkowskian inner product by simply adding the requirement that Φ_2 leave $\overrightarrow{op}_2$ fixed. This map serves our purposes because it takes $\Phi_1(\overrightarrow{oq}_1)$ to $\overrightarrow{oq}_2$, as required:

$$\Phi_2(\Phi_1(\overrightarrow{oq}_1)) = \Phi_2(a\,\overrightarrow{op}_2 \,+\, u) = a\,\Phi_2(\overrightarrow{op}_2) \,+\, \Phi_2(u) = b\,\overrightarrow{op}_2 \,+\, v = \overrightarrow{oq}_2.$$

□

The second lemma that we need to complete the proof of proposition 3.3.1 is the following.

Proposition 3.3.3. *Let $h : [0, \infty) \to \mathbb{R}$ be a map that that is additive and continuous. (Additivity here means that, for all x and y, $h(x+y) = h(x)+h(y)$.) Then, for all x,*

$$h(x) = Kx \tag{3.3.7}$$

where $K = h(1)$.

Proof. It follows by induction, of course, that

$$h(x_1 + x_2 + ... + x_n) = h(x_1) + h(x_2) + ... + h(x_n),$$

for all $n \geq 1$, and all $x_1, x_2, ..., x_n$ in $[0, \infty)$. Using just this condition (i.e., without appealing to the continuity of h), we can show that (3.3.7) holds for all *rational* x. The argument proceeds in three stages. First, for all $n \geq 1$, we have

$$h(n) \;=\; h(\underbrace{1 + ... + 1}_{n}) \;=\; \underbrace{h(1) + ... + h(1)}_{n} \;=\; n\; h(1) \;=\; K\, n.$$

(This also holds, trivially, if $n = 0$, since $h(0) = h(0 + 0) = h(0) + h(0)$, and so $h(0) = 0 = k\ 0$.) Next, for all $m \geq 1$,

$$h(1) \;=\; h(\underbrace{(1/m) + ... + (1/m)}_{m}) \;=\; \underbrace{h(1/m) + ... + h(1/m)}_{m} \;=\; m\; h(1/m)$$

and, so, $h(1/m) = (1/m)\ h(1) = K\ (1/m)$. If follows that, for all $n \geq 0$ and all $m \geq 1$,

$$\begin{aligned} h(n/m) &= h(\underbrace{(1/m) + ... + (1/m)}_{n}) = \underbrace{h(1/m) + ... + h(1/m)}_{n} \\ &= n\ h(1/m) = n\,(K/m) = K\,(n/m). \end{aligned}$$

Thus, as claimed, (3.3.7) holds for all *rational* x. If we *now* invoke continuity, we can extend the claim to all reals. This is clear since every real r in the interval $[0, \infty)$ can be realized as the limit of a sequence of rationals $\{q_i\}$ in that interval and, so,

$$h(r) = h(\lim q_i) = \lim h(q_i) = \lim (Kq_i) = K \lim q_i = K\, r.$$

□

It turns out that the proposition fails if the assumption of continuity is dropped. There exist maps $h : \mathbb{R} \to \mathbb{R}$ (and so also maps $h : [0, \infty) \to \mathbb{R}$) that are additive, but not continuous. (See, for example, Gelbaum and Olmsted [2, p. 33].) For such maps, there *cannot* exist a constant K such that $h(x) = Kx$ for all x (since the latter condition implies continuity).

Problem 3.3.1. *Formulate and prove a corresponding uniqueness result for Euclidean angular measure.*

3.4 Uniqueness Results for the Relative Simultaneity Relation

We noted in section 3.2, when discussing the decomposition of vectors at a point into their "temporal" and "spatial" components relative to a timelike vector there, that we were taking for granted the standard identification of relative simultaneity with orthogonality. Here we return to consider the justification of that identification.

In what follows, let $(\mathbf{A}, \langle\ ,\ \rangle)$ be an n-dimensional ($n \geq 2$) Minkowskian space (with underlying point set A and vector space $\mathbf{V}$) endowed with a temporal orientation.

Consider a timelike line L in A. What pairs of points p, q in A should qualify as being "simultaneous relative to L"? That is the question we are considering. The standard answer, once again, is that they should do so precisely if $\overrightarrow{pq} \perp L$.

In traditional discussions of relative simultaneity, the standard answer is often cast in terms of "epsilon" values. The connection is easy to see. Let p be any point that is not on our timelike line L. Then (as we know, for example, from problem 3.1.5), there exist unique points r and s on L (distinct from one another) such that $\overrightarrow{rp}$ and $\overrightarrow{ps}$ are future-directed null vectors. (See figure 3.4.1.) Now let q be any point on L. (We think of it as a candidate for being judged simultaneous with p relative to L.) Then $\overrightarrow{rq} = \epsilon\, \overrightarrow{rs}$ for some $\epsilon \in \mathbb{R}$. Hence, since $\overrightarrow{ps}$ and $\overrightarrow{pr}$ are null,

$$0 = \langle \overrightarrow{ps}, \overrightarrow{ps} \rangle = \langle \overrightarrow{pr} + \overrightarrow{rs}, \overrightarrow{pr} + \overrightarrow{rs} \rangle = 2\langle \overrightarrow{pr}, \overrightarrow{rs} \rangle + \langle \overrightarrow{rs}, \overrightarrow{rs} \rangle.$$

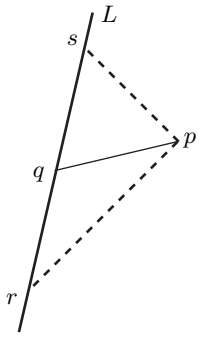


Figure 3.4.1: The $\epsilon = \frac{1}{2}$ characterization of relative simultaneity: p and q are simultaneous relative to L iff q is midway between r and s.

It follows that

$$\langle \overrightarrow{pq}, \overrightarrow{rs} \rangle \;=\; \langle \overrightarrow{pr}+\overrightarrow{rq},\, \overrightarrow{rs} \rangle \;=\; \langle \overrightarrow{pr}+\epsilon\overrightarrow{rs},\, \overrightarrow{rs} \rangle \;=\; \langle \overrightarrow{pr}, \overrightarrow{rs} \rangle + \epsilon \langle \overrightarrow{rs}, \overrightarrow{rs} \rangle \;=\; (\epsilon - \frac{1}{2}) \langle \overrightarrow{rs}, \overrightarrow{rs} \rangle.$$

Thus,

$$\epsilon = \frac{1}{2} \iff \overrightarrow{pq} \perp \overrightarrow{rs}. \tag{3.4.1}$$

So the standard (orthogonality) relation of relative simultaneity in special relativity may equally well be described as the "$\epsilon = \frac{1}{2}$" relation of relative simultaneity.

Yet another equivalent formulation involves the "one-way speed of light". Suppose a light ray travels from r to p with speed c_+ relative to L, and from p to s with speed c_- relative to L. We saw in section 3.2 that *if* one adopts the standard criterion of relative simultaneity, then it follows that $c_+ = c_-$. (Indeed, in this case, both c_+ and c_- turn out to be 1.) The converse is true as well. For if $c_+ = c_-$, then, as determined relative to L, it should take as much time for light to travel from r to p as from p to s. And in that case, a point q on L should be judged simultaneous with p relative to L precisely if it is midway between r and s. So we are led, once again, to the "$\epsilon = \frac{1}{2}$" relation of relative simultaneity.

Now is adoption of the standard relation a matter of convention, or is it in some significant sense forced on us?

There is a large literature devoted to this question. (Classic statements of the conventionalist position can be found in Reichenbach [9] and Grünbaum [4]. Grünbaum has recently responded to criticism of his views in [3]. An overview of the debate with many references can be found in Janis [7].) It is not my purpose to review it here, but I do want to draw attention to certain remarks of Howard Stein [12, pp. 153-4] that seem to me particularly insightful.

He makes the point that determinations of conventionality require a context.

> There are really two distinct aspects to the issue of the "conventionality" of Einstein's concept of relative simultaneity. One may assume the position of Einstein himself at the outset of his investigation – that is, of one confronted by a problem, trying to find a theory that will deal with it satisfactorily; or one may assume the position of (for instance) Minkowski – that is, of one confronted with a theory already developed, trying to find its most adequate and instructive formulation.

The problem Einstein confronted was (in part) that of trying to account for our apparent inability to detect any motion of the earth with respect to the "aether". A crucial element of his solution was the proposal that we think about simultaneity a certain way (i.e., in terms of the "$\epsilon = \frac{1}{2}$ criterion"), and resolutely follow through on the consequences of doing so. Stein emphasizes just how different that proposal looks when we consider it, not from Einstein's initial position, but rather from the vantage point of the finished theory, i.e., relativity theory conceived as an account of invariant spacetime structure.

> [For] Einstein, the question (much discussed since Reichenbach) whether the evidence really shows that that the speed of light *must* be regarded as the same in all directions and for all observers is not altogether appropriate. A person devising a theory does not have the responsibility, at the outset, of showing that the theory being developed is the only possible one given the evidence. [But] once Einstein's theory had been developed, and had proved successful in dealing with all relevant phenomena, the case was quite transformed; for we know that *within* this theory, there is only one "reasonable" concept of simultaneity (and in terms of that concept, the velocity of light is indeed as Einstein supposed); therefore an alternative will only present itself if someone succeeds in constructing, not simply a different empirical criterion of simultaneity, but an essentially different (and yet viable) theory of electrodynamics of systems in motion. No serious alternative theory is in fact known.
> (emphasis in original)

Our goal in the remainder of this section is to formulate two elementary uniqueness results, closely related to one another, that capture the sense in which "there is only one 'reasonable' concept of (relative) simultaneity" within the framework of Minkowski spacetime. (Many propositions of this form can be found in the literature. See Budden [1] for a review. Ours are intended only as examples.)

Once again, let L be a timelike line in A, and let Sim_L be the standard relation of simultaneity relative to L. (So $(p, q) \in Sim_L$ iff $\overrightarrow{pq} \perp L$, for all p and q in A.) Further, let S be an arbitrary two-place relation on A that we regard as a candidate for the relation of "simultaneity relative to L". The first uniqueness result asserts that if S satisfies three conditions, including an invariance condition, then $S = Sim_L$. The first two conditions are straightforward.

(S1) S is an equivalence relation (i.e., S is reflexive, symmetric, and transitive).

(S2) For all points $p \in A$, there is a unique point $q \in L$ such that $(p, q) \in S$.

If S satisfies (S1), it has an associated family of equivalence classes. We can think of them as "simultaneity slices" (as determined relative to L). Then (S2) asserts that every simultaneity slice intersects L in exactly one point. Note that if $S = Sim_L$, then (S1) and (S2) are satisfied. For in this case, the equivalence classes associated with S are hyperplanes orthogonal to L, and these clearly intersect L in exactly one point. (A *hyperplane* is a subspace of V whose dimension is one less than that of V. So if V is four-dimensional, hyperplanes are three-dimensional subspaces.)

The third, invariance condition is intended to capture the requirement that S is determined by, or definable in terms of, the background geometric structure of Minkowski spacetime and by L itself. The one subtle point here is whether temporal orientation is taken to count as part of that background geometric structure or not. Let's assume for the moment that it does not.

Let $\varphi\colon A \to A$ be an isometry of $(\mathbf{A}, \langle\ ,\ \rangle)$, i.e., an affine space isomorphism that preserves the Minkowski inner product $\langle\ ,\ \rangle$. We will say it is an *L-isometry* if, in addition, it preserves L, i.e., if, for all points p in A, $\ p \in L \iff \varphi(p) \in L$. We will be interested in L-isometries of three types.

(a) translations along L

In this case, there exist points r, s on L such that, for all p, $\varphi(p) = p + \overrightarrow{rs}$.

(b) isometries that leave L fixed

In this case, $\varphi(p) = p$ for all $p \in L$, and the restriction of φ to any hyperplane orthogonal to L is a reflection or rotation.

(c) temporal reflections with respect to hyperplanes orthogonal to L.

In this case, there is a point o on L such that, for all p, if $p = o + v + w$, where v is parallel to L and w is orthogonal to it (the representation is unique by proposition 3.1.1), $\varphi(p) = o - v + w$.

(It turns out that every L-isometry can be expressed as a composition of L-isometries of these three basic types. But that fact will not be needed in what follows.) We will say that our two-place relation S is *L-invariant* if it is preserved under all L-isometries, i.e., if for all L-isometries $\varphi : A \to A$, and all points $p, q \in A$,

$$(p,\ q) \in S \iff (\varphi(p),\ \varphi(q)) \in S. \tag{3.4.2}$$

We can now formulate the first uniqueness result. (It is a close variant of one presented in Hogarth [6].)

Proposition 3.4.1. *Let L be a timelike line, and let S be a two-place relation on A that satisfies conditions (S1) and (S2), and is L-invariant. Then $S = Sim_L$.*

Proof. Assume the three conditions hold. For every point $p \in A$, let $f(p)$ be the unique point q on L such that $\overrightarrow{pq} \perp L$. ($q$ is the intersection of L with the hyperplane through p orthogonal to L.) So, clearly, the following conditions hold.

(i) For all $p \in A, (p, f(p)) \in Sim_L$.
(ii) For all $p, p' \in A$, $(p, p') \in Sim_L \iff f(p) = f(p')$.

We claim the following condition holds as well.

(iii) For all $p \in A$, $(p, f(p)) \in S$.

For suppose p is a point in A. By (S2), there is a unique point q on L such that $(p, q) \in S$. Now let $\varphi : A \to A$ be a temporal reflection with respect to the hyperplane orthogonal to L that contains p and $f(p)$. (See figure 3.4.2.) Then φ is an L-isometry, $\varphi(p) = p$, and $\overrightarrow{f(p)\,\varphi(q)} = \overrightarrow{q\,f(p)}$ or, equivalently, $\varphi(q) = f(p) + \overrightarrow{q\,f(p)}$. Since $(p, q) \in S$, it follows by L-invariance of S that $(p, \varphi(q)) = (\varphi(p), \varphi(q)) \in S$. But q and $\varphi(q)$ are both on L. Hence, by the uniqueness condition in (S2), $\varphi(q) = q$. Therefore, $\overrightarrow{f(p)\,q} = \overrightarrow{f(p)\,\varphi(q)} = \overrightarrow{q\,f(p)}$. So $\overrightarrow{f(p)\,q} = \mathbf{0}$ and, therefore, $q = f(p)$. Thus $(p, f(p)) = (p, q) \in S$. So we have (iii).

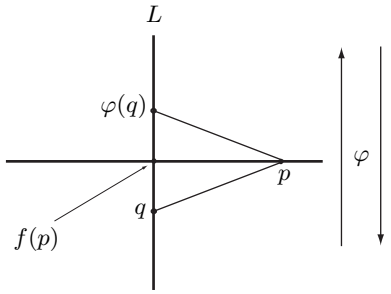


Figure 3.4.2: In the proof of proposition 3.4.1, φ is a temporal reflection with respect to a hyperplane orthogonal to L that contains p and $f(p)$. It leaves p fixed, but moves q.

Our conclusion now follows easily from (i), (ii), and (iii). To see this, assume first that $(p, p') \in S$. By (iii), we have $(p', f(p')) \in S$. So, by the transitivity of S, $(p, f(p')) \in S$. But we also have $(p, f(p)) \in S$, by (iii) again. Since $f(p)$ and $f(p')$ are both on L, it follows from the uniqueness condition in (S2) that $f(p) = f(p')$. Hence, by (ii), $(p, p') \in Sim_L$. Thus we have $S \subseteq Sim_L$. Assume, conversely, that $(p, p') \in Sim_L$. It follows, by (ii), that $f(p) = f(p')$. Hence, by (iii), we have $(p, f(p)) \in S$ and $(p', f(p)) = (p', f(p')) \in S$. Hence, since S is reflexive and transitive, $(p, p') \in S$. Thus $Sim_L \subseteq S$, and we are done. □

Notice that we have not used the full force of L-invariance in our proof. We have only used the fact that S is preserved under L-isometries of type (c).

Suppose now that we *do* want to consider temporal orientation as part of the background structure that may play a role in the determination of S. Then we need to recast the invariance condition. Let us say that an L-isometry $\varphi : A \to A$ is an *$(L, \uparrow)$-isometry* if it (also) preserves temporal orientation, i.e., if for all timelike vectors $\overrightarrow{pq}$, $\overrightarrow{\varphi(p)\,\varphi(q)}$ is co-oriented with $\overrightarrow{pq}$. And let us say that S is

(L, $\uparrow$)-invariant if it is preserved under all (L, $\uparrow$)-isometries. (So, to be (L, $\uparrow$)-invariant, S must be preserved under all L-isometries of type (a) and (b), but need not be preserved under those of type (c).)

(L, $\uparrow$)-invariance is a weaker condition than L-invariance and, in fact, is too weak to deliver the uniqueness result we want. There are many two-place relations S on A other than Sim_L that satisfy (S1), (S2), and are (L, $\uparrow$)-invariant. They include, for example, ones whose associated "simultaneity slices" are "flat cones" (see figure 3.4.3) that are preserved under L-isometries of type (a) and (b), but not (c).

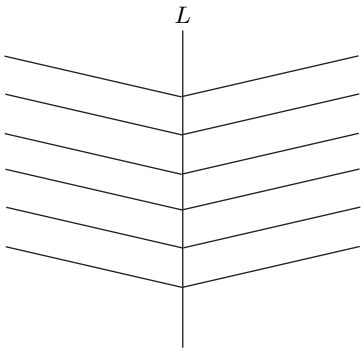


Figure 3.4.3: The "flat cones" displayed are the simultaneity slices associated with a two-place relation S that satisfies conditions (S1) and (S2), and is (L, $\uparrow$)-invariant (but not L-invariant).

But we can still get a uniqueness result if we change the set-up slightly, and think of simultaneity as determined, not relative to individual timelike lines, but, rather, relative to families of parallel timelike lines. Let us officially take a *frame* in A to be a set of parallel timelike lines $\mathcal{L}$ that is maximal in the sense that every point in A falls on one (and only one) of them. With hardly any work, we can recast our previous notions in terms of frames rather than lines.

In what follows, let $\mathcal{L}$ be some fixed frame. Given any two lines L and L' in $\mathcal{L}$, $Sim_L = Sim_{L'}$. (Since L and L' are parallel, any vector orthogonal to one must be orthogonal to the other.) So we can, without ambiguity, make reference to $Sim_{\mathcal{L}}$ (the standard relation of simultaneity relative to $\mathcal{L}$). Let $\varphi: A \to A$ be an isometry of $(\mathbf{A}, \langle\ ,\ \rangle)$. We say it is an *$\mathcal{L}$-isometry* if, for all L in $\mathcal{L}$, the line $\varphi[L]$ is also in $\mathcal{L}$. And we say that it it is an ($\mathcal{L}$, $\uparrow$)-*isometry* if, in addition, it preserves temporal orientation.

If L is a line in $\mathcal{L}$, then the set of $\mathcal{L}$-isometries certainly includes all L-isometries of types (a), (b), and (c) above. But it includes, in addition, (d) translations taking L to some other line in $\mathcal{L}$, and (e) isometries that leave fixed the points on some line in $\mathcal{L}$ other than L. If we restrict attention to ($\mathcal{L}$, $\uparrow$)-isometries, we lose maps of type (c), but we retain those of types (a), (b), (d), and (e). Invariance under this larger class *is* sufficient to drive a uniqueness result.

We say (of course) that S is $\mathcal{L}$-*invariant* if it is preserved under all $\mathcal{L}$-isometries, and $(\mathcal{L}, \uparrow)$-invariant if it is preserved under all $(\mathcal{L}, \uparrow)$-isometries. Our second uniqueness result comes out as follows. (It is closely related to propositions in Spirtes [11], Stein [12], and Budden [1].)

Proposition 3.4.2. *Let $\mathcal{L}$ be a frame, and let S be a two-place relation on A. Suppose S satisfies (S1) and, for some L in $\mathcal{L}$, satisfies (S2). Further, suppose S is $(\mathcal{L}, \uparrow)$-invariant. Then $S = Sim_{\mathcal{L}}$.*

Proof. The proof is very much like that of proposition 3.4.1. Assume S satisfies the hypotheses of the proposition. Then there is a line L in $\mathcal{L}$ such that S satisfies (S2). Just as before, for every point $p \in A$, let $f(p)$ be the unique point q on L such that $\overrightarrow{pq} \perp L$. ($q$ is the intersection of L with the hyperplane through p that is orthogonal to L.) Once again, the following three conditions hold.

(i) For all $p \in A, (p, f(p)) \in Sim_L$.
(ii) For all $p, p' \in A$, $(p, p') \in Sim_L \iff f(p) = f(p')$.
(iii) For all $p \in A$, $(p, f(p)) \in S$.

The first two are immediate. Only the third requires argument. But once we have verified (iii), we will be done. For the rest of the argument – that (i), (ii), and (iii) collectively imply $S = Sim_L$ (and, so, $S = Sim_{\mathcal{L}}$) – goes through intact.

Let p be a point in A. By (S2), there is a unique point q on L such that

$$(p, q) \in S. \tag{3.4.3}$$

If $p \in L$, then $f(p) = p$, and so $(p, f(p)) \in S$ automatically (because S is reflexive). So we may assume that $p \notin L$. Let $\varphi_1 \colon A \to A$ be an L-isometry of type (b) – either a reflection or rotation – that leaves L intact and flips (about L) the timelike two-plane containing L and p. (See figure 3.4.4.) So we have

$$\begin{aligned} \varphi_1(q) &= q \\ \varphi_1(p) &= f(p) + \overrightarrow{p\,f(p)}. \end{aligned}$$

Next, let $\varphi_2 \colon A \to A$ be a translation that takes q to p. It is an $(\mathcal{L}, \uparrow)$-isometry of type (d). The action of φ_2 on any point r in A is given by $\varphi_2(r) = r + \overrightarrow{qp}$.

The composed map $(\varphi_2 \circ \varphi_1)$ is an $(\mathcal{L}, \uparrow)$-isometry. So, it follows from (3.4.3) and the $(\mathcal{L}, \uparrow)$-invariance of S that $((\varphi_2 \circ \varphi_1)(p), (\varphi_2 \circ \varphi_1)(q)) \in S$. But

$$\begin{aligned} (\varphi_2 \circ \varphi_1)(p) &= (f(p) + \overrightarrow{p\,f(p)}) + \overrightarrow{qp} \;=\; f(p) + \overrightarrow{q\,f(p)} \\ (\varphi_2 \circ \varphi_1)(q) &= \varphi_2(q) \;=\; p. \end{aligned}$$

Thus, we have $(f(p) + \overrightarrow{q\,f(p)},\, p) \in S$. But $f(p) + \overrightarrow{q\,f(p)}$ is a point on L (since q and $f(p)$ are). So by the uniqueness condition in (S2) (and by (3.4.3) and the symmetry of S), it follows that $f(p) + \overrightarrow{q\,f(p)} = q$. But this condition implies that $f(p) = q$. (Recall proposition 2.2.1.) So (by (3.4.3)), we may conclude that $(p,\, f(p)) \in S$. This gives us (iii), as needed. □

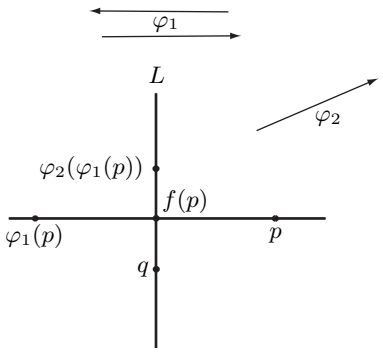


Figure 3.4.4: In the proof of proposition 3.4.2, φ_1 is a reflection or rotation that keeps L intact and flips (about L) the timelike two-plane containing L and p; φ_2 is a translation that takes q to p.

The move from proposition 3.4.1 to proposition 3.4.2 involves a trade-off. We drop the requirement that S be invariant under maps of type (c), but add the requirement that it be invariant under those of type (d) and (e).

Problem 3.4.1. *In our proof of proposition 3.4.2, we did not use the full strength of the assumption that S is $(\mathcal{L}, \uparrow)$-invariant. We only used the fact that S is invariant under $\mathcal{L}$-isometries of types (b) and (d). Show with an alternate proof that it suffices to assume that it is invariant under all $\mathcal{L}$-isometries of type (e).*

4 From Minkowskian Geometry to Hyperbolic Plane Geometry

In section 4.1, we first present Tarski's axiomatization of first-order Euclidean plane geometry, and a variant axiomatization of first-order hyperbolic (i.e., Lobatchevskian) plane geometry. Then we formulate (without proof) parallel completeness theorems for the two axiom systems. The second completeness theorem will be formulated in terms of the Klein-Beltrami model for hyperbolic geometry. (This brief excursion into the metamathematics of plane geometry should be of some interest in its own right.) In section 4.2 we establish the connection with Minkowskian geometry. There we use a three-dimensional Minkowskian space to give a second, more intuitive model for hyperbolic plane geometry and show that it is isomorphic to the Klein-Beltrami model. (In doing so we gain a sense of "where the latter comes from".)

4.1 Tarski's Axioms for first-order Euclidean and Hyperbolic Plane Geometry

We start with an axiomatization for Euclidean geometry given by Alfred Tarski [14]. It is formulated in a formal first-order language L with two relation symbols: a three-place relation symbol 'B' and a four-place relation symbol 'C'. Intuitively, we think of '$Bxyz$' as asserting that the three points x, y, z are collinear with y between x and z. (The case where y is identical with x or z is allowed.) We think of '$Cxyzw$' as asserting that the distance between x and y equals that between z and w.

In this formalization of geometry, we work with only one type of object, namely points. All statemement about "lines", "angles", and other geometric objects are recast in terms of statements about points that determine those lines, angles, and other objects. The formalization is "first-order" in that, intuitively, we allow quantification over points, but not over sets of points, or sets of sets of points, and so forth. One could add to L a six place relation symbol 'A' and understand '$Axyzuvw$' to mean that (i) the points x, y, z are distinct, (ii) the points u, v, w are distinct, and (iii) the angle $\angle(x, y, z)$ is congruent to the angle $\angle(u, v, w)$. But Tarski chose not to do so. No loss is involved because one can define congruence of angular measure in terms of 'B' and 'C'. The axioms are the following.

(1) Identity Axiom for Betweenness

$(\forall x)(\forall y)(Bxyx \rightarrow x = y)$

(2) Transitivity Axiom for Betweenness

$(\forall x)(\forall y)(\forall z)(\forall w)(Bxyz \ \& \ Byzw \rightarrow Bxyw)$

(3) Connectivity Axiom for Betweenness

$(\forall x)(\forall y)(\forall z)(\forall w)(Bxyz \ \& \ Bxyw \ \& \ x \neq y \rightarrow (Bxzw \ \vee \ Bxwz))$

(4) Reflexivity Axiom for Congruence

$(\forall x)(\forall y)Cxyyx$

(5) Identity Axiom for Congruence

$(\forall x)(\forall y)(\forall z)(Cxyzz \rightarrow x = y)$

(6) Transitivity Axiom for Congruence

$(\forall x)(\forall y)(\forall z)(\forall u)(\forall v)(\forall w)(Cxyzu \ \& \ Cxyvw \rightarrow Czuvw)$

(7) Pasch's Axiom

$(\forall x)(\forall y)(\forall z)(\forall u)(\forall v)(Bxyz \ \& \ Buzv \rightarrow (\exists w)(Bxwv \ \& \ Buyw))$

(8) Euclid's Axiom (*This is Tarski's name for the axiom, but it does not, in fact, appear in Euclid.*)

$(\forall x)(\forall y)(\forall z)(\forall u)(\forall v)(Bxyz \ \& \ Buyv \ \& \ x \neq y$
$\rightarrow (\exists w)(\exists t)(Bxvw \ \& \ Bxut \ \& \ Btzw))$

(9) Five Segment Axiom

$(\forall x_1)(\forall x_2)(\forall y_1)(\forall y_2)(\forall z_1)(\forall z_2)(\forall u_1)(\forall u_2)$
$(Cx_1y_1x_2y_2 \ \& \ Cy_1z_1y_2z_2 \ \& \ Cx_1u_1x_2u_2 \ \& \ Cy_1u_1y_2u_2$
$\& \ Bx_1y_1z_1 \ \& \ Bx_2y_2z_2 \ \& \ x_1 \neq y_1 \ \rightarrow \ Cz_1u_1z_2u_2)$

(10) Axiom of Segment Construction

$(\forall x)(\forall y)(\forall u)(\forall v)(\exists z)(Bxyz \ \& \ Cyzuv)$

(11) Lower Dimension Axiom

$(\exists x)(\exists y)(\exists z)(\neg Bxyz \ \& \ \neg Byzx \ \& \ \neg Bzxy)$

(12) Upper Dimension Axiom

$(\forall x)(\forall y)(\forall z)(\forall u)(\forall v)(Cxuxv \ \& \ Cyuyv \ \& \ Czuzv \ \& \ u \neq v$
$\rightarrow (Bxyz \vee Byzx \vee Bzxy))$

(13) Completeness Axiom(s)

Let ϕ be a well-formed formula in L in which the variables $x, v, w, ...$ occur free, but not y, z, or u. Further let ψ be a well-formed formula in L in which the variables $y, v, w, ...$ occur free, but not x, z, or u. Then the following is an axiom:

$(\forall v)(\forall w)...[(\exists z)(\forall x)(\forall y)(\phi \ \& \ \psi \rightarrow Bzxy)$
$\rightarrow (\exists u)(\forall x)(\forall y)(\phi \ \& \ \psi \rightarrow Bxuy)]$

Problem 4.1.1. *Exhibit a sentence ϕ_{par} in the language L that captures the "parallel postulate", the assertion that given a line L_1 and a point p not on L_1, there is a unique line L_2 that contains p and does not intersect L_1.*

We consider three theories: T_{Abs} (absolute geometry), T_{Euc} (Euclidean geometry), and T_{Hyp} (hyperbolic geometry). Let ϕ_{Euc} be Euclid's axiom. Then we take

$$\begin{aligned} T_{Abs} &= \text{the set of all axioms listed above except for Euclid's axiom, i.e, axioms} \\ &\quad (1)\text{–}(7), (9)\text{–}(12), \text{and all instances of schema (13).} \\ T_{Euc} &= T_{Abs} \cup \{\phi_{Euc}\} \\ T_{Hyp} &= T_{Abs} \cup \{\neg\phi_{Euc}\}. \end{aligned}$$

We also consider, in turn, two interpretations of the language L: $\mathbf{E}$ and $\mathbf{K}$. The first is the "standard interpretation" for T_{Euc}. For the domain $|\mathbf{E}|$ of $\mathbf{E}$ we take $\mathbb{R}^2$. For the assignments to 'B' and 'C' we take the usual relations of betweenness and congruence in Euclidean geometry:

$$\begin{aligned} \mathbf{E}(B) &= \{(p,q,r) \in |\mathbf{E}|^3 \colon \overrightarrow{pq} = k\,\overrightarrow{pr} \text{ for some } k \text{ with } 0 \leq k \leq 1\} \\ \mathbf{E}(C) &= \{(p,q,r,s) \in |\mathbf{E}|^4 \colon d(p,q) = d(r,s)\}. \end{aligned}$$

(Here, and in what follows, $d \colon |\mathbf{E}| \times |\mathbf{E}| \to \mathbb{R}$ is the standard Euclidean distance function. If $p = (p_1, p_2)$ and $q = (q_1, q_2)$, $d(p,q) = [(q_1 - p_1)^2 + (q_2 - p_2)^2]^{\frac{1}{2}}$.) The basic completeness theorem for T_{Euc} is the following.

Proposition 4.1.1. *For all sentences ϕ in L,*

$$T_{Euc} \vdash \phi \iff \phi \text{ is true under interpretation in } \boldsymbol{E}.$$

(Here we assume that we have in place some (sound and complete) derivation system for first-order logic. For any set of sentences T in L, and any individual sentence ϕ in L, we understand $T \vdash \phi$ to be the relation that holds if there is a derivation of ϕ from T in that derivation system.) To prove the soundness (left to right) half of the proposition, it suffices to check that all the sentences in T_{Euc} are true in E. That is relatively straightforward. It is the converse direction that requires work. (A sketch of the proof can be found in the Tarski article cited at the beginning of the section.)

Our second interpretation of L is the Klein-Beltrami model for T_{Hyp}. Its domain $|\mathbf{K}|$ is the interior of the unit circle in $\mathbb{R}^2$, i.e., the set $\{p \in \mathbb{R}^2 \colon d(p, \mathbf{0}) < 1\}$, where $\mathbf{0} = (0,0)$. The assignment to 'B' is just the restriction of the Euclidean betweenness relation to $|\mathbf{K}|$, i.e.,

$$\mathbf{K}(B) = \{(p,q,r) \in |\mathbf{K}|^3 \colon \overrightarrow{pq} = k\,\overrightarrow{pr} \text{ for some } k \text{ with } 0 \leq k \leq 1\}.$$

To specify the assignment to 'C', we first need to define the *hyperbolic distance function* $d_K \colon |\mathbf{K}| \times |\mathbf{K}| \to \mathbb{R}$ on $|\mathbf{K}|$. Given two points p and q in $|\mathbf{K}|$, if $p = q$, we set $d_K(p,q) = 0$. If $p \neq q$, the line determined by p and q intersects the unit circle in two points. If r is the point on the circle so situated that p is between r and q, and s is the one such that q is between p and s (see figure 4.1.1), then we take the *cross-ratio* of p and q to be the fraction

$$CR(p,q) = \frac{d(p,r)\,d(q,s)}{d(p,s)\,d(q,r)}$$

and set $d_K(p,q) = -\frac{1}{2}\log(CR(p,q))$. The assignment to '$C$' in the Klein-Beltrami model is the congruence relation determined by d_K, i.e.,

$$\mathbf{K}(C) = \{(p,q,r,s) \in |\mathbf{K}|^4 \colon d_K(p,q) = d_K(r,s)\}.$$

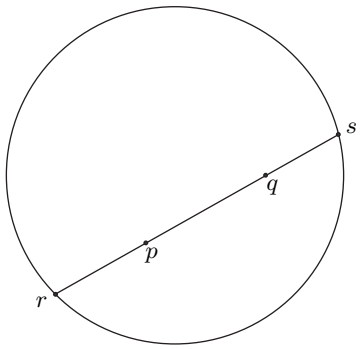


Figure 4.1.1: The Klein-Beltrami model of hyperbolic plane geometry.

Let's first observe that the function d_K has certain properties that make it at least a plausible candidate for a "distance function" on $|\mathbf{K}|$.

(1) For all p and q in $|\mathbf{K}|$, $d_K(p,q) \geq 0$, and $d_K(p,q) = 0$ iff $p = q$.

This follows from the fact that if $p \neq q$, then $CR(p,q) < 1$ and, hence, $\log(CR(p,q)) < 0$.

(2) For all p_1, p_2, and p_3 in $|\mathbf{K}|$, if the points are collinear with p_2 between p_1 and p_3,

$$d_K(p_1,p_3) = d_K(p_1,p_2) + d_K(p_2,p_3).$$

This too is easy to check. If $p_2 = p_1$ or $p_2 = p_3$, the assertion is trivial. Otherwise if follows from the linearity of the log function. (Of course, since the three points are collinear, the extremal points on the boundary of the circle that figure in the definition in the cross ratio are the same for the three pairs $\{p_1,p_2\}, \{p_2,p_3\}$, and $\{p_1,p_3\}$.)

(3) If a point p in $|\mathbf{K}|$ is held fixed, and a second point q "moves to the boundary of the circle along a fixed line segment", then, in this limit, $d_K(p,q)$ goes to ∞.

In the limit under consideration, the extremal points r and s in the definition of $CR(p,q)$ are held fixed, and q converges to s. So, in the limit, $CR(p,q)$ goes to 0, and $\log(CR(p,q))$ goes to $(-\infty)$. This gives us (3).

It would be nice to add to the list the assertion that d_K satisfies the triangle inequality. Certainly it does. But it is not so easy to prove this directly. In

the next section we will show, in a sense, where the odd-looking function d_K "comes from" and, in so doing, it will become more or less obvious that it satisfies the triangle inequality. The basic completeness theorem for T_{Hyp} (the proof of which is similar to that for T_{Euc}) is the following.

Proposition 4.1.2. *For all sentences ϕ in L,*

$$T_{Hyp} \vdash \phi \iff \phi \text{ is true under interpretation in } \boldsymbol{K}.$$

To prove the soundness (left to right) half of the proposition, one needs to check that all the sentences in T_{Hyp} are true in K. Note, for example, that Euclid's axiom is not true under interpretation in K (and so its negation is true in that interpretation). To see this, it suffices to consider the counterexample in figure 4.1.2. Again, it is the converse direction that requires work. (A proof can be found in Szmielew [13].)

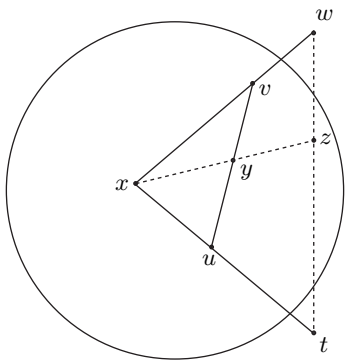


Figure 4.1.2: "Euclid's axiom" is false under interpretation in the the Klein-Beltrami model. (The points w and t – whose existence is required for the assertion to be true – do not lie within the domain of the model.)

We note a few simple consequences of the completeness theorems before moving on.

Corollary 4.1.1. *For all sentences ϕ in L,*

(i) $T_{Abs} \vdash \phi \iff \phi$ *is true in both* $\boldsymbol{E}$ *and* $\boldsymbol{K}$.

(ii) $T_{Abs} \vdash (\phi \leftrightarrow \phi_{Euc}) \iff \phi$ *is true in* $\boldsymbol{E}$*, but false in* $\boldsymbol{K}$.

(iii) $T_{Abs} \vdash (\phi \leftrightarrow \neg\phi_{Euc}) \iff \phi$ *is false in* $\boldsymbol{E}$*, but true in* $\boldsymbol{K}$.

Proof. The claims all follow by simple principles of first-order logic. For (i), assume first that $T_{Abs} \vdash \phi$. Then $T_{Euc} \vdash \phi$ and $T_{Hyp} \vdash \phi$ (since T_{Abs} is a subset of both T_{Euc} and T_{Hyp}). So it follows by propositions 4.1.1 and 4.1.2 that ϕ is true in **E** and true in **K**. Conversely, assume ϕ is true in both interpretations. Then $T_{Euc} \vdash \phi$ and $T_{Hyp} \vdash \phi$ by those propositions, i.e., $T_{Abs} \cup \{\phi_{Euc}\} \vdash \phi$ and $T_{Abs} \cup \{\neg\phi_{Euc}\} \vdash \phi$. So (by the "deduction theorem" for our derivation system),

$T_{Abs} \vdash (\phi_{Euc} \to \phi)$ and $T_{Abs} \vdash (\neg\phi_{Euc} \to \phi)$. Hence (by basic principles of sentential logic), $T_{Abs} \vdash \phi$.

For (ii) assume first that $T_{Abs} \vdash (\phi \leftrightarrow \phi_{Euc})$. Then, by part (i), $(\phi \leftrightarrow \phi_{Euc})$ is true in both **E** and **K**. So ϕ and ϕ_{Euc} have the same truth value in the two interpretations. But ϕ_{Euc} is true in **E** and false in **K**. So ϕ too is true in **E** and false in **K**. Conversely, assume that ϕ is true in **E**, but false in **K**. Then, since ϕ_{Euc} has those same truth values in **E** and **K**, the biconditional $(\phi \leftrightarrow \phi_{Euc})$ must be true in both interpretations. Hence, by part (i) again, it follows that $T_{Abs} \vdash (\phi \leftrightarrow \phi_{Euc})$.

The proof for (iii) is very much the same as that for (ii). □

It follows from the corollary, of course, that in our formalization of Euclidean geometry we can substitute for ϕ_{Euc} any sentence in L that is true in **E**, but false in **K**. In particular, we can substitute the parallel postulate. (And in our formalization of hyperbolic geometry we can substitute for $\neg\phi_{Euc}$ any sentence in L that is true in **K**, but false in **E**.)

4.2 The Hyperboloid Model of Hyperbolic Plane Geometry

In what follows, let $(A, \langle\ ,\ \rangle)$ be a three-dimensional Minkowskian space endowed with a temporal orientation. Let o be an arbitrary point in A, and (as in section 3.3) let H_o^+ be the set of all points p in A such that $\overrightarrow{op}$ is a future-directed unit timelike vector. In this section we consider an interpretation **H** of L whose domain $|\mathbf{H}|$ is H_o^+, and show that it is isomorphic to the Klein-Beltrami model **K**. We call it the *hyperboloid* model of T_{Hyp}.

We take the assignment of **H** to 'B' to be:

$$\mathbf{H}(B) = \{(p, q, r) \in |\mathbf{H}|^3 \colon \overrightarrow{oq} = a\,\overrightarrow{op} + b\,\overrightarrow{or} \text{ for some numbers } a \geq 0 \text{ and } b \geq 0\}.$$

That is, we take q to qualify as "between" p and r if the vectors $\overrightarrow{op}$, $\overrightarrow{oq}$, and $\overrightarrow{or}$ belong to a two-dimensional subspace with the second between the first and the third. If we take $d_H : \mathbf{H} \times \mathbf{H} \to \mathbb{R}$ to be the distance function defined by $\cosh d_H(p, q) = \langle\overrightarrow{op}, \overrightarrow{oq}\rangle$, then the assignment of H to 'C' comes out to be

$$\mathbf{H}(C) = \{(p, q, r, s) \in |\mathbf{H}|^4 \colon d_H(p, q) = d_H(r, s)\}.$$

(The condition defining d_H, recall, is equivalent to stipulation that $d_H(p, q)$ be the (hyperbolic) angular measure $\measuredangle(p, o, q)$.)

There is a variant way to think about $\mathbf{H}(C)$ that is helpful. Given points p and q in $|\mathbf{H}|$, we take the *line segment* in $|\mathbf{H}|$ connecting them to be the set of all points r in $|\mathbf{H}|$ that are between p and q (in the sense of $\mathbf{H}(B)$). (See figure 4.2.1.) This line segment has a length with respect to the inner-product $\langle\ ,\ \rangle$. This length, we claim, is precisely $d_H(p, q)$.

To see this, note that we can express $\overrightarrow{oq}$ in the form $\overrightarrow{oq} = (\cosh\theta)\,\overrightarrow{op} + (\sinh\theta)\,w$, where $\theta = \cosh^{-1}(\langle\overrightarrow{op}, \overrightarrow{oq}\rangle) = d_H(p, q)$, and w is a unit spacelike vector orthogonal to $\overrightarrow{op}$. Now consider the curve $\gamma : [0, \theta] \to H_o^+$ defined by

$\gamma(s) = o + (\cosh s)\,\overrightarrow{op} + (\sinh s)\,w$. Clearly, $\gamma(0) = o + \overrightarrow{op} = p$, $\gamma(\theta) = o + \overrightarrow{oq} = q$, and the image of γ is just the "line segment" in which we are interested. Since $\gamma'(s) = (\sinh s)\,\overrightarrow{op} + (\cosh s)\,w$, $\|\gamma'(s)\| = 1$, and the length of the segment is

$$\|\gamma\| = \int_0^\theta \|\gamma'(s)\|\,ds = \theta = d_H(p,q).$$

Thus, as claimed, $d_H(p,q)$ is the length of the "line segment" in $|\mathbf{H}|$ connecting p to q as determined by the inner product $\langle\ ,\ \rangle$. Though we will not stop to do so, one can prove that, of all differentiable curves in H_o^+ connecting p and q, it is those whose images are "line segments" that have minimal length with respect to $\langle\ ,\ \rangle$. (Thus one can think of those segments as geodesics (or geodesic segments) with respect to the metric structure induced on H_o^+ by $\langle\ ,\ \rangle$.) This makes it clear that d_H satisfies the triangle inequality.

Given points p and q in $|\mathbf{H}|$, we know what the "line segment" in $|\mathbf{H}|$ connecting them is. We get the (full) *line* in $|\mathbf{H}|$ containing them, naturally enough, by extending the segment, i.e., by adding all points r in $|\mathbf{H}|$ such that either (p,q,r) or (r,p,q) stands in the relation $\mathbf{H}(B)$. Equivalently, we can characterize it as the set of all points of form $o + (\cosh s)\,\overrightarrow{op} + (\sinh s)\,w$ where w is as in the preceding paragraph and s is now *any* real number. Yet a third equivalent characterization is available. We can think of the line containing p and q as just the intersection of H_o^+ with the two dimensional affine subspace of A that contains o, p, and q. (See figure 4.2.1.)

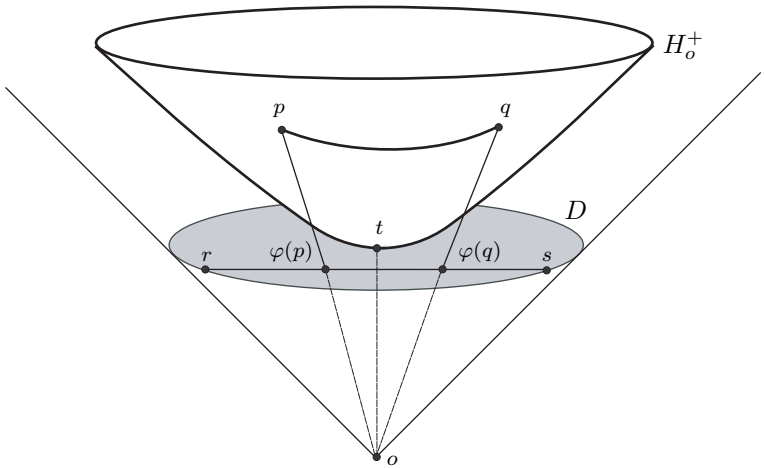


Figure 4.2.1: The indicated downward projection of H_o^+ onto the shaded disk D determines an isomorphism between the hyperboloid model $\mathbf{H}$ and the Klein-Beltrami model $\mathbf{K}$.

Now we proceed to show how one gets from $\mathbf{K}$ to $\mathbf{H}$, i.e., from the Klein-Beltrami model of hyperbolic plane geometry to the hyperboloid model. Let t be

a point (any point) in $|\mathbf{H}|$, and let D be the set of all points d such that $\overrightarrow{td} \perp \overrightarrow{ot}$ and $\|\overrightarrow{td}\| < 1$. We think of D as a copy of the unit disk that is the domain of $\mathbf{K}$. Consider the map $\varphi\colon |\mathbf{H}| \to A$, defined by setting $\varphi(p) = o + \langle \overrightarrow{op}, \overrightarrow{ot} \rangle^{-1} \overrightarrow{op}$ for all points p in $\mathbf{H}$. It is not hard to check that φ determines a bijection between $|\mathbf{H}|$ and D. (See problem 4.2.1.) Intuitively, it is a downward projection map. It assigns to p that point where the line through p and o meets the disk D.

Problem 4.2.1. *Verify that the map φ defined in the preceding paragraph is, as claimed, a bijection between H_o^+ and D.*

Notice next that φ preserves the betweenness relation, i.e., given any three points p, m, q in $|\mathbf{H}|$, m is between p and q in the sense of $\mathbf{H}(B)$ iff $\varphi(m)$ is between $\varphi(p)$ and $\varphi(q)$ in the sense of $\mathbf{K}(B)$. Formally, this is the assertion that: $\overrightarrow{om} = a\,\overrightarrow{op} + b\,\overrightarrow{oq}$ for some numbers $a \geq 0$ and $b \geq 0$ iff $\overrightarrow{\varphi(p)\,\varphi(m)} = k\,\overrightarrow{\varphi(p)\,\varphi(q)}$ for some number k, where $0 \leq k \leq 1$. One can certainly give an analytic proof of this fact. But it should be intuitively clear. Given points p and q in $|\mathbf{H}|$, consider the two-dimensional affine subspace containing them and the point o. The intersection of that subspace with H_o^+ is a line in H; its intersection with D plays the role of a line in $\mathbf{K}$; and our projection map takes the first intersection set to the second. So it is clear that φ takes lines in the first interpretation to lines in the second (and preserves the order of points on a line). This is precisely our claim.

Now, finally, we consider the distance functions d_H on $|\mathbf{H}|$ and d_K on D. We claim that for all points p and q in $|\mathbf{H}|$,

$$d_H(p, q) = d_K(\varphi(p), \varphi(q)). \tag{4.2.1}$$

(This will imply, of course, that φ preserves the congruence relation, i.e., given any four points p_1, q_1, p_2, q_2 in $|\mathbf{H}|$, the pairs $\{p_1, q_1\}$ and $\{p_2, q_2\}$ are congruent in the sense of $\mathbf{H}(C)$ iff the pairs $\{\varphi(p_1), \varphi(q_1)\}$ and $\{\varphi(p_2), \varphi(q_2)\}$ are congruent in the sense of $\mathbf{K}(C)$.) To set this up, let p and q be given, and let r and s be points on the boundary of D as called for in the definition of the cross ratio of $\varphi(p), \varphi(q)$ – with $\varphi(p)$ between r and $\varphi(q)$, and $\varphi(q)$ between $\varphi(p)$ and s. (See figure 4.2.1.) Now $d_H(p, q) = \cosh^{-1}(\langle \overrightarrow{op}, \overrightarrow{oq} \rangle)$ and

$$d_K(\varphi(p), \varphi(q)) \;=\; -\frac{1}{2}\log(CR(\varphi(p), \varphi(q))) \;=\; -\frac{1}{2}\log \frac{\|\overrightarrow{\varphi(p)\,r}\|\,\|\overrightarrow{\varphi(q)\,s}\|}{\|\overrightarrow{\varphi(p)\,s}\|\,\|\overrightarrow{\varphi(q)\,r}\|}.$$

We proceed by deriving an expression for the latter and showing it equal to the former. This involves a long, but relatively straightforward, computation. We start by deriving expressions for r and s. Since the points are collinear with $\varphi(p)$ and $\varphi(q)$ (and are positioned the way they are), there exist numbers $\alpha_+ > 1$ and $\alpha_- < 0$ such that

$$\begin{aligned} s &= \varphi(p) + \alpha_+\,\overrightarrow{\varphi(p)\,\varphi(q)} \\ r &= \varphi(p) + \alpha_-\,\overrightarrow{\varphi(p)\,\varphi(q)}. \end{aligned}$$

We can determine α_+ and α_- using the fact that $\overrightarrow{or}$ and $\overrightarrow{os}$ are null vectors, and will do so in a moment. But first we note that

$$\begin{aligned}
\|\overrightarrow{\varphi(p)\, r}\| &= |\alpha_-|\, \|\overrightarrow{\varphi(p)\, \varphi(q)}\| = (-\alpha_-)\, \|\overrightarrow{\varphi(p)\, \varphi(q)}\| \\
\|\overrightarrow{\varphi(q)\, s}\| &= \|\overrightarrow{\varphi(p)\, s}\| - \|\overrightarrow{\varphi(p)\, \varphi(q)}\| \;=\; \alpha_+\, \|\overrightarrow{\varphi(p)\, \varphi(q)}\| - \|\overrightarrow{\varphi(p)\, \varphi(q)}\| \\
&= (\alpha_+ - 1)\, \|\overrightarrow{\varphi(p)\, \varphi(q)}\| \\
\|\overrightarrow{\varphi(p)\, s}\| &= \alpha_+\, \|\overrightarrow{\varphi(p)\, \varphi(q)}\| \\
\|\overrightarrow{\varphi(q)\, r}\| &= \|\overrightarrow{\varphi(p)\, r}\| + \|\overrightarrow{\varphi(p)\, \varphi(q)}\| \;=\; (1 - \alpha_-)\, \|\overrightarrow{\varphi(p)\, \varphi(q)}\|.
\end{aligned}$$

Hence

$$CR(\varphi(p), \varphi(q)) = \frac{(-\alpha_-)(\alpha_+ - 1)}{\alpha_+(1 - \alpha_-)} = \frac{\alpha_- \,-\, \alpha_-\alpha_+}{\alpha_+ \,-\, \alpha_-\alpha_+}.$$

Since $\overrightarrow{os} = \overrightarrow{o\, \varphi(p)} + \overrightarrow{\varphi(p)\, s} = \overrightarrow{o\, \varphi(p)} + \alpha_+\, \overrightarrow{\varphi(p)\varphi(q)}$ is null, we have

$$0 = \|\overrightarrow{o\, \varphi(p)}\|^2 + 2\alpha_+ \langle \overrightarrow{o\, \varphi(p)},\, \overrightarrow{\varphi(p)\, \varphi(q)} \rangle - (\alpha_+)^2\, \|\overrightarrow{\varphi(p)\, \varphi(q)}\|^2.$$

Using abbreviations

$$\begin{aligned}
A &= \|\overrightarrow{\varphi(p)\, \varphi(q)}\|^2 \\
B &= -2\, \langle \overrightarrow{o\, \varphi(p)}, \overrightarrow{\varphi(p)\, \varphi(q)} \rangle \\
C &= -\|\overrightarrow{o\, \varphi(p)}\|^2,
\end{aligned}$$

we have $A(\alpha_+)^2 + B\alpha_+ + C = 0$. A similar analysis yields the same equation for α_-. Thus α_+ and α_- are just the two roots of a single quadratic equation. It follows that

$$\alpha_\pm = \frac{-B \,\pm\, (B^2 - 4AC)^{\frac{1}{2}}}{2A}$$

and therefore

$$\frac{\alpha_- - \alpha_-\alpha_+}{\alpha_+ - \alpha_-\alpha_+} = \frac{(-B - 2C) - (B^2 - 4AC)^{\frac{1}{2}}}{(-B - 2C) + (B^2 - 4AC)^{\frac{1}{2}}}.$$

It remains to derive expressions for the two terms appearing on the right side of the equation. For the first, we have

$$\begin{aligned}
-B &= 2\langle \overrightarrow{o\, \varphi(p)},\, \overrightarrow{\varphi(p)\, \varphi(q)} \rangle \;=\; 2\, \langle \overrightarrow{o\, \varphi(p)},\, -\overrightarrow{o\, \varphi(p)} \,+\, \overrightarrow{o\, \varphi(q)} \rangle \\
&= -2\, \|\overrightarrow{o\, \varphi(p)}\|^2 + 2\, \langle \overrightarrow{o\, \varphi(p)},\, \overrightarrow{o\, \varphi(q)} \rangle \;=\; 2C + 2\, \langle \overrightarrow{o\, \varphi(p)}, \overrightarrow{o\, \varphi(q)} \rangle.
\end{aligned}$$

So $(-B - 2C) = 2\, \langle \overrightarrow{o\, \varphi(p)},\, \overrightarrow{o\, \varphi(q)} \rangle$ and, hence,

$$\begin{aligned}
A &= -\langle -\overrightarrow{o\, \varphi(p)} + \overrightarrow{o\, \varphi(q)},\, -\overrightarrow{o\, \varphi(p)} + \overrightarrow{o\, \varphi(q)} \rangle \;=\; C + (-B - 2C) - \|\overrightarrow{o\, \varphi(q)}\|^2 \\
&= -(B + C) - \|\overrightarrow{o\, \varphi(q)}\|^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}(B^2-4AC) &= B^2+4C(B+C)+4C\|\overrightarrow{o\varphi(q)}\|^2 = (B+2C)^2+4C\,\|\overrightarrow{o\,\varphi(q)}\|^2\\ &= 4\,\langle\overrightarrow{o\,\varphi(p)},\,\overrightarrow{o\,\varphi(q)}\rangle^2-4\|\overrightarrow{o\,\varphi(p)}\|^2\|\,\overrightarrow{o\,\varphi(q)}\|^2.\end{aligned}$$

Finally, given our definition of φ, we have

$$\begin{aligned}\overrightarrow{o\,\varphi(p)} &= \langle\overrightarrow{op},\overrightarrow{ot}\rangle^{-1}\,\overrightarrow{op}\\ \overrightarrow{o\,\varphi(q)} &= \langle\overrightarrow{oq},\overrightarrow{ot}\rangle^{-1}\,\overrightarrow{oq}.\end{aligned}$$

So

$$\begin{aligned}(-B-2C) &= 2\langle\overrightarrow{op},\overrightarrow{ot}\rangle^{-1}\langle\overrightarrow{oq},\overrightarrow{ot}\rangle^{-1}\langle\overrightarrow{op},\overrightarrow{oq}\rangle\\ &= 2\langle\overrightarrow{op},\overrightarrow{ot}\rangle^{-1}\langle\overrightarrow{oq},\overrightarrow{ot}\rangle^{-1}(\cosh\measuredangle(p,o,q))\\ (B^2-4AC)^{\frac{1}{2}} &= 2\langle\overrightarrow{op},\overrightarrow{ot}\rangle^{-1}\langle\overrightarrow{oq},\overrightarrow{ot}\rangle^{-1}(\langle\overrightarrow{op},\overrightarrow{oq}\rangle^2-1)^{\frac{1}{2}}\\ &= 2\langle\overrightarrow{op},\overrightarrow{ot}\rangle^{-1}\langle\overrightarrow{oq},\overrightarrow{ot}\rangle^{-1}(\sinh\measuredangle(p,o,q))\end{aligned}$$

and therefore,

$$\begin{aligned}CR(\varphi(p),\varphi(q)) &= \frac{(-B-2C)-(B^2-4AC)^{\frac{1}{2}}}{(-B-2C)+(B^2-4AC)^{\frac{1}{2}}}\\ &= \frac{\cosh\measuredangle(p,o,q)-\sinh\measuredangle(p,o,q)}{\cosh\measuredangle(p,o,q)+\sinh\measuredangle(p,o,q)} = e^{-2\measuredangle(p,o,q)}.\end{aligned}$$

Hence

$$d_K(\varphi(p),\varphi(q)) = -\frac{1}{2}\log(CR(\varphi(p),\varphi(q))) = \measuredangle(p,o,q) = d_H(p,q).$$

This gives us (4.2.1) above, and we are done.

References

[1] T. Budden. Geometric simultaneity and the continuity of special relativity. *Foundations of Physics Letters*, 11:343–357, 1998.

[2] B. Gelbaum and J. Olmsted. *Counterexamples in Analysis*. Dover Publications, 1964.

[3] A. Grünbaum. David Malament and the conventionality of simultaneity: A reply. *Foundations of Physics, forthcoming*. Also available on the Pittsburgh Philosophy of Science Archive: http://philsci-archive.pitt.edu/archive/00000184/.

[4] A. Grünbaum. *Philosophical Problems of Space and Time*. Reidel, 2nd enlarged edition, 1973.

[5] J. Hintikka, editor. *The Philosophy of Mathematics*. Oxford University Press, 1969.

[6] M. Hogarth. Conventionality of simultaneity: Malaments result revisited. *Foundations of Physics Letters*, 18:491–497, 2005.

[7] A. Janis. Conventionality of simultaneity. In E. Zalta, editor, *Stanford Encyclopedia of Philosophy*, 2002. URL=http://plato.stanford.edu/archives/fall2002/entries/spacetime-convensimul/.

[8] S. Lang. *Linear Algebra*. Springer Verlag, 3rd edition, 1987.

[9] H. Reichenbach. *The Philosophy of Space and Time*. Dover, 1958.

[10] J. Roe. *Elementary Geometry*. Oxford University Press, 1993.

[11] P. Spirtes. Conventionalism in the philosophy of Henri Poincaré. unpublished Ph. D. thesis, University of Pittsburgh, 1981.

[12] H. Stein. On relativity theory and the openness of the future. *Philosophy of Science*, 58:147–167, 1991.

[13] W. Szmielew. Some metamathematical problems concerning elementary hyperbolic geometry. In L. Henkin, P. Suppes, and A. Tarski, editors, *The Axiomatic Method, with Special Reference to Geometry and Physics*. North Holland, 1959.

[14] A. Tarski. What is elementary geometry. In L. Henkin, P. Suppes, and A. Tarski, editors, *The Axiomatic Method, with Special Reference to Geometry and Physics*. North Holland, 1959. The article was reprinted in [5].