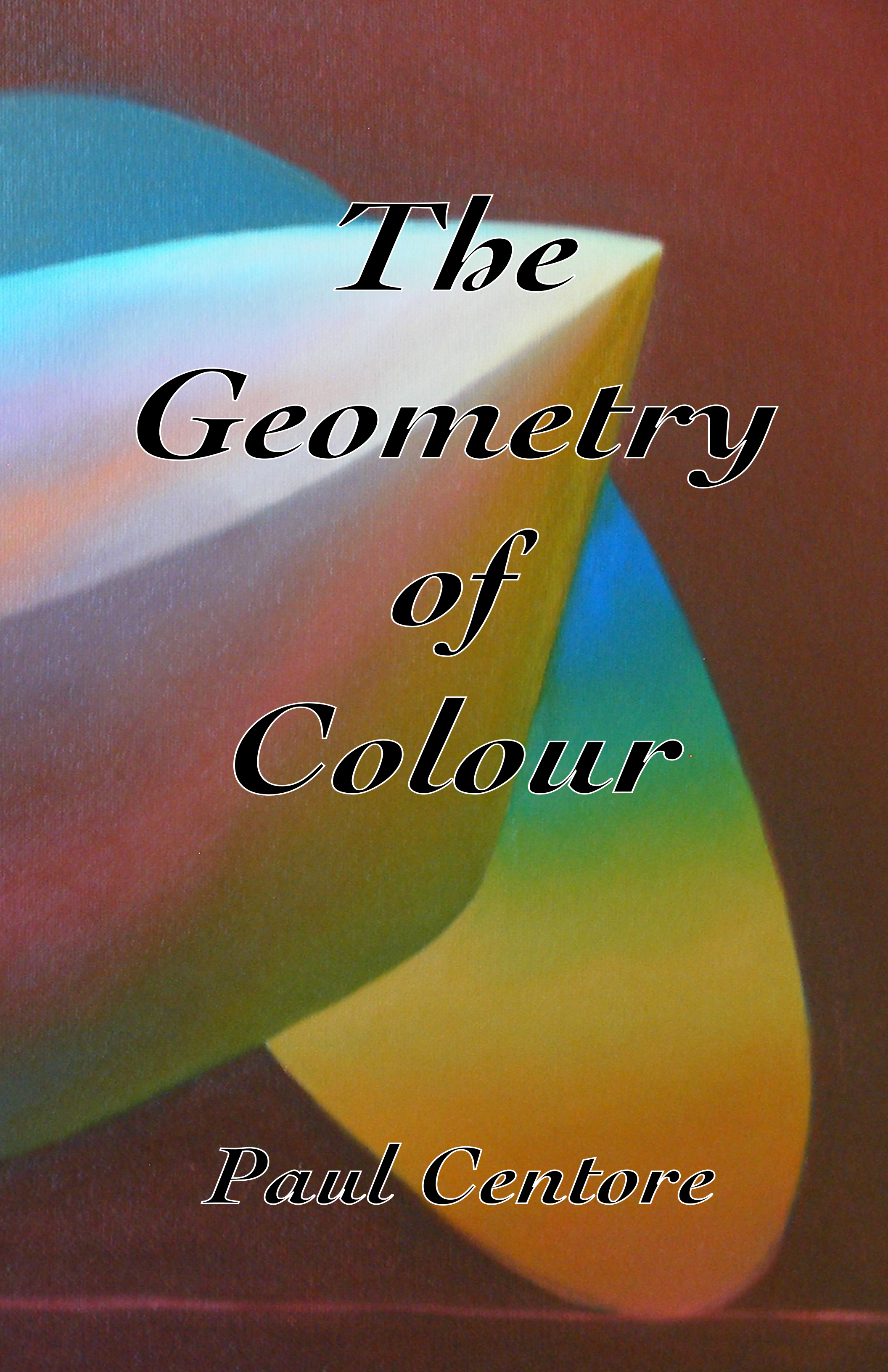

The Geometry of Colour

Paul Centore

The Geometry of Colour

Paul Centore

Suggested Cataloging Data
Centore, Paul
The Geometry of Colour/P. Centore
221 p., 21.6 by 14.0 cm
ISBN 978-0-9990291-0-7
1. Color 2. Color—Mathematics 3. Colorimetry 4. Convex Geometry I. Centore, Paul, 1968- II. Title
QC495.C397 2017
535.6

Printed in the United States of America

ISBN 978-0-9990291-0-7

ISBN 9780999029107

9 780999 029107

Introduction

Colour is a universal experience, which has been investigated since antiquity. Today, colour science rests firmly on both empirical findings and theoretical development. Unlike many technical fields, colour science considers not only physical factors, but also human perception, and the relationships between the two. A basic part of colour science is colour matching, which identifies when two side-by-side lights, viewed in isolation through an aperture, produce the same colour.

The Geometry of Colour formulates colour matching mathematically, emphasizing geometric constructions and understanding. The motivation for writing it was the unifying insight that many apparently disparate objects in colour science share a common zonohedral structure. The book aims at both rigor and intuition. Fortunately, many colour science objects can be explicitly constructed in a three-dimensional vector space, and, while linear algebra insures rigorous definitions, a premium is placed on a concrete visual and spatial presentation—ideally, a motivated reader could build literal models, for example with foam board and glue.

This book is intended for mathematicians, colour scientists, and, as much as possible, curious non-specialists. Familiarity with basic linear algebra is assumed, although anybody who understands ordinary vector operations, and is comfortable thinking visually, could follow most of the book. Mathematicians with no preconceptions about colour will likely follow the derivations easily, but find some colour concepts, which arise from empirical considerations, puzzling and unmotivated. Though indispensable, colour-matching experiments are artificial, and their relationship to everyday colour appearance is still unclear. Readers looking to understand colour in realistic situations will likely be dissatisfied: this book will not do much to distinguish red from blue, nor dull colours from bright ones. Unlike mathematicians, colour scientists already understand the place of colour matching in colour science, but will probably find familiar concepts expressed in terms that are unfamiliar, and perhaps off-putting. The mathematical excursions, however, eventually pay large dividends, by identifying a common structure in seemingly unrelated colour science concepts: object-colour solids, illuminant gamuts, and electronic displays are all shown to be zonohedra.

The Geometry of Colour is organized as follows. The first two chapters present only mathematics, and no colour science. The first chapter presents convexity in a vector space setting, with no metric assumptions; the tools of linear functionals and hyperplanes are introduced. The second chapter discusses Minkowski sums and zonohedra. The Minkowski sum "adds" sets in a vector space geometrically, by sweeping one set over the other. A zonohedron is a special Minkowski sum, in which all the summands, or generators, are vectors in $\mathbb{R}^3$. Zonohedra are convex, rotationally symmetric, polytopes whose faces are generically parallelograms.

The third chapter discusses the physics needed for colour science: humans perceive a light stimulus as having a colour, so a vector space of light stimuli is constructed. Also, light often reflects off coloured surfaces before entering the human eye, so a vector space of reflectance functions is constructed. The fourth chapter introduces human colour-matching behavior, as codified in the CIE 1931 Standard Observer, which defined a linear transformation from physical light stimuli to the three-dimensional space of perceived colours. The set of physical stimuli has a high dimension, so often multiple stimuli, called *metamers*, produce the same colour perception.

The fifth chapter applies the convexity and zonohedra of the first two chapters to the Standard Observer. Zonohedral *object-colour solids* are constructed, consisting of those colours in colour space that result when a given light source reflects off a physical object. Such a zonohedron's generators are the *spectrum locus vectors*, which correspond to the colour perceptions that would result from objects which reflect light at only one wavelength. The convex cone of the locus vectors is called the *spectrum cone*, and a two-dimensional section of this cone, after a coordinate transformation, is the ubiquitous *chromaticity diagram*. Each object-colour solid fits perfectly into the spectrum cone at the origin. The colours on the boundary of an object-colour solid are called *optimal colours*, and the Optimal Colour Theorem, states that they have a special form, originally given by Erwin Schrödinger. The zonohedral development is used to prove this theorem, and further results about metamerism.

The sixth chapter of the book applies the zonohedral approach to *computational colour constancy*, which arises when working with cameras. Cameras are usually designed to mimic the human eye, so many constructions for human vision have camera analogues. The analogues of human object-colour solids are called *illuminant gamuts*, and are also zonohedra. The chapter constructs illuminant gamuts and related camera objects, via the Minkowski sum, and discusses their relevance in some camera applications. The seventh chapter deals with electronic displays, showing that a *display gamut*, which is the set of all the colours a display can produce, is a zonohedron, and that a dissection of this gamut into parallelepipeds can be used for some practical applications.

This book developed over several years, and was originally motivated by a search for an intuitive, visual derivation of the Schrödinger form for optimal colours. Surprisingly, the resulting zonohedral form also led to results in electronic displays, and then in computational colour constancy. After further applications and development, the mass of results, and the diverse topics in colour science that they connected, warranted a book.

In mathematics, questions of originality are sometimes difficult, not to mention delicate. Often multiple independent investigators glimpse or vaguely intuit an important idea, without fully articulating it. The zonohedral approach to colour science likely follows this rule. While I conceived and developed it on my own, probably other researchers had similar insights previously, so I hesitate to claim much originality. In any event, this book contributes to the literature by developing the zonohedral approach explicitly, in a unified treatment.

And lastly, one regret. As a graduate student at the University of Toronto, I attended the weekly geometry seminars. Though long retired, the eminent Professor H. S. M. Coxeter regularly attended, and delivered quite a few talks himself. At that time I had no inkling of the geometry hidden in colour, but was inspired by Professor Coxeter's surprising ability to turn concrete observations into sophisticated abstract mathematics. Professor Coxeter passed away ten years ago, so I regret that I can't present this book's results to him. He liked surprising applications, and he liked zonohedra—and colour science is a surprising application of zonohedra.

—Paul Centore
Gales Ferry, Connecticut
June 4, 2017

Contents

Chapter 1

Convexity in Vector Spaces

1.1 Introduction

This book elucidates some geometric structures that arise from colour science. In particular, we will focus on colour-matching experiments, in which two visual stimuli, though they are very different physically, produce the same colour. Both the set of visual colour stimuli and the set of colour perceptions can be formulated as convex subsets of vector spaces, and the relationships between them can be formulated as linear transformations involving those subsets and vector spaces. This chapter and the next provide mathematical tools for understanding and working with convex sets in a vector space setting.

A basic understanding of linear algebra, at about the level of a second-year university course, will be assumed. Concepts such as linear dependence/independence, bases, and subspaces will be used freely, without explanation. As much as possible, however, the presentation will emphasize the visual and spatial aspects, so much of it can be followed by any reader who is familiar with basic vector operations (addition and scalar multiplication). Fortunately, colour space is three-dimensional, so helpful pictures can be drawn of many of the objects of interest, which should further aid understanding.

The basic object of linear algebra is a finite-dimensional vector space over the real numbers, which we will denote $\mathbb{R}^n$, where n indicates the dimension of the space. $\mathbb{R}^2$ and $\mathbb{R}^3$ can be visualized as a flat plane and space, respectively. A vector is often pictured as an arrow that extends from the origin of a vector space to some point in that space; we could just as well interpret a vector as the point at the head of that arrow, so the terms *point* and *vector* will be used interchangeably.

Convexity is an important concept, which can be defined naturally in a vector space. A subset of a vector space is convex if that subset contains the line segment between any pair of points in that subset. Convex sets have many convenient properties, such as a well-defined dimension. Bounded convex sets have topologically simple boundaries, and can be approximated arbitrarily well by polytopes, which are multi-dimensional versions of polygons and polyhedra. Furthermore, many bounded convex sets of interest can be generated by a finite set of points, called vertices. Even some unbounded convex sets, such as convex cones, can be finitely generated (by a set of rays rather than points). The most important property for us is that convex sets, and much of their internal structure, are preserved under linear transformations. As a result, a convex set can be transferred from one vector space to another, and still remain convex; likely it will undergo some structural adjustments, which will provide important information about the transformation.

Convexity is a well-developed area of mathematics, with a long history and a wide variety of important results. This chapter's treatment of convexity is far from comprehensive, presenting only the material needed to study the geometry of colour matches. Many results are stated without proof. For a more comprehensive treatment, that is both rigorous and readable, and that proves all the statements made in this chapter, consult Steven Lay's *Convex Sets and Their Applications.*

This chapter is organized as follows. First we discuss linear functionals, and describe how they divide a vector space into a set of parallel hyperplanes; hyperplanes are useful because they can separate convex sets and act as bounds for them. Next, we define and describe convex sets in vector spaces, and show how any convex set can be generated by a minimal (and often finite) set of points. Then we deal with two special kinds of convex sets, called convex cones and polytopes, and show how hyperplanes can be used to delimit them. These concepts are used in the next chapter, which presents a further special kind of convex set, called a zonohedron. Next, a linear transformation between two vector spaces is shown to preserve convexity, and much of the internal structure of cones and polytopes. These facts become important later in the book, when the set of human colour perceptions is shown to be a subset of a three-dimensional vector space, and individual colours are the images, under a linear transformation, of functions over the visible electromagnetic spectrum. Finally, a warning is given against implicitly attributing Euclidean ideas such as distance and angle to the vector spaces defined in this book: somewhat counterintuitively, combinatorial and linear relationships are sufficient to investigate colour matching mathematically.

1.2 Functionals and Hyperplanes

Given two vector spaces, $\mathbf{V}$ and $\mathbf{W}$, of any dimension, a special kind of function, called a *linear transformation*, can be defined between them. Formally $L : \mathbf{V} \to \mathbf{W}$ is a linear transformation if and only if

$$L(\alpha \mathbf{v}_1) = \alpha L(\mathbf{v}_1), \text{ and} \tag{1.1}$$

$$L(\mathbf{v}_1 + \mathbf{v}_2) = L(\mathbf{v}_1) + L(\mathbf{v}_2), \tag{1.2}$$

for any two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ in $\mathbf{V}$, and any real number α. When $\mathbf{V}$ and $\mathbf{W}$ are the same vector space, so L goes from $\mathbf{V}$ to itself, L is also referred to as a *linear operator*. When $\mathbf{W}$ is the real line $\mathbb{R}^1$, L is also referred to as a *linear functional*, or just a *functional*. This section will describe how a linear functional subdivides a vector space into a set of parallel hyperplanes.

A *hyperplane* $\mathcal{H}$, sometimes also called an *affine hyperplane*, in a vector space $\mathbf{V}$ of dimension n, is a translation, by an arbitrary vector, of a subspace of dimension $n-1$. Every subspace $\mathbb{S}_{n-1}$ of dimension $n-1$ contains the origin of $\mathbf{V}$, and divides the vector space into two regions, or *half-spaces*, one on either side of $\mathbb{S}_{n-1}$. Some simple examples are a line through the origin in $\mathbb{R}^2$, and a plane through the origin in $\mathbb{R}^3$. That subspace can be translated by a vector $\mathbf{v}$, simply by adding $\mathbf{v}$ to every vector in $\mathbb{S}_{n-1}$. The result (unless $\mathbf{v}$ is already in $\mathbb{S}_{n-1}$) is that the subspace is shifted so that it no longer contains the origin. The shifted subspace is a hyperplane that is parallel to the original subspace. Some simple examples of hyperplanes are the line $x + y = 1$ in $\mathbb{R}^2$, and the plane $x + y + z = 1$ in $\mathbb{R}^3$. Every hyperplane $\mathcal{H}$, whether it contains the origin or not, divides $\mathbf{V}$ into two regions, one on either side of $\mathcal{H}$. This division can be used to separate two disjoint convex sets, or to separate one convex set from some region of the vector space; such separations are not always possible with non-convex sets.

Hyperplanes are intimately connected with linear functionals. Suppose we have a (non-zero) linear functional F, from $\mathbf{V}$ to $\mathbb{R}$, and a real number α. Then the set of vectors in $\mathbf{V}$ which are mapped by F to α is called the *pre-image* of α, and denoted $F^{-1}(\alpha)$. Since $\mathbf{V}$ has dimension n and $\mathbb{R}$ has dimension 1, the kernel of F, denoted ker F, is a subspace of dimension $n-1$. A basic result of linear algebra is that

$$F^{-1}(\alpha) = \ker F + \mathbf{v}, \tag{1.3}$$

where $\mathbf{v}$ is any vector such that $F(\mathbf{v}) = \alpha$. Since $\ker F$ is a subspace of dimension $n-1$, it follows that $F^{-1}(\alpha)$ is a hyperplane.

The above argument applies to any real numbers α and β, so $F^{-1}(\alpha)$ and $F^{-1}(\beta)$ are both hyperplanes of the form given in Equation (1.3). It follows from this form that all the hyperplanes corresponding to pre-images under F are parallel. Since F is defined for every vector in $\mathbf{V}$, the hyperplanes *foliate* $\mathbf{V}$: each vector in $\mathbf{V}$ belongs to one and only one hyperplane. Furthermore, $F^{-1}(0)$ is just another way of writing ker F, so $F^{-1}(0)$ is a hyperplane through the origin.

In $\mathbb{R}^3$, one might visualize the set of hyperplanes corresponding to a linear functional as an infinite stack of pancakes, extending both upwards and downwards. The stack could be tilted at any angle. The pancakes, or hyperplanes, are assigned numbers by the functional. The hyperplane, through the origin is assigned the number 0; a hyperplane somewhat above that is assigned the number 1, and hyperplanes between those two are sequentially assigned numbers from 0 to 1. The hyperplane twice as far from the origin as hyperplane number 1 is assigned number 2, and so on to infinity. On the other side of the kernel hyperplane, the numbering is similar, but assigns negative values.

This interpretation will be useful later. To prove that a convex set is restricted to a half-space, it suffices to find a linear functional F that is positive on every point in the convex set; the hyperplane $F^{-1}(0)$, or just as conveniently the hyperplane for any negative number, then defines a half-space which contains the convex set.

The above construction can be reversed: one can start with a hyperplane $\mathcal{H}$, and construct a linear functional F. Suppose for the sake of discussion that $\mathcal{H}$ does not contain the origin. Then arbitrarily choose a vector $\mathbf{v}$ in $\mathcal{H}$ and write

$$\mathcal{H} = \mathbb{S}_{n-1} + \mathbf{v}, \tag{1.4}$$

where $\mathbb{S}_{n-1}$ is a subspace of dimension $n-1$. Now define $F(\mathbf{s}) = 0$, for every $\mathbf{s}$ in $\mathbb{S}_{n-1}$, and define $F(\mathbf{v}) = 1$. $\mathbf{v}$ is linearly independent of $\mathbb{S}_{n-1}$, and a basis for $\mathbb{S}_{n-1}$ would have $n-1$ linearly independent vectors. By giving the values of F on a set of n linearly independent vectors, F is completely defined as a functional from $\mathbf{V}$ to $\mathbb{R}$. (This construction can easily be modified if one starts with a hyperplane through the origin, by choosing a parallel hyperplane that does not go through the origin.) Although F is well-defined, it is not unique: one could choose a different $\mathbf{v}$ in $\mathcal{H}$, or let $F(\mathbf{v})$ be any non-zero real number. The choice of functionals, however, is limited. Suppose F_1 and F_2 are two functionals that are constant on $\mathcal{H}$; then it is easy to see that $F_1 = kF_2$, for some non-zero constant k.

To sum up, a non-zero linear functional F on $\mathbf{V}$ allows V to be foliated into a set of parallel hyperplanes, which can be naturally indexed by the real numbers. The kernel of F is not only a hyperplane, but also a subspace of dimension $n-1$, which contains the origin; this subspace has index 0. Conversely, given any hyperplane $\mathcal{H}$ in $\mathbf{V}$, a linear functional can be found, whose value on that hyperplane is constant; this linear functional is unique up to a scaling factor.

1.3 Convex Sets in Vector Spaces

1.3.1 Convex Sets

Although convex sets were originally defined in Euclidean spaces, we will see that a Euclidean structure is not needed, and that a vector space by itself has sufficient structure. The main requirement to discuss convexity is that a unique line segment can be drawn between any pair of points. In Euclidean geometry, of course, this requirement is an axiom: the line segments are just ordinary straight lines. In a vector space, line segments can be constructed from the vector space axioms. Suppose that $\mathbf{v}_1$ and $\mathbf{v}_2$ are two vectors in a vector space $\mathbf{V}$ of dimension n. Then define the line segment between them by

$$\left\{ \sum_{i=1}^{2} \alpha_i \mathbf{v}_i \,\middle|\, 0 \leq \alpha_i \leq 1 \ \forall i \text{ and } \sum_{i=1}^{2} \alpha_i = 1 \right\}. \tag{1.5}$$

Geometrically, this set gives a parametrized path from the interval $[0, 1]$ into $\mathbf{V}$, as can be seen by rewriting Equation (1.5) as

$$\{\alpha_1 \mathbf{v}_1 + (1 - \alpha_1)\mathbf{v}_2 | 0 \leq \alpha_1 \leq 1\}. \tag{1.6}$$

The path starts at $\mathbf{v}_2$, when $\alpha_1 = 0$, and ends at $\mathbf{v}_1$, when $\alpha_1 = 1$. The vectors defined by Equation (1.5) are called *convex combinations* of $\mathbf{v}_1$ and $\mathbf{v}_2$. If the vector space did have a standard Euclidean structure, given by an inner product, then this path would correspond to

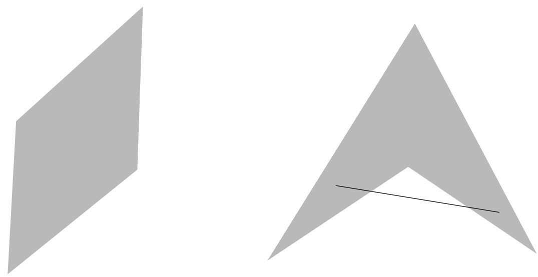

Figure 1.1: Examples of Convex and Non-Convex Sets

the Euclidean straight line segment between the points. In fact, this statement holds for an arbitrary (non-degenerate) inner product, so the path can reasonably be thought of as a straight line segment. The path is symmetric: it could just as easily start at $\mathbf{v}_1$ and end at $\mathbf{v}_2$, as can be seen by reparametrizing so that $\alpha_2 = 1 - \alpha_1$. The set of points in either parametrization is identical.

This algebraic line segment is sufficient to make the following definition: a subset $\mathcal{K}$ of $\mathbf{V}$ is *convex* if and only if, for every pair of points $\mathbf{v}_1$ and $\mathbf{v}_2$ in $\mathcal{K}$, the straight line segment given by Equation (1.5) is also in $\mathcal{K}$. To avoid technicalities, convex sets are sometimes required to be closed, or one can automatically take the closure of a convex set; this book will typically assume implicitly that convex sets are closed.

Convex sets can exist in any dimension, but are easily understood in two dimensions. Figure 1.1 shows a diamond, which is convex, and a chevron, which is not convex. To see that the diamond is convex, choose any two points, and draw the line segment between them; no matter which two points are chosen, the line segment between them is always within the diamond. This property does not hold for the chevron: the line segment shown joins two points of the chevron, but lies partially outside the shape. Thus the chevron is not convex. Similar examples can easily be constructed in any dimension.

1.3.2 Convex Hulls

A convex set called the *convex hull* can be constructed from an arbitrary set $\mathcal{Q}$ of points in a vector space. The convex hull is the smallest convex set that contains all the points in $\mathcal{Q}$. As a set, it is given by

$$\operatorname{hull}(\mathcal{Q}) = \left\{ \sum_{i=1}^{m} \alpha_i \mathbf{v}_i \,\middle|\, \mathbf{v}_i \in \mathcal{Q}\ \forall i, 0 \le \alpha_i \le 1\ \forall i, \text{ and } \sum_{i=1}^{m} \alpha_i = 1 \right\}, \tag{1.7}$$

where m is any finite number. Note that Equation (1.7) is just Equation (1.5), with 2 replaced by m. In general, a *convex combination* of m vectors is a linear combination of those vectors, with the requirements that each coefficient in the combination is between 0 and 1 inclusive, and that the coefficients sum to 1. Equation (1.7) then says that the convex hull of $\mathcal{Q}$ consist of all convex combinations of all finite subsets of $\mathcal{Q}$. If $\mathcal{Q}$ consists of three (non-collinear) vectors, then their convex hull is just the triangle for which those vectors are vertices. Similarly, the diamond in Figure 1.1 is the convex hull of its four vertices.

Convex hulls can be constructed inductively by adding points one by one. Figure 1.2 shows an example. To begin with, $\mathbf{v}_1$ is a single point, which is itself a convex set, shown on the far left of Figure 1.2. Now add the second point $\mathbf{v}_2$. The convex hull is a convex set that contains both $\mathbf{v}_1$ and $\mathbf{v}_2$, so it must contain the line segment joining $\mathbf{v}_1$ and $\mathbf{v}_2$, but no further points are necessary. This new convex set is second on the left in Figure 1.2. Then add $\mathbf{v}_3$. Convexity requires that the convex hull contains every line segment that joins $\mathbf{v}_3$ to any point on the line segment between $\mathbf{v}_1$ and $\mathbf{v}_2$. The result, unless the three points are collinear, is a triangle. To add the fourth point, $\mathbf{v}_4$, one must also add all line segments between $\mathbf{v}_4$ and any point in the triangle. If $\mathbf{v}_4$ is in the same plane as the triangle, the result will be another plane figure, such as the diamond in Figure 1.1. Possibly, $\mathbf{v}_4$ is in the interior of the triangle, so it is already a convex combination of the other three points; in this case the convex hull contains no new points and remains the triangle, shown on the far right of Figure 1.2. If $\mathbf{v}_4$ is outside the plane of the triangle, then the result will be a tetrahedron, extending into three-dimensional space. This process can be continued indefinitely, sometimes increasing the dimension and sometimes not.

The set $\mathcal{Q}$ is said to be a set of *generators* for $\operatorname{hull}(\mathcal{Q})$. There can be multiple sets of generators for the same convex hull. For instance, let $\mathcal{Q}$ be the union of the diamond's vertices and the diamond's center. Then $\mathcal{Q}$ generates the diamond, but the center of the diamond is superfluous—the set $\mathcal{Q}$ with the diamond's center removed would also generate the diamond. A convex set cannot always be written as the convex hull of a finite set of generators. For instance, a circular disc is the convex hull of all the points on its circumference (an infinite set),

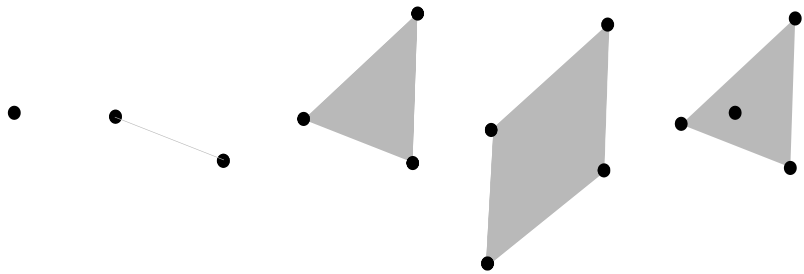

Figure 1.2: Constructing a Convex Hull Inductively

but not the convex hull of any smaller set of points.

1.3.3 Convex Cones

An important kind of convex set in a vector space is a *ray.* A ray is a straight line that starts at the origin, and continues on indefinitely in some direction. Formally, let $\mathbf{v}$ be a vector in $\mathbf{V}$. Then the set

$$\mathrm{ray}(\mathbf{v}) = \{\alpha\mathbf{v} | 0 \leq \alpha\} \tag{1.8}$$

defines the ray in the direction $\mathbf{v}$. Informally speaking, a ray is half of the one-dimensional subspace generated by $\mathbf{v}$; any one-dimensional subspace is a straight line that goes through the origin and continues to infinity in two opposite directions. The choice of $\mathbf{v}$ in Equation (1.8) is not unique: any positive multiple of $\mathbf{v}$ would generate the same set. Like convex sets, rays can exist in a vector space of arbitrary dimension.

A *convex cone* is the convex hull of a set of rays (or, to be technically correct, it is the convex hull of the union of all the points contained in all those rays). The set of rays is said to *generate* its convex cone. Without loss of generality, we can also define the convex cone of an arbitrary set of points $\mathcal{Q}$ as the convex cone of the rays resulting from each vector in $\mathcal{Q}$. Formally, define

$$\mathrm{cone}(\mathcal{Q}) = \left\{ \sum_{i=1}^{m} \alpha_i \mathbf{v}_i \middle| \mathbf{v}_i \in \mathcal{Q},\ 0 \leq \alpha_i \right\}, \tag{1.9}$$

where m is any finite number. Equation (1.9) can easily be modified to apply to a set of rays: just define a set $\mathcal{Q}$ such that each ray is generated by one vector in $\mathcal{Q}$. It is easy to see that the same convex set will

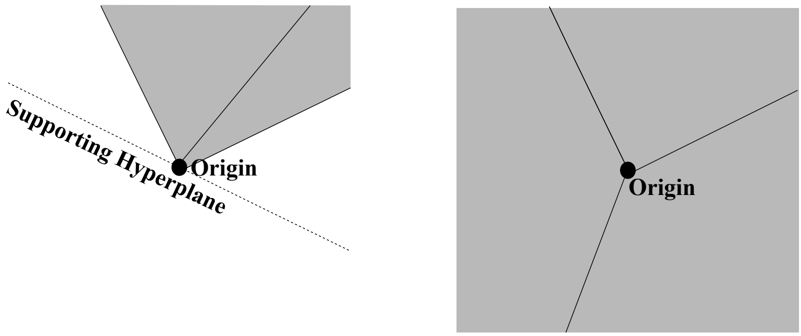

Figure 1.3: Proper and Improper Convex Cones

result, regardless of the choice of the rays' generating vectors. The linear combinations in Equation (1.9) might be called *non-negative combinations*, because their coefficients are all non-negative—but otherwise arbitrary, as opposed to convex combinations, which require in addition that all coefficients sum to 1.

Suppose there are two rays. Then their convex cone is the infinite wedge between them. Note that this wedge is in the smaller of the two angles between the rays. In the special case in which the two rays are in opposite directions, their convex cone collapses to a single line through the origin. If the two rays are in a vector space of dimension greater than two, then the convex cone is restricted to the two-dimensional plane that the rays span.

If a third ray is added, then the convex cone is usually a triangular pyramid whose apex is the origin and whose "base" is infinitely far away. An instructive example occurs when the three rays are in a two-dimensional vector space. Figure 1.3 breaks this instance into the *proper* and *improper* cases, which give qualitatively different results. In this figure, the solid black lines represent rays through the origin, and the grey regions are the convex cones. At the left of the figure, all three rays are on the same side of a line through the origin; then the convex cone is again a wedge, with one of the rays inside it. The outside line is a hyperplane, which separates the convex cone from the rest of the vector space; such a hyperplane is called a *supporting hyperplane*. This hyperplane is not unique; many choices could be made for it.

In the second case, no such separating line exists; then the convex cone is the entire plane. The dividing line between the two cases occurs when the total angle of the cone at the origin is 180°. In the case on the left, the bounding rays could be spread apart until the angle between them reaches 180°. A supporting hyperplane would still exist, and the convex cone would contain at most a half-space. Once the angle exceeds 180°, however, the convex cone contains the entire vector space. There are no intermediate cases. There could not be, for example, a convex cone of angle 240°. In higher dimensions, the separating line becomes a supporting hyperplane, and the cone can have higher dimension than two. All the same considerations apply when there are more than three rays: a supporting hyperplane can be either found or not, no matter how many rays are in the set.

The cone on the left is called a *proper cone*, because it fits our intuitive idea of what a cone should be better than the cone on the right. Formally, a cone is said to be proper if it does not contain any one-dimensional subspaces; geometrically, of course, one-dimensional subspaces are just straight lines through the origin. Equivalently, if a proper cone contains a non-zero vector $\mathbf{v}$, then it cannot also contain $-\mathbf{v}$. This condition allows the construction of a functional which is always non-negative, whose kernel is a supporting hyperplane through the origin. An improper cone contains both $\mathbf{v}$ and $-\mathbf{v}$ for some vector, so any (non-zero) functional on an improper cone takes on both positive and negative values.

An important case of a proper convex cone in a vector space $\mathbf{V}$ is a non-negative octant, denoted $\mathcal{O}$. Non-negative octants are not unique—there are different ones for different choices of basis. Once a particular basis is fixed, $\mathcal{O}$ consists of the convex cone generated by the basis rays, where a basis ray is a ray in the direction of a basis vector. Equivalently, $\mathcal{O}$ contains any vector whose coordinates, when written in the fixed basis, are all non-negative. The convexity of $\mathcal{O}$ is evident, because a convex combination of any two non-negative vectors is again a non-negative vector, as can be seen from Equation (1.5).

A simple example is the vector space $\mathbb{R}^3$, in Cartesian coordinates. The non-negative octant is then a "cube" that extends to infinity. It has three infinite edges, which are the basis rays in the x, y, and z directions. The co-ordinate vectors for these rays are $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$. Clearly every non-negative vector can be written as a non-negative linear combination of these three vectors, and every non-negative linear combination gives a non-negative vector, showing that the non-negative octant is identical with the convex cone of the basis rays. This example easily generalizes to higher dimensions.

It is easy to see geometrically that there exists a supporting hyperplane, in fact many supporting hyperplanes, for $\mathcal{O}$, satisfying the first case in Figure 1.3. One such hyperplane, $\mathcal{H}$, is the kernel of the linear functional F, given by

$$F(\mathbf{v}) = \sum_{i=1}^{n} v_i, \tag{1.10}$$

where the v_i's are the coordinates of $\mathbf{v}$ in terms of a fixed basis, given by vectors $\{\mathbf{b}_1, \mathbf{b}_2, ...\mathbf{b}_n\}$:

$$\mathbf{v} = \sum_{i=1}^{n} v_i \mathbf{b}_i. \tag{1.11}$$

In $\mathbb{R}^3$ with Cartesian coordinates, this functional would be written as $F(x, y, z) = x + y + z$.

The hyperplane $\mathcal{H}$ is the kernel of F, given by

$$\mathcal{H} = \ker F = \{\mathbf{v} \in V | F(\mathbf{v}) = 0\}. \tag{1.12}$$

Since F applied to any non-negative vector (except the origin) is greater than 0, and since F is 0 on any vector in $\mathcal{H}$, it follows from the previous discussion that all of $\mathcal{O}$ is on one side of $\mathcal{H}$ (excepting the origin itself, which is contained in $\mathcal{H}$). $\mathcal{H}$ is therefore a supporting hyperplane that separates the non-negative octant from a half-space of $\mathbf{V}$.

1.3.4 Convex Polytopes

A particularly simple and important kind of convex set is called a *convex polytope*, which is defined to be the convex hull of a finite set of points. A polytope can be thought of as an n-dimensional generalization of polygons and polyhedra. A convex polyhedron consists of a three-dimensional interior, bounded by two-dimensional polygonal faces; each bounding polygon is itself convex and is bounded by convex line segments, which are themselves bounded by isolated vertices. Similarly, a general n-dimensional convex polytope is bounded by a set of $(n-1)$-dimensional convex polytopes, which are themselves bounded by $(n-2)$-dimensional convex polytopes, and so on, until the vertices are reached.

Since a convex polytope $\mathcal{P}$ is the convex hull of a finite set, we can write

$$\mathcal{P} = \text{hull}\left(\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_m\}\right). \tag{1.13}$$

It is possible, however, that some $\mathbf{v}_i$'s are superfluous, like the point in the interior of the triangle at the right of Figure 1.2. A particular $\mathbf{v}_i$ is superfluous if it can be written as a convex combination of other vectors in the set. The same polytope will result if all the superfluous members of the set are removed, so we can assume without loss of generality that the generating set is *minimal*, that is, that all its elements are necessary. Furthermore, it can be shown that the minimal generating set of a polytope is unique. Geometrically, each member of the minimal generating set is a vertex of the polytope, and each vertex is a member of the minimal generating set.

A polytope's vertices have the special property of being *exposed points*. A point $\mathbf{v}$ of a convex set $\mathcal{K}$ is said to be exposed if there exists a supporting hyperplane $\mathcal{H}$ of $\mathcal{K}$, such that $\mathcal{H} \cap \mathcal{K} = \{\mathbf{v}\}$. Geometrically, this condition says that $\mathcal{K}$ is on one side of $\mathcal{H}$, but that $\mathbf{v}$ is the *only* point of $\mathcal{K}$ that is contained in $\mathcal{H}$. It can also be shown that there exists a supporting hyperplane that intersects the polytope in exactly one edge, or exactly one face; a polytope has the attractive property that all its bounding subsets, of any dimension, are exposed in the same way its vertices are.

This result will be useful later on, when we consider optimization over polytopes. A standard optimization problem, dealt with in linear programming, is to maximize a linear functional over a convex polytope. If the linear functional is seen as a stack of hyperplanes, then its maxima all occur on a hyperplane that supports the polytope. If the supporting hyperplane intersects the polytope in only one point, then the maximum is unique, and is an exposed point of the polytope. For a vertex of the polytope, the converse is also possible: we can construct a linear functional whose unique maximum occurs at that vertex.

1.4 Linear Transformations and Convexity

Linear transformations, as defined by Equations (1.1) and (1.2), are basic to linear algebra, because they preserve much of a vector space's structure. This section will show that they also preserve much of the structure of convex sets.

An important result is that the linear image of a convex set is again convex. Formally, let $\mathcal{K}$ be a convex set in the vector space $\mathbf{V}$, and let L be a linear transformation from $\mathbf{V}$ to $\mathbf{W}$; then the set $L(\mathcal{K})$ is also convex. To prove this result, let $\mathbf{w}_1$ and $\mathbf{w}_2$ be any two vectors in

$\mathbf{W}$, that are also in $L(\mathcal{K})$. By Equation (1.6), $L(\mathcal{K})$ is convex if and only if $\alpha_1\mathbf{w}_1 + (1-\alpha_1)\mathbf{w}_2$ is also in $L(\mathcal{K})$, for any α between 0 and 1. Since $\mathbf{w}_1$ is in $L(\mathcal{K})$, there must be at least one $\mathbf{v}_1$ in $\mathcal{K}$ such that $L(\mathbf{v}_1) = \mathbf{w}_1$. Similarly there is at least one $\mathbf{v}_2$ such that $L(\mathbf{v}_2) = \mathbf{w}_2$. The convexity of $\mathcal{K}$ then implies that $\alpha\mathbf{v}_1 + (1-\alpha)\mathbf{v}_2$ is in $\mathcal{K}$. By the linearity of L,

$$L(\alpha\mathbf{v}_1 + (1-\alpha)\mathbf{v}_2) = \alpha L(\mathbf{v}_1) + (1-\alpha)L(\mathbf{v}_2) \tag{1.14}$$
$$= \alpha\mathbf{w}_1 + (1-\alpha)\mathbf{w}_2, \tag{1.15}$$

implying that $\alpha\mathbf{w}_1 + (1-\alpha)\mathbf{w}_2$ is in $L(\mathcal{K})$. This statement is sufficient to prove that $L(\mathcal{K})$ is convex, as was to be shown.

A general linear transformation can significantly distort a shape in a vector space. For example, the linear image of a circular disc could be a very elongated elliptical disc, and the linear image of the diamond in Figure 1.1 could be a square or a parallelogram with a different orientation. Despite these distortions, the convexity of the shapes is preserved: a diamond could not be sent to a chevron, for example.

Not only do linear transformations preserve convexity, but they also preserve some of the internal structure of a convex set. Suppose, for example, that $\mathcal{K}$ is the convex hull of a set of vectors $\mathcal{P} = \{\mathbf{v}_1, \mathbf{v}_2, ...\mathbf{v}_m\}$ in $\mathbf{V}$. Then it is easy to see that $L(\mathcal{K})$ is the convex hull of $L(\mathcal{P})$, in the vector space $\mathbf{W}$. Thus $\{L(\mathbf{v}_1), L(\mathbf{v}_2), ...L(\mathbf{v_m})\}$ is a generating set for $L(\mathcal{K})$. Note, however, that the minimality of a generating set might not be preserved. That is, even though $\{\mathbf{v}_1, \mathbf{v}_2, ...\mathbf{v}_m\}$ is the smallest set in $\mathbf{V}$ that can generate $\mathcal{K}$, it does not follow that $\{L(\mathbf{v}_1), L(\mathbf{v}_2), ...L(\mathbf{v_m})\}$ is the smallest set in $\mathbf{W}$ that can generate $L(\mathcal{K})$. For example, the vertices of a cube are a minimal generating set for the cube. Under a linear transformation, one face of the cube could be preserved, but a vertex of the opposite face could be sent to the interior of the original face. The linear image would then be a square, which, like the original cube, is convex. A square, however, has only four minimal generators (its vertices), while there are eight images of the cube's vertices. While those eight image points generate the square, any image points in the interior of the square are superfluous, so the image points as a whole are not a minimal generating set.

Linear transformations similarly preserve much of the structure of convex cones. Suppose that $\mathcal{K}$ is a convex cone in $\mathbf{V}$, and that $\mathcal{K}$ is generated by the rays corresponding to the vectors $\{\mathbf{v}_1, \mathbf{v}_2, ...\mathbf{v}_m\}$ in $\mathbf{V}$. Then it is straightforward to show not only that $L(\mathcal{K})$ is a convex cone in **W**, but also that the rays generated by $\{L(\mathbf{v}_1), L(\mathbf{v}_2), ...L(\mathbf{v_m})\}$ are a generating set for the new convex cone. Again, some of those image rays might be superfluous, if L sends one of the $\mathbf{v}_i$'s to the interior of the new cone. The fact that linear transformations preserve convex cones will be directly relevant to colour investigations in Chapter 4, when we investigate a transformation from the vector space of radiometric functions to the vector space of perceived colours.

1.5 Euclidean Notions

Although this book emphasizes geometric naturalness in its presentation, some very natural Euclidean notions, such as distance and angle, do not appear, and in fact are not needed. Somewhat surprisingly, the combinatorial structure—the fact that certain vectors sum to certain other vectors—and the linear relationships between vector spaces contain all the information needed for developing colour matching mathematically. A general vector space, in fact, has no notion of length or angle, unless one imposes some additional properties, such as an inner product. The reader should be warned against using the standard dot product to define magnitudes or perpendicular projections, or to interpret conclusions in those terms: the dot product implicitly assumes that an orthonormal basis exists, while the vector spaces used in this book provide no concepts of right angle or unit length.

1.6 Chapter Summary

This chapter has introduced the concept of convexity, in a vector space setting. The following standard definitions and results of linear algebra were used:

1. A linear functional F is a linear transformation from a vector space V to the real numbers $\mathbb{R}$,
2. A hyperplane $\mathcal{H}$ (or more properly, an *affine* hyperplane) in a vector space V of dimension n is an $(n-1)$-dimensional vector subspace, that has been translated by the addition of an arbitrary vector $\mathbf{v}$,
3. Let α be a real number, and let F be a non-degenerate functional (i.e. F takes on some non-zero values) on V. Then the pre-image $F^{-1}(\alpha)$ is a hyperplane $\mathcal{H}$ in V,

4. Suppose F is a non-degenerate functional on V. Then the set of all hyperplanes, that are pre-images under F of an arbitrary real number α, foliate V : each vector $\mathbf{v}$ in V belongs to exactly one hyperplane. The hyperplanes can be indexed by assigning the value $F(\mathcal{H})$ to each hyperplane $\mathcal{H}$ in the foliation. Under this indexing, the hyperplane $F^{-1}(0)$ contains the origin, and the hyperplanes with positive indices are stacked linearly, in the order given by their indices, away from the origin; a symmetrical result holds for hyperplanes with negative indices,
5. If two functionals produce the same hyperplane foliation, then the two functionals are scalar multiples of each other.

A vector space $\mathbf{V}$ contains sufficient structure to define and construct convex sets. The following standard definitions and results were introduced:

1. Given two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ in $\mathbf{V}$, the line segment between them is the set defined by
$$\{\alpha_1\mathbf{v}_1 + (1-\alpha_1)\mathbf{v}_2 | 0 \leq \alpha_1 \leq 1\}. \tag{1.16}$$
Using α as an index parametrizes the line segment continuously, as an image of the real interval $[0,1]$,
2. A subset $\mathcal{K}$ of V is convex if, for every $\mathbf{v}_1$ and $\mathbf{v}_2$ in $\mathcal{K}$, the line segment between $\mathbf{v}_1$ and $\mathbf{v}_2$ is contained in $\mathcal{K}$,
3. Let $\mathcal{Q}$ be a set of points in V. Then the convex hull of $\mathcal{Q}$, denoted hull$(\mathcal{Q})$, is the smallest convex set that contains every point in $\mathcal{Q}$. As a set, hull$(\mathcal{Q})$ is given by
$$\text{hull}(\mathcal{Q}) = \left\{ \sum_{i=1}^{m} \alpha_i\mathbf{v}_i \middle| \mathbf{v}_i \in \mathcal{Q}\ \forall i, 0 \leq \alpha_i \leq 1\ \forall i \text{ and } \sum_{i=1}^{m} \alpha_i = 1 \right\}. \tag{1.17}$$
Algebraically, hull$(\mathcal{Q})$ is the set of convex combinations of arbitrary finite subsets of $\mathcal{Q}$, where a convex combination of vectors is a linear combination of those vectors, whose coefficients are between 0 and 1 inclusive, and which sum to 1,
4. Let $\mathcal{Q}$ be a set of points in V. Then the convex cone of $\mathcal{Q}$, denoted cone$(\mathcal{Q})$, is the set
$$\text{cone}(\mathcal{Q}) = \left\{ \sum_{i=1}^{m} \alpha_i\mathbf{v}_i \middle|, \mathbf{v}_i \in \mathcal{Q},\ 0 \leq \alpha_i \right\}, \tag{1.18}$$
5. A convex polytope (sometimes called just a polytope) is the convex hull of a finite set of points,

6. Geometrically, a convex polytope can be written as the convex hull of its vertices, which are a unique subset of the polytope.
7. Each vertex $\mathbf{v}$ of a convex polytope $\mathcal{P}$ is an exposed point, that is: there exists a hyperplane $\mathcal{H}$ such that $\mathcal{H}$ contains $\mathbf{v}$, but $\mathcal{H}$ contains no other point of $\mathcal{P}$.

Linear transformations affect convex subsets of an underlying vector space as follows:

1. Linear transformations preserve convexity. Suppose a linear transformation L goes from a vector space $\mathbf{V}$ to a vector space $\mathbf{W}$, and that $\mathcal{K}$ is a convex set in $\mathbf{V}$. Then $L(\mathcal{K})$ is a convex set in $\mathbf{W}$.
2. Furthermore, if $\mathcal{Q}$ is a set of points in $\mathbf{V}$, then a linear transformation L preserves some of the internal structure of convex sets:

$$L(\mathrm{hull}(\mathcal{Q})) = \mathrm{hull}(L(\mathcal{Q})), \text{ and} \tag{1.19}$$

$$L(\mathrm{cone}(\mathcal{Q})) = \mathrm{cone}(L(\mathcal{Q})). \tag{1.20}$$

Finally, the reader should keep in mind that a general vector space, and in particular the vector spaces treated in this book, have no concept of distance or angle, so the book's conclusions should not be understood in Euclidean terms. Instead, combinatorial and linear relationships provide all the structure needed.

Chapter 2

Zonohedra

2.1 Introduction

A *zonohedron* $\mathcal{Z}$ is a special kind of convex polytope, that consists of all the linear combinations of a set of vectors in $\mathbb{R}^3$, provided that the coefficients in each combination are between 0 and 1 inclusive. The set of such combinations is a special instance of the *Minkowski sum*, sometimes also called the *vector sum*. Zonohedra occur in colour science when simple colours, such as the primaries of an electronic display or single-wavelength radiometric functions, combine to produce new colours. Each simple colour can be represented as a vector in three-dimensional perceptual colour space. The vector for a colour combination is a linear combination of vectors for the simple colours, with the physical restriction that the coefficients in the combination cannot exceed 1. The set of all possible colour mixtures is thus a zonohedron. This chapter derives some geometric properties of zonohedra. Later chapters will use the geometric properties to derive further results about colour.

The Minkowski sum of two subsets of a vector space is a new subset, consisting of all sums of pairs of vectors, where the first vector belongs to the first subset and the second vector belongs to the second subset. This definition can easily be extended to include an arbitrary number of subsets, rather than just two, and the order of the subsets is immaterial. Typically, the Minkowski sum is applied to convex subsets, in which case it produces another convex subset. Apart from its simple algebraic definition, the Minkowski sum also has a concrete geometric interpretation: the Minkowski sum of two subsets is, up to a translation, the set produced by placing a copy of the first subset over every point of the second subset. More dynamically, one subset is “swept” over the entirety of the other subset, and any point that is covered by the sweeping action is contained in the Minkowski sum.

Geometrically, a vector can be interpreted as either an isolated point in a vector space, or as the line segment that joins that point to the origin. The Minkowski sum can be applied to a set of vectors, called *generating vectors*, by interpreting those vectors as line segments, which are convex subsets of the overall space. In three dimensions, the Minkowski sum of a set of vectors is called a zonohedron.

The Minkowski sum of two vectors is the filled parallelogram that is swept out when one vector slides over the other. Visually, one can picture the first vector moving continuously as its tail slides along the line segment corresponding to the second vector. Then sweep a third vector, that is not coplanar with the first two, over the parallelogram, producing a solid parallelepiped. This solid is the Minkowski sum of all three vectors, and is a zonohedron. In $\mathbb{R}^3$, a fourth vector could be Minkowski-summed with the parallelepiped, producing a new, more complicated, zonohedron, and in fact any number of vectors could be similarly Minkowki-summed to produce a zonohedron.

Zonohedra have many useful properties. For instance, a zonohedron’s faces are parallelograms, and its edges are translated copies of the generating vectors. Each vertex can be written uniquely as the sum of all the generating vectors that lie on one side of a plane through the origin. A zonohedron is encircled by bands called *zones*, where each zone corresponds to a generating vector. Furthermore, a zonohedron is centrally symmetric. This chapter constructs Minkowski sums and zonohedra intuitively, and derives such properties. While zonohedra generalize naturally to *zonotopes* in arbitrary dimensions, we will restrict ourselves to $\mathbb{R}^3$, which is sufficient for colour matching. While no results in this chapter are new, their systematic and concrete presentation from first principles is believed to be the only one available so far.

2.2 The Minkowski Sum

2.2.1 Definition and Properties

In visual terms, the Minkowski sum, sometimes also called the *vector sum*, can be pictured as a way of adding shapes or solids, or, indeed, arbitrary subsets of a vector space. It is a simple, intuitive construction, both algebraically and geometrically. Formally, let $\mathcal{A}$ and $\mathcal{B}$ be two non-empty subsets of a vector space $\mathbb{R}^n$. Then their Minkowski sum, denoted $\oplus$, is defined as

$$\mathcal{A} \oplus \mathcal{B} = \left\{ \mathbf{v}_{\mathcal{A}} + \mathbf{v}_{\mathcal{B}} \middle| \mathbf{v}_{\mathcal{A}} \in \mathcal{A} \text{ and } \mathbf{v}_{\mathcal{B}} \in \mathcal{B} \right\}. \tag{2.1}$$

In this equation, $\mathbf{v}_{\mathcal{A}}$ and $\mathbf{v}_{\mathcal{B}}$ are both vectors in $\mathbb{R}^n$, and the plus sign on the right-hand side indicates vector addition.

The commutativity of the vector addition in Equation (2.1) implies that the Minkowski sum is commutative, so $\mathcal{A} \oplus \mathcal{B} = \mathcal{B} \oplus \mathcal{A}$. Furthermore, the associativity of vector addition implies that $(\mathcal{A} \oplus \mathcal{B}) \oplus \mathcal{C} = \mathcal{A} \oplus (\mathcal{B} \oplus \mathcal{C})$, if there were a third subset $\mathcal{C}$. Together, commutativity and associativity imply that the Minkowski sum of any number of summands is well-defined, regardless of the order of those summands.

Although the Minkowski sum is defined for arbitrary subsets $\mathcal{A}$ and $\mathcal{B}$, it is most commonly applied to convex subsets, which will be the case of interest for this book. When $\mathcal{A}$ and $\mathcal{B}$ are both convex, then $\mathcal{A} \oplus \mathcal{B}$ is also convex. To demonstrate this statement, let $\mathbf{v}$ and $\mathbf{w}$ be two vectors in $\mathcal{A} \oplus \mathcal{B}$, and let α be a scalar between 0 and 1; convexity will follow if we can show that $\alpha\mathbf{v} + (1-\alpha)\mathbf{w} \in \mathcal{A} \oplus \mathcal{B}$. Since $\mathbf{v}$ and $\mathbf{w}$ are both in the Minkowski sum of $\mathcal{A}$ and $\mathcal{B}$, there must exist vectors such that

$$\mathbf{v} = \mathbf{v}_{\mathcal{A}} + \mathbf{v}_{\mathcal{B}}, \tag{2.2}$$

$$\mathbf{w} = \mathbf{w}_{\mathcal{A}} + \mathbf{w}_{\mathcal{B}}, \tag{2.3}$$

where the subscripts indicate which subsets the vectors belong to. Since $\mathcal{A}$ and $\mathcal{B}$ are both convex, it follows that

$$\alpha\mathbf{v}_{\mathcal{A}} + (1-\alpha)\mathbf{w}_{\mathcal{A}} \in \mathcal{A}, \tag{2.4}$$

$$\alpha\mathbf{v}_{\mathcal{B}} + (1-\alpha)\mathbf{w}_{\mathcal{B}} \in \mathcal{B}. \tag{2.5}$$

Equation (2.4) gives a vector in $\mathcal{A}$, and Equation (2.5) gives a vector in $\mathcal{B}$, so their sum is in $\mathcal{A} \oplus \mathcal{B}$:

$$\alpha(\mathbf{v}_{\mathcal{A}} + \mathbf{v}_{\mathcal{B}}) + (1-\alpha)(\mathbf{w}_{\mathcal{A}} + \mathbf{w}_{\mathcal{B}}) \in \mathcal{A} \oplus \mathcal{B}. \tag{2.6}$$

Substituting Equations (2.2) and (2.3) into Equation (2.6) gives

$$\alpha\mathbf{v} + (1-\alpha)\mathbf{w} \in \mathcal{A} \oplus \mathcal{B}, \tag{2.7}$$

implying that $\mathcal{A} \oplus \mathcal{B}$ is convex, as was to be shown.

2.2.2 Some Two-Dimensional Examples

Though Minkowski sums are defined in arbitrary dimensions, some simple two-dimensional examples provide geometric motivation. For the first example, let $\mathcal{A}$ be a triangle in the positive quadrant of $\mathbb{R}^2$, and let $\mathcal{B}$ be a small disc centered on the origin, as shown in the left of Figure 2.1. Let $\mathbf{v}_{\mathcal{A}}$ be a point in the triangle. It follows easily from the definition that $\mathcal{A} \oplus \mathcal{B}$ is the union of all the sets $\mathbf{v}_{\mathcal{A}} + \mathcal{B}$, over all points $\mathbf{v}_{\mathcal{A}}$ in $\mathcal{A}$. The set $\mathbf{v}_{\mathcal{A}} + \mathcal{B}$ is a small disc that is centered on $\mathbf{v}_{\mathcal{A}}$. This disc is a translation of the set $\mathcal{B}$, and is contained in $\mathcal{A} \oplus \mathcal{B}$.

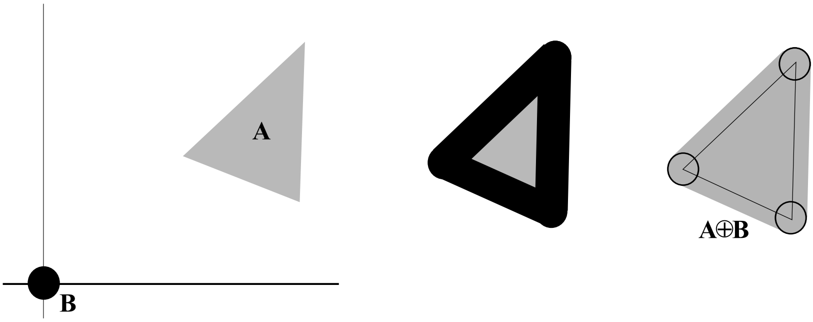

Figure 2.1: An Example of a Minkowski Sum

If $\mathbf{v}_{\mathcal{A}}$ is well in the interior of $\mathcal{A}$, then $\mathbf{v}_{\mathcal{A}} + \mathcal{B}$ is contained completely in $\mathcal{A}$. If $\mathbf{v}_{\mathcal{A}}$ is on the perimeter of $\mathcal{A}$, however, then $\mathbf{v}_{\mathcal{A}} + \mathcal{B}$ extends somewhat outside $\mathcal{A}$. The middle of Figure 2.1 shows the translated discs along the triangle's edges; these discs are included in the Minkowski sum. One can picture the disc sliding along each edge, and sweeping out a band on either side of the edge; the band on the triangle's exterior produces a parallel contour that expands the triangle. Once the disc reaches a vertex, the band changes direction abruptly, and its contour is now an arc of the disc rather than a straight line.

The Minkowski sum is the union of the triangle's interior, the bands along the three edges, and the circular sectors at the three vertices, as shown on the right in Figure 2.1. A unifying interpretation is that the Minkowski sum consists of all the points the disc would cover, if it were swept over every point on the triangle. Note that the sweeping is done by fixing a point on the disc, in this case its center, and letting that center sweep over the triangle. The origin was chosen as a convenient location for this example because, when the center is at the origin, adding it to $\mathcal{A}$ does not move $\mathcal{A}$.

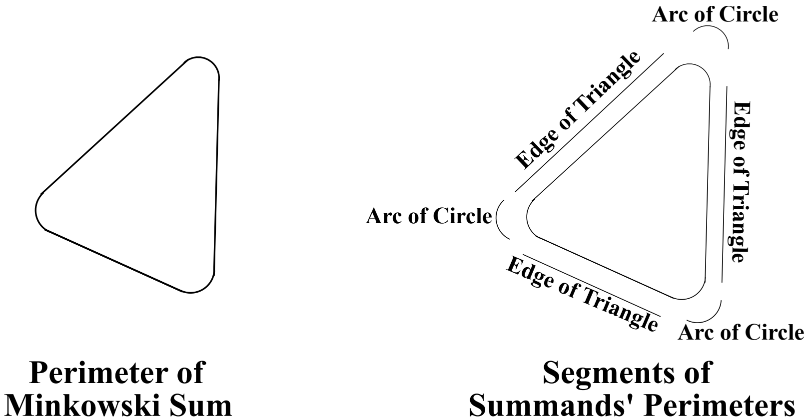

Figure 2.2: The Perimeter of a Minkowski Sum

An interesting property, which applies when $\mathcal{A}$ and $\mathcal{B}$ are convex sets, is that the perimeter of the Minkowski sum consists of reassembled pieces of the perimeters of the two summand sets. Figure 2.2 shows the pieces for the Minkowski sum in Figure 2.1. The three straight pieces are just translations of the triangle's three edges, from which one could reconstruct the triangle. Similarly the three rounded pieces are just translations of segments of the disc's circumference, from which one could reconstruct the disc.

In the special case when one of the summand sets is a single vector, the Minkowski sum is just a translation of the other set by that vector. In the trivial case when one summand set consists solely of the origin, the Minkowski sum is just an unmodified copy of the other set. Suppose that the disc in Figure 2.1 was not centered at the origin, but somewhere else, as shown in Figure 2.3. Then the new Minkowki sum would just be a shifted version of the original sum, as shown on the right, but would be otherwise identical to the Minkowski sum in Figure 2.1. A translated version of the triangle would produce a similar shift. Often the shape of the Minkowski sum is of more interest than its location. In that case, the sum can be "normalized" by translating it to a convenient location, or the shape can be studied without regard to location.

The discussion so far has been in two dimensions, but should allow the Minkowski sum in three dimensions, and even in arbitrary dimen-

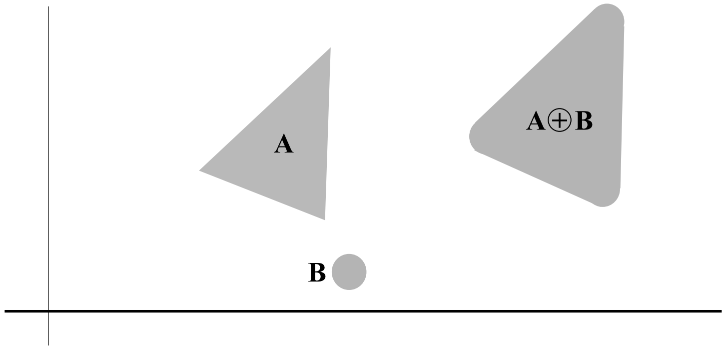

Figure 2.3: Location vs. Shape for a Minkowski Sum

sions, to be easily visualized. The sum has been interpreted as the set of all the points covered when one summand is swept over the other summand, and this interpretation applies to $\mathbb{R}^3$ just as well as $\mathbb{R}^2$, except that areas now become volumes. In $\mathbb{R}^n$, when $n > 3$, volumes can be replaced with their higher-dimensional analogues.

2.3 Zonohedra

Zonohedra are a special kind of Minkowski sum, which occur when all the summands are vectors in $\mathbb{R}^3$. In this context, a vector in $\mathbb{R}^3$, rather than being thought of as a point in space, is thought of as the line segment joining that point to the origin; the line segment contains all the points that would be covered if the vector were drawn as an arrow with its tail at the zero vector. When speaking of zonohedra, the terms *vector* and *segment* will be used interchangeably. Given a finite set of vectors in $\mathbb{R}^3$, then, the Minkowski sum of the resulting line segments is called a *zonohedron*. A similar Minkowski sum is called a *zonogon* in $\mathbb{R}^2$, and a *zonotope* in an arbitrary $\mathbb{R}^n$.

This section and those following will derive many interesting properties of zonohedra. For instance, a zonohedron's faces are, in the generic case, all parallelograms, and each edge is a translated copy of a generating vector. A zonohedron is encircled by many crisscrossing bands called *zones*. A zone consists of all the parallelogram faces, such that one edge of the parallelogram is a translation of a given generating vector. Every vertex of a zonohedron is the sum of all the generating vectors that are on one side of a plane through the origin. Furthermore, a zonohedron is centrally symmetric: its shape at the origin and its shape at the farthest vertex are reversed, but otherwise identical. *Cyclic* zonohedra are a useful subclass, whose zones take a convenient form; cyclic zonohedra arise when the generating vectors are all on the boundary of a convex cone.

Zonotopes occur naturally when a number of ingredients, in a general sense, are combined, again in a general sense, to make a new mixture, and the maximum quantity of each ingredient is limited. Each ingredient can be represented by a vector whose head represents the maximum possible quantity of that ingredient while the tail, at the origin, means that none of that ingredient is used. The set of all possible mixtures is the Minkowski sum of the ingredients' vectors, and is thus a zonotope.

Later in this book, surface colours and other objects of interest will be expressed as just such mixtures. At each wavelength, a surface colour reflects between 0 and 100 percent of the incoming light. The colour corresponding to each wavelength is a vector in a three-dimensional perceptual space. The contribution at each wavelength is an "ingredient" in the overall colour, which is the "mixture" of the reflected light at each wavelength. The set of all surface colours, when viewed in a certain illumination, is called an *object-colour solid*, and this interpretation implies that it is a zonohedron. By similar constructions, an electronic *display gamut*, which is the set of all the colours that that display can produce, is also a zonohedron. The zonohedral form of these objects of colour science is the central insight of this book. Later chapters will use the formalism of this chapter to construct colour zonohedra, and then derive results in colour science from zonohedral properties.

2.3.1 Definitions

Suppose we denote a set of m vectors, or equivalently m line segments starting at the origin, in $\mathbb{R}^n$, by

$$\mathcal{G} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_m\}. \tag{2.8}$$

Then the *zonotope* $\mathcal{Z}$ generated by $\mathcal{G}$ is defined to be the Minkowski sum of those segments:

$$\mathcal{Z}(\mathcal{G}) = \mathbf{v}_1 \oplus \mathbf{v}_2 \oplus ... \oplus \mathbf{v}_m. \tag{2.9}$$

$\mathcal{G}$ is referred to as the set of *generating vectors*, or just *generators*, for Z. When $n = 2$, $\mathbb{R}^n$ is the plane, and $\mathcal{Z}$ is called a *zonogon.* When $n = 3$, $\mathbb{R}^n$ is three-dimensional space, and $\mathcal{Z}$ is called a *zonohedron.*

Since the line segment corresponding to a vector $\mathbf{v}$ is given by the set

$$\{\alpha\mathbf{v} | 0 \leq \alpha \leq 1\}, \tag{2.10}$$

it follows that an equivalent definition for a zonotope is

$$\mathcal{Z}(\mathcal{G}) = \left\{ \sum_{i=1}^{m} \alpha_i \mathbf{v}_i \middle| 0 \leq \alpha_i \leq 1 \ \forall i \right\}. \tag{2.11}$$

This definition of a zonotope is similar to the definition of a convex hull given by Equation 1.7, except that the coefficients in a convex hull must sum to 1, while zonotope coefficients lack that restriction, although they must still be between 0 and 1. Such a linear combination, where all the coefficients are between 0 and 1 inclusive, is called a *zonal combination.* Equation (2.11) makes calculations easier, because the α_i's provide a convenient coordinatization. For geometric understanding, however, Equation (2.9) is preferred, because it indicates how a zonotope is constructed.

While the set of vectors $\mathcal{G}$ is arbitrary, this book will simplify calculations by further assuming that each vector in $\mathcal{G}$ is in the non-negative octant $\mathcal{O}$ relative to some basis. Then all the coordinates of the $\mathbf{v}_i$'s are non-negative, and the zonotope as a whole is contained in the non-negative octant. Such a zonotope is referred to as a *non-negative zonotope.* Non-negative zonotopes or zonohedra are adequate for colour science, whose geometric objects are non-negative by construction. This approach avoids technical complications, such as two generating vectors that are in exactly opposite directions.

One immediate conclusion, using the results of Section 2.2.1, is that a zonotope is always convex, since it is the Minkowski sum of line segments, which are themselves convex. Zonotopes, and in particular zonohedra, have considerably more structure, however, which this chapter will elucidate, after investigating some examples. The first examples will be some simple zonogons in $\mathbb{R}^2$, which will then be generalized to zonohedra in $\mathbb{R}^3$.

2.4 Construction of Zonogons

A zonogon is a two-dimensional zonotope. For a simple example, start with two vectors, $\mathbf{v}_1$ and $\mathbf{v}_2$, in the non-negative octant of $\mathbb{R}^2$, as shown on the left of Figure 2.4. The Minkowski sum of the two vectors, or of their corresponding segments, is the area swept out when the first segment is moved continuously, with its tail following the entire length of the second segment. The right side of Figure 2.4 shows the result: the zonogon is a parallelogram with one vertex at the origin.

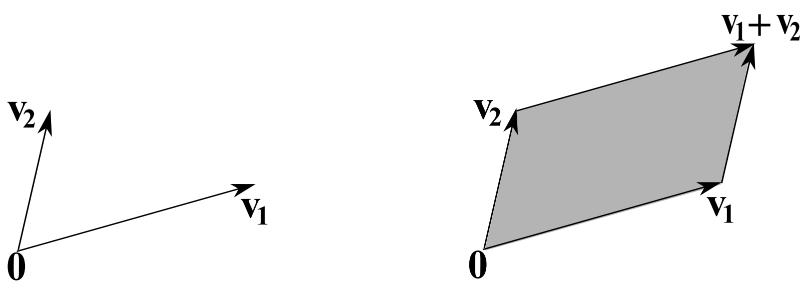

Figure 2.4: A Zonogon with Two Generating Vectors

This construction is symmetric: if the second vector swept over the first, the same parallelogram would result. Algebraically, the symmetry results from the commutativity of the Minkowski sum. The construction also increases area: even though each segment has zero area, their Minkowski sum has positive area. In this simple case, Equation (2.11) indicates a coordinatization for the parallelogram: any point is a unique linear combination of $\mathbf{v}_1$ and $\mathbf{v}_2$, both of whose coefficients are between 0 and 1.

Now let us add a third vector, $\mathbf{v}_3$, as shown in the left of Figure 2.5, and find the zonogon generated by all three vectors. Since the Minkowski sum is both commutative and associative, the order of its summands is immaterial, and we can write

$$\mathbf{v}_1 \oplus \mathbf{v}_2 \oplus \mathbf{v}_3 = (\mathbf{v}_1 \oplus \mathbf{v}_2) \oplus \mathbf{v}_3. \tag{2.12}$$

The Minkowski sum $\mathbf{v}_1 \oplus \mathbf{v}_2$ already appears as a parallelogram in Figure 2.4, so let $\mathbf{v}_3$ sweep over that parallelogram. The resulting zonogon is the irregular, filled hexagon on the right in Figure 2.5.

This zonogon construction is again symmetric, this time in three vectors instead of two.. Had we started with the parallelogram $\mathbf{v}_1 \oplus \mathbf{v}_3$,

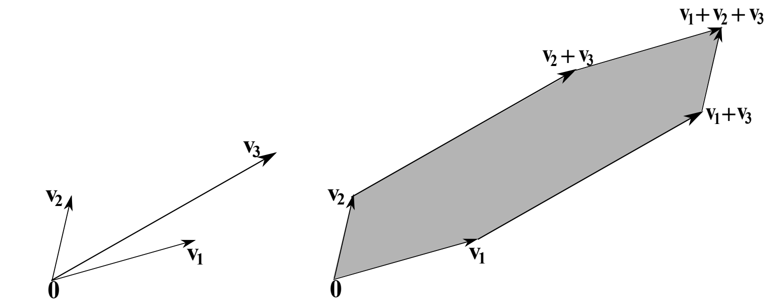

Figure 2.5: A Zonogon with Three Generating Vectors

for example, and then swept the vector $\mathbf{v}_2$ over it, we would have produced the same hexagon. The α_i's from Equation (2.11) again suggest coordinates for the points of the hexagon, but this time the coordinatization is not unique, because the generating vectors are not linearly independent. Different coordinatization schemes correspond to different decompositions of the hexagon into parallelograms; Figure 2.6 shows two decompositions.

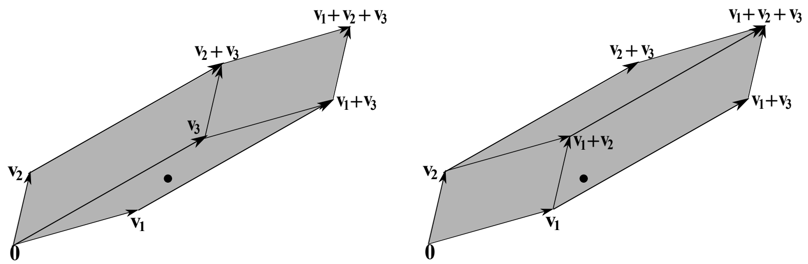

Figure 2.6: Two Decompositions of the Zonogon from Figure 2.5

The marked point would have coordinates

$$[\alpha_1, \alpha_2, \alpha_3] = [1/2, 0, 1/2] \tag{2.13}$$

in the first decomposition, but coordinates

$$[\alpha_1, \alpha_2, \alpha_3] = [1, 1/3, 1/8] \tag{2.14}$$

in the second decomposition. While decompositions provide a convenient coordinate framework, other sets of α's, that are not tied to a decomposition, are also possible coordinates for the indicated point.

Figure 2.7 shows another interesting property: the hexagon's perimeter consists entirely of translated copies of the generating vectors. Each vector appears twice, on opposite sides of the hexagon. In two dimensions, a set of non-negative generating vectors can be placed in clockwise order. From the left side of Figure 2.5, for instance, the clockwise ordering is $\mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_1$. When traced out from the origin, the boundary consists of two copies of the generators, in clockwise order. In Figure 2.7, the edge sequence is $\mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_1$, which is just the clockwise ordering listed twice.

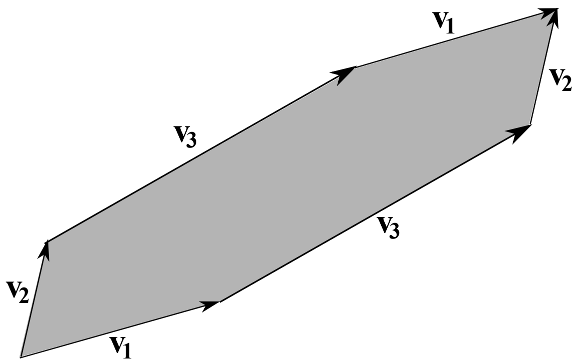

Figure 2.7: Translated Generating Vectors Make Up a Zonogon's Boundary

This interpretation also shows that the coordinates of the vertices have a special form: each vertex can be written as a sum of generating vectors. In terms of Equation (2.11), a vertex can be written as a linear combination of generators, such that each coefficient α_i in the combination is either 0 or 1. Formally, define the *nodes* of a zonotope as the set

$$\mathcal{N}(\mathcal{Z}) = \left\{ \sum_{i=1}^{m} \alpha_i \mathbf{v}_i \middle| \alpha_i \in \{0, 1\} \ \forall i \right\}. \tag{2.15}$$

Although every vertex is a node, it is not the case that every node is a vertex; nodes can also appear in the interior of the zonogon. Figure 2.6 shows, for instance, that $\mathbf{v}_3$ and $\mathbf{v}_1 + \mathbf{v}_2$ are both interior nodes. We will later use hyperplanes to derive expressions for a zonotope's vertices.

As another consequence of the non-negative restriction, a zonogon, and more generally a zonotope, has a unique minimal point, the origin, which occurs when every α_i is 0, and a unique maximal, or *terminal*, point, which occurs when every α_i is 1. Even though a general vector space implies no notion of distance, the terminal point is, loosely speaking, the "farthest" point from the origin. The *center* of the zonogon is halfway between the origin and the terminal point. The zonogon is symmetric around the center: a 180° rotation about the center would map the zonogon to itself.

From the examples already given, adding further generating vectors to produce new zonogons is straightforward: simply sweep the new vector over the zonogon for the previous vectors. The result will be a larger, more complicated, polygon that is still convex and centrally symmetric, but whose boundary now includes two copies of the new vector.

2.5 Vectors in General Position

A special case occurs when one generator is a scalar multiple of another generator. In this case, an edge of the zonogon consists of those two vectors laid end to end, and a node that is not a vertex appears on the boundary. This configuration makes sense combinatorially, but is redundant geometrically, so we will use the concept of general position to avoid it.

A set of vectors $\mathcal{G}$ in $\mathbf{R}^n$ is said to be in *general position* if every subset of n vectors is linearly independent. Geometrically, a set of vectors in $\mathbf{R}^2$ is in general position if no two of them are scalar multiples of each other, and vectors in $\mathbf{R}^3$ are in general position if no three of them are coplanar; analogues hold for higher dimensions. Requiring $\mathcal{G}$ to be in general position, or checking that an empirically found $\mathcal{G}$ is in general position, sidesteps many technical complications and allows for simpler descriptions. If generators in $\mathbb{R}^2$ are in general position, for instance, then each edge of a zonogon is a translated copy of a single generator. Otherwise, an edge might consist of two generators, laid end to end. The property of being in general position is generic: an arbitrary set is most likely already in general position, and, if it is not, a negligible adjustment will put the set in general position.

In the colour science applications to be studied, the generating vectors are determined by empirical measurements, which are only accurate to some ε, so there is no information lost in adjusting any of the measurements by an amount that is much smaller than ε. As a result, we can always obtain a set in general position, that agrees well with the measured data. Strictly speaking, any measurement uncertainty, no matter how small, makes it impossible to determine that two measured vectors are scalar multiples of each other, or that three measured vectors are coplanar; decisions about linear dependence therefore proceed from theory or modeling. In fact, we will see a case in which a set of vectors is nearly, but not quite, coplanar, and it will not be obvious whether the non-coplanarity is genuine, or an artifact of data smoothing.

2.6 Construction of Zonohedra

A zonohedron is constructed similarly to a zonogon, but in three dimensions instead of two. Like the previous section, we will build a non-negative zonohedron generator by generator, letting interesting properties emerge naturally.

2.6.1 An Example

Begin with two generators in $\mathbb{R}^3$. The zonohedron they generate is just the parallelogram shown in Figure 2.8, which is similar to Figure 2.4. That parallelogram is located in the plane spanned by the two generators. Now add a third generator, $\mathbf{v}_3$, such that the three vectors are in general position, and are thus not coplanar. The Minkowski sum can still be found by sweeping $\mathbf{v}_3$ over every point of the parallelogram formed by the first two generators. The result is the solid parallelepiped shown in Figure 2.9. Each edge of the parallelepiped is a translated copy of one of the generating vectors. The faces of the parallelepiped are parallelograms, three of which meet at the origin. Each such parallelogram is the Minkowski sum of two of the generators. Like any Minkowski sum, the construction of the parallelogram is symmetric. We could have started with any parallelogram, and swept the remaining generator over it, to create the same final parallelepiped.

Now suppose we add a fourth generator, $\mathbf{v}_4$. To construct the new zonohedron, let the fourth generator sweep over the current parallelepiped, producing the solid body shown in Figure 2.10. The fourth generator is drawn as a dotted line starting at the origin—the generating vector itself does not appear on the boundary of the zonohedron, though translated copies of it do.

Figure 2.11 shows another interpretation of the construction of the zonohedron in Figure 2.10. Suppose the boundary of the paral-

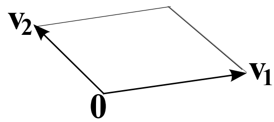

Figure 2.8: A Zonohedral Parallelogram

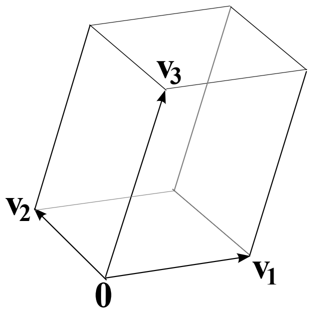

Figure 2.9: A Zonohedral Parallelepiped

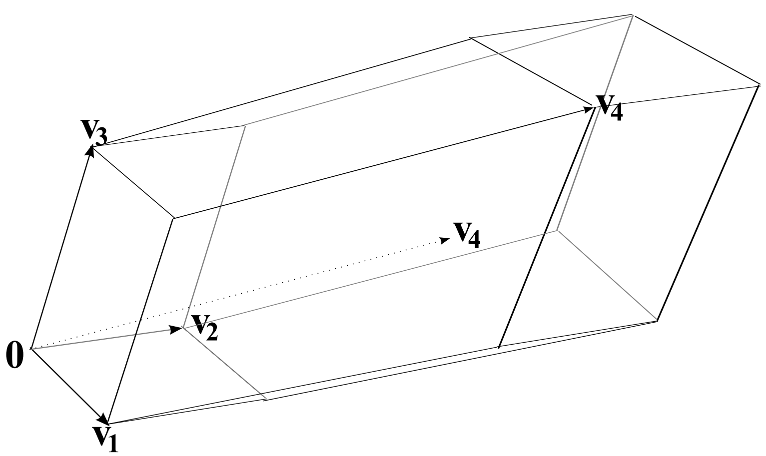

Figure 2.10: A Zonohedron with Four Generators

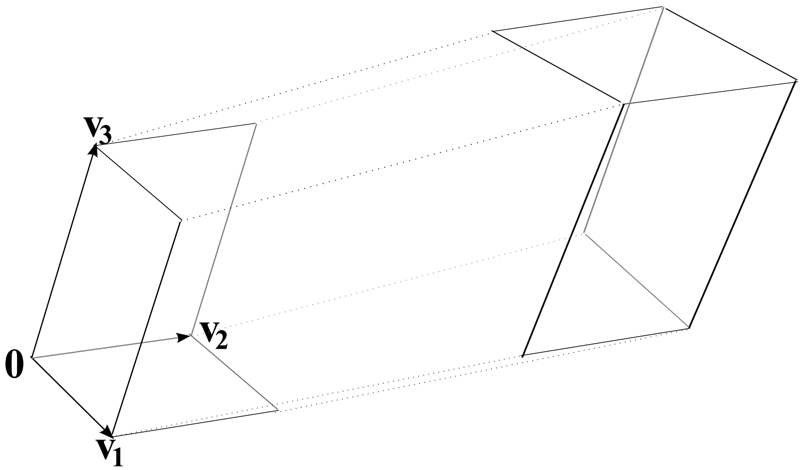

Figure 2.11: Another Construction of the Zonohedron in Figure 2.10

lelepiped in Figure 2.9 were broken up into two sections, with each section containing three faces. Then one of those sections could be translated away from the other section, along the dotted lines, by the vector $\mathbf{v}_4$, as shown in Figure 2.11. The total volume swept out by this translation would comprise the zonohedron in Figure 2.10, and it is easy to see how copies of $\mathbf{v}_4$ appear as edges in that zonohedron.

Clearly this construction could be continued indefinitely, with as many generators as desired. The construction also makes clear the effects of a generator's magnitude, versus the effects of its direction. Suppose the generator $\mathbf{v}_4$ maintained its direction, but that it was multiplied by 2, doubling its magnitude. Then the construction in Figure 2.11 would still hold, but the two pieces of the parallelepiped would be twice as far apart. Similarly, shrinking $\mathbf{v}_4$, without changing its direction, would move the two pieces closer together. The relationships between the zonohedron's vertices, edges, and faces, however, would be unaltered; these relationships are sometimes referred to as its *combinatorial structure*, and have historically been a major motivation for studying zonohedra.

Changing the direction of $\mathbf{v}_4$, on the other hand, could require breaking up the parallelepiped into two new pieces, different from the previous ones, and could change the combinatorial structure. Algebraically, the current $\mathbf{v}_4$ is a positive linear combination of the other three generators, so it is "inside" the parallelepiped. The final zonohedron thus has three edges meeting at the origin. If $\mathbf{v}_4$ was not a positive linear combination, then it would be "outside," and four edges would meet at the origin. These considerations will become important later when discussing cyclic zonohedra.

2.6.2 Zonohedral Generators in General Position

Recall that a set of generating vectors in an n-dimensional vector space is in general position if every subset of n vectors is linearly independent. In three dimensions, the general position requirement avoids the case where one face of a zonohedron consists of three or more generators; instead, each face is a parallelogram, each edge of which is a translated copy of a generator. Since a parallelogram's four edges break up into two pairs of parallel vectors, exactly two generators are needed to specify each parallelogram face.

To see the relevance of general position, begin with the six-sided zonogon shown in Figure 2.5. The three generators shown on the left all lie in $\mathbb{R}^2$. Extend this zonogon to a zonohedron by adding a fourth generator, which does not lie in the plane of the first three generators. Figure 2.12 shows the result, with the original zonogon in grey. The zonohedron is constructed by sweeping $\mathbf{v}_4$ over that zonogon, resulting in a translated copy, that is also shown in grey. The first three vectors are in general position in $\mathbb{R}^2$, because no two of them form a linearly dependent set. They are not in general position in $\mathbb{R}^3$, however, because the subset consisting of all three of them *is* linearly dependent. Geometrically, the lack of general position prevents the grey face of the zonohedron from being a parallelogram.

The construction method shows that a zonohedron's faces are all parallelograms if and only if the generators are in general position. Formally, we will present a demonstration that uses induction on the number of generators. Suppose that there are k generators, and that the zonohedron they generate only has parallelograms for faces. Now add a $(k+1)^{\text{th}}$ generator, which does not form a linearly dependent set when combined with any two of the original generators. The new zonohedron is constructed, as in Figure 2.11, by breaking the original zonohedron into two pieces (with breaks occurring only along edges, not on faces), translating one piece by $\mathbf{v}_{k+1}$, and introducing a new set of faces between the pieces. By the construction, each new face must have two edges parallel to $\mathbf{v}_{k+1}$, and two edges that are parallel to one of the original edges. Because of the assumption of general position,

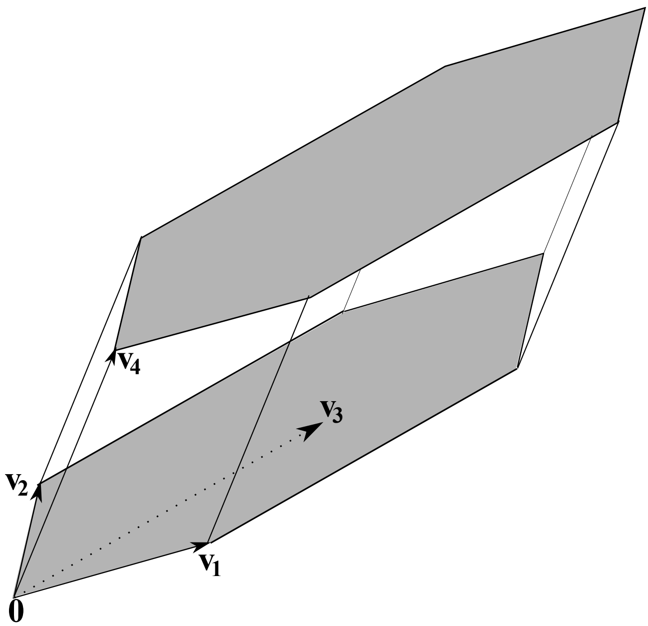

Figure 2.12: A Zonohedron when Generators are Not in General Position

the new vector cannot be in the affine plane spanned by any pair of the original vectors, so it cannot be in the plane of any of the zonohedron's original faces. Thus all the new faces are in different planes from all the old faces. Since each new face has two pairs of parallel sides, each new face is a parallelogram, in a distinct plane from all previous faces. General position is thus sufficient to insure that all the faces of the new zonohedron are parallelograms.

Conversely, suppose that the generators are not in general position, so that three generators are in the same plane. Since the order of the generators is irrelevant, start the construction of the zonohedron with those three generators. They will generate a non-parallelogram face, such as appears in Figure 2.5. This face will be preserved (or perhaps be extended, if other vectors are also in that plane) as further vectors are added, and so will appear in the final zonohedron. General position is thus necessary if all the faces of the final zonohedron are parallelograms. The sufficiency and necessity together imply the "if and only if" statement in the preceding paragraph, as was to be shown.

Apart from introducing many difficulties in computations, non-parallelogram faces are non-generic, in the sense discussed earlier, and can often be eliminated. In particular, when the generating vectors result from a measurement process, their coordinates are always somewhat uncertain, so no information is lost by adjusting one them by an amount that is orders of magnitude smaller than the measurement variability. If three vectors are coplanar, then modifying one vector very slightly moves it out of the plane spanned by the other two.

In Figure 2.12, for example, $\mathbf{v}_3$ could be raised a tiny amount so that it is no longer in the same plane as $\mathbf{v}_1$ and $\mathbf{v}_2$. Figure 2.13 shows the result. The point $\mathbf{v}_3 + \mathbf{v}_4$ in Figure 2.13 is now slightly above the upper grey face, while in Figure 2.12 it was right on that face. The irregular hexagon has been decomposed into three parallelogram faces, all in slightly different planes. Furthermore, the difference from the original zonohedron can be made as small as desired. Because of this flexibility, generators will frequently be assumed to be in general position for the rest of this book, even when it is not explicitly stated.

2.6.3 Zones

A *zone* is a sort of encircling band that lies on a zonohedron's surface. Formally, a zone corresponds to a generating vector and is defined to be the set of all the zonohedron's faces that contain a translated copy of that vector as one of their edges. Figure 2.11 motivates the concept. That figure inserts a band of six parallelograms between two separated halves of a parallelepiped. Each parallelogram face of the inserted band incorporates a copy of the vector $\mathbf{v}_4$ as an edge. The band of six parallelograms is the zone corresponding to the generator $\mathbf{v}_4$. Figure 2.14 shows the zone shaded in grey. In this view of the zonohedron, the lighter faces are seen from the outside; the darker faces, which are partially hidden by the lighter faces, can only be seen by looking through the zonohedron's outer surface.

Constructing a zone is straightforward. As an example, start with any copy of $\mathbf{v}_4$, as an edge $\mathcal{E}_1$ on the zonohedron. That edge will be the common edge to two parallelogram faces, $\mathcal{P}_1$ and $\mathcal{P}_2$. Include both those faces in the zone. The parallelogram face $\mathcal{P}_2$ will have a second edge $\mathcal{E}_2$, which is also a translated copy of $\mathbf{v}_4$, and which is located across $\mathcal{P}_2$, on the opposite side from the original edge $\mathcal{E}_1$. $\mathcal{E}_2$ will be the edge common to the parallelogram face $\mathcal{P}_2$, and to another parallelogram face $\mathcal{P}_3$. Include $\mathcal{P}_3$ in the zone, and continue, adding

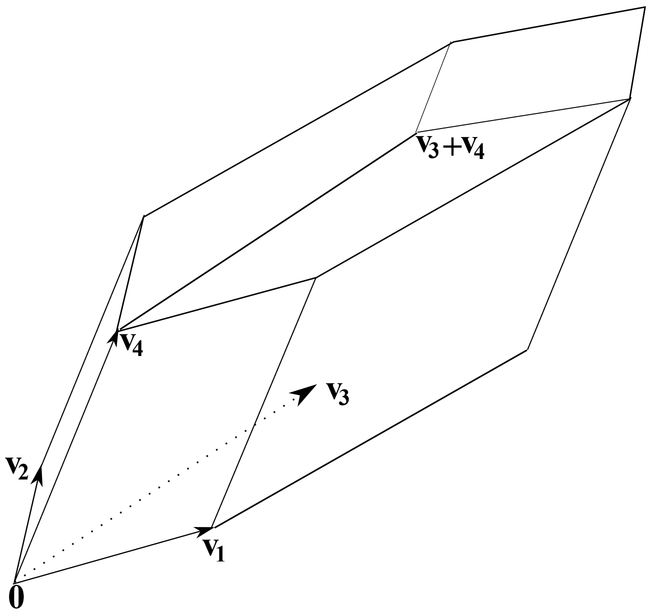

Figure 2.13: An Adjustment of Figure 2.12, such that Generators are in General Position

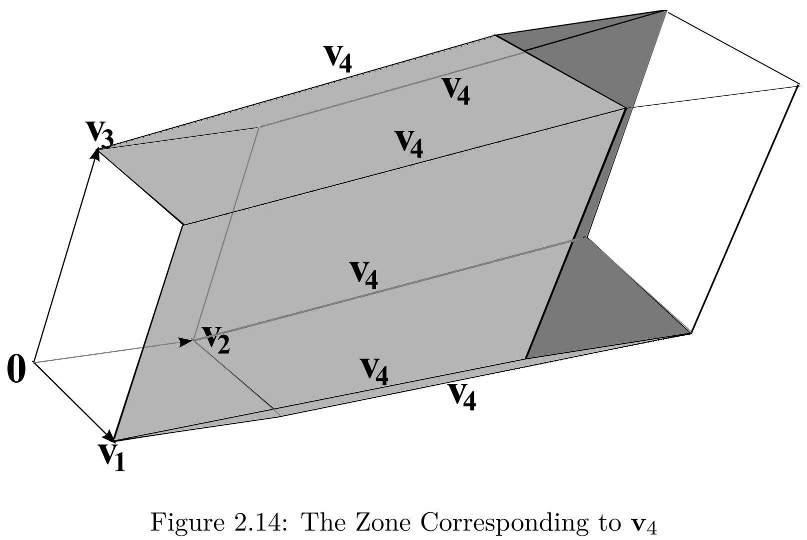

Figure 2.14: The Zone Corresponding to $\mathbf{v}_4$

a new parallelogram face at each step. The process must eventually terminate, if only because a zonohedron has a finite number of edges and faces. By a similar finiteness argument, and the fact that each edge bounds exactly two parallelograms, the zone must eventually end where it started. The result is that a zone is a band of parallelograms that encircles the zonohedron.

A priori, it seems possible that a zone could consist of more than one band. The zonohedron's convexity, however, implies that only one band can occur, as will now be shown. Recall that the order of the summands in a Minkowski sum is immaterial. We can therefore reorder the m generating vectors if necessary so that the generating vector that corresponds to a zone of interest is the last, or m^{th}, summand. Denote the zonohedron of the first $m-1$ generators by $\mathscr{Z}_{m-1}$. Then the total zonohedron $\mathscr{Z}$ is given by

$$\mathscr{Z} = \mathscr{Z}_{m-1} \oplus \mathbf{v}_m. \tag{2.16}$$

Now project $\mathscr{Z}_{m-1}$ along $\mathbf{v}_m$ onto a two-dimensional plane, producing a filled polygon $\mathbb{P}$, with boundary $\mathbb{B}$. (More visually, $\mathbb{P}$ could be seen as the shadow cast by $\mathscr{Z}_{m-1}$, when illuminated by light rays parallel to $\mathbf{v}_m$.) The pre-image of each side of $\mathbb{P}$ is exactly one edge of $\mathscr{Z}_{m-1}$, but not a part of a face—if it were part of a face, that face would contain two edges which were coplanar with $\mathbf{v}_m$; since edges are translated copies of generators, the generators would not be in general position. By a similar argument, the pre-image of each vertex of $\mathbb{P}$ is exactly one vertex of $\mathscr{Z}_{m-1}$. Since projection is continuous, the pre-image of $\mathbb{B}$ is an unbroken sequence of edges of $\mathscr{Z}_{m-1}$. Any two consecutive edges in the sequence meet at a common vertex of $\mathscr{Z}_{m-1}$. Topologically, the sequence of edges forms a circle.

To add $\mathbf{v}_m$ to $\mathscr{Z}_{m-1}$, in the Minkowski sense, split $\mathscr{Z}_{m-1}$ apart along this sequence of edges, as was done in Figure 2.11, and insert copies of $\mathbf{v}_m$ at each vertex (which will become two separated vertices after splitting) in the sequence. The inserted copies form the m^{th} zone. Since the sequence of edges is a topological circle, the inserted copies form a single topological cylindrical band. All inserted copies belong to this cylinder, showing that the m^{th} zone, and similarly any other zone, contains exactly one band.

Another surprising result is that any two zones intersect; in fact they contain two faces in common, on opposite sides of the zonohedron. Figure 2.15, for instance, shows the band for $\mathbf{v}_1$. A comparison with Figure 2.14 shows that the 1^{st} and the 4^{th} zones intersect at the two faces whose edges are copies of $\mathbf{v}_1$ and $\mathbf{v}_4$. An equivalent statement of

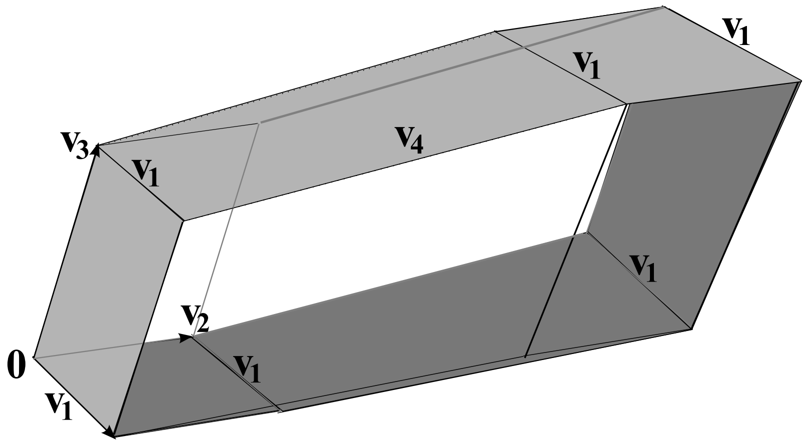

Figure 2.15: The Zone Corresponding to $\mathbf{v}_1$

this result is that, for any pair of generators $\mathbf{v}_i$ and $\mathbf{v}_j$, a zonohedron contains two parallelogram faces whose edges are translations of $\mathbf{v}_i$ and $\mathbf{v}_j$. Again, an argument from construction makes this result clear. Begin the construction of the zonohedron with $\mathbf{v}_i$ and $\mathbf{v}_j$, so the first stage will be a parallelogram with edges $\mathbf{v}_i$ and $\mathbf{v}_j$. The Minkowski sum with the third generator will be a parallelepiped, two opposite faces of which are ij-parallelograms. Further Minkowski sums will add more zones to the zonohedron, and will maintain those two faces, though moving them steadily farther apart. The faces will remain after the zonohedron is complete, and the i^{th} and j^{th} zones will each contain both of them.

2.7 Zonohedra as Convex Polytopes

Both geometric and algebraic arguments show that a zonohedron is a convex polytope. Geometrically, the method of construction shows that a zonohedron's boundary consists of a finite number of line segments and parallelogram faces, both of which consist of convex combinations of their vertices. Since there are finitely many edges and faces, there are also finitely many vertices, which generate all the edges and faces, not to mention the solid zonohedron itself.

Algebraically, Equation (2.11) can be rewritten in terms of the set of nodes:

$$\mathcal{Z}(\mathcal{G}) = \text{hull}(\mathcal{N}(\mathcal{Z})). \tag{2.17}$$

To see this result, express the zonotope in the form given in Equation (2.11). Any point $\mathbf{p}$ in the zonotope can be then be expressed by a sequence $\alpha_1, \alpha_2, ..., \alpha_m$ of coefficients in that equation, where each α_i is between 0 and 1. Formally,

$$\begin{aligned} \mathbf{p} &= \sum_{i=1}^{m} \alpha_i \mathbf{v}_i && (2.18) \\ &= \alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + ... + \alpha_m \mathbf{v}_m. && (2.19) \end{aligned}$$

Note that this expression might not be unique—often many such sequences define the same point.

Without loss of generality, the vectors $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_m$ can be reordered if necessary so that

$$\alpha_1 \leq \alpha_2 \leq \alpha_3 \leq ... \leq \alpha_m. \tag{2.20}$$

To show that $\mathcal{Z}$ is the convex hull of the nodes of $\mathcal{Z}$, it is sufficient to express $\mathbf{p}$ as a convex combination of a set of nodes. In other words, write $\mathbf{p}$ as a linear combination of nodes, whose coefficients are all between 0 and 1, and all sum to 1. To write $\mathbf{p}$ in this form, rearrange Equation (2.19):

$$\begin{aligned} \mathbf{p} = \alpha_1 \sum_{i=1}^{m} \mathbf{v}_i + (\alpha_2 - \alpha_1) \sum_{i=2}^{m} \mathbf{v}_i + ... \\ ... + (\alpha_3 - \alpha_2) \sum_{i=3}^{m} \mathbf{v}_i + ... + (\alpha_m - \alpha_{m-1}) \sum_{i=m}^{m} \mathbf{v}_i && (2.21) \\ = \sum_{j=1}^{m} \left((\alpha_j - \alpha_{j-1}) \sum_{i=j}^{m} \mathbf{v}_i \right), && (2.22) \end{aligned}$$

where α_0 is set to 0. Each summation inside the parentheses in Equation (2.22) is a sum of generating vectors, so, by Equation (2.15), they are all nodes. Furthermore, the inequalities in Equation (2.20) insure that each coefficient inside the parentheses is between 0 and 1. The sum of all the coefficients inside the parentheses is a telescoping series

$$\alpha_1 + \alpha_2 - \alpha_1 + \alpha_3 - \alpha_2 + ... + \alpha_{m-1} - \alpha_{m-2} + \alpha_m - \alpha_{m-1}, \tag{2.23}$$

which sums to α_m.

Equation (2.22) nearly achieves our goal of writing $\mathbf{p}$ as a sum of nodes, with coefficients between 0 and 1, that sum to 1. To take the final step, use the fact that the zero vector $\mathbf{0}$ is always in the set of nodes, and add a multiple of $\mathbf{0}$ to Equation (2.22):

$$\mathbf{p} = \sum_{j=1}^{m} \left((\alpha_j - \alpha_{j-1}) \sum_{i=j}^{m} \mathbf{v}_i \right) + (1 - \alpha_m)\mathbf{0}. \tag{2.24}$$

This addition does not change the vector produced by the linear combination of nodes, but does guarantee that its coefficients sum to 1, as needed. We have therefore succeeded in writing a point $\mathbf{p}$ as a convex combination of nodes.

The nodes in Equation (2.22) are all sums of sequences of vectors in $\mathcal{G}$, where the sequences have different lengths. What constitutes a "sequence," however, depends on how the vectors in $\mathcal{G}$ are ordered, and the ordering used for Equation (2.22) was chosen to insure that Equation (2.20) held. The α_i's in Equation (2.18) could be chosen to induce any desired ordering, so the entire set of orderings, and thus the entire set of nodes, is needed.

The constructions show that every vertex of a zonohedron is the sum of a set of generating vectors, so every vertex is a node. Even simple examples, however, like the zonogon in Figure 2.6 show that not every node is a vertex—in fact, many nodes occur in the interior.

The fact that a zonohedron, or in fact a general zonotope, is the convex hull of its finite set of nodes establishes that a zonotope is a convex polytope. This result will allow us to reach conclusions about possible coordinate representations of points in zonotopes, and, in later chapters, those representations will imply that some colours can only be produced in certain ways.

2.8 Coefficient Sequences for Zonohedra

Equation (2.11) expresses a point $\mathbf{p}$ in a zonotope by a sequence

$$(\alpha_1, \alpha_2, ..., \alpha_m), \tag{2.25}$$

where the α's are the coefficients in a linear combination of generating vectors that produces $\mathbf{p}$. As simple examples like Figure 2.6 show, that sequence might not be unique. In fact, when there are more generators than dimensions, there is often a multitude of such sequences. A sequence could be used as a set of coordinates for a point in a zonotope. This section shows that, under the assumption of general position, a point on the boundary of a zonohedron has a unique coordinate sequence. The result is first shown for the zonohedron's vertices, each of which can be written, in just one way, as the sum of all the generating vectors which lie on one side of a hyperplane through the origin. That result is then used to derive similar uniqueness results for points on edges or faces.

2.8.1 Vertices

Section 2.7 showed that a zonotope is a convex polytope and Section 1.3.4 explained that the vertices of a convex polytope are exposed points: for each vertex $\mathbf{v}$ there exists a supporting hyperplane $\mathcal{H}_{\mathbf{v}}$ such that $\mathcal{H}_{\mathbf{v}}$ intersects the polytope only at $\mathbf{v}$; apart from that intersection, the polytope lies completely on one side of $\mathcal{H}_{\mathbf{v}}$. This section will prove from that description that each vertex of a zonotope is represented by a unique coordinate sequence. The coordinate sequence contains only 0s and 1s, so each vertex is the sum of a unique subset of the generating vectors, consisting of those that are on one side of $\mathcal{H}_{\mathbf{v}}$, after translating $\mathcal{H}_{\mathbf{v}}$ to the origin.

The proof interprets a vertex as the solution to an optimization problem that is studied in linear programming: find a point $\mathbf{p}_{\max}$ on a convex polytope at which a given linear functional takes on a maximum value. Section 1.2 characterized a linear functional as a stack of parallel hyperplanes, each of which is the pre-image under that functional of a certain real number. The functional values corresponding to a hyperplane increase as the hyperplanes get further from the origin. Geometrically, then, $\mathbf{p}_{\max}$ must lie on a supporting hyperplane; all the hyperplanes on one side are of smaller value, and none of the hyperplanes on the other side intersect the polytope.

In the case at hand, the polytope is a zonotope $\mathcal{Z}$. Let us specify a vertex $\mathbf{v}$, and find its coordinate sequence. First, choose a supporting hyperplane $\mathcal{H}_{\mathbf{v}}$. The hyperplane defines, up to a scalar factor, a functional $F_{\mathbf{v}}$; choose the scalar factor so that $F_{\mathbf{v}}(\mathcal{H}_{\mathbf{v}})$ is positive. By the argument in the previous paragraph, $\mathbf{v}$ is the unique solution to the problem of maximizing $F_{\mathbf{v}}$ over $\mathcal{Z}$.

Every point in $\mathcal{Z}$ has the form given by Equation (2.11), so we can write

$$\mathbf{v} = \sum_{i=1}^{m} \alpha_i \mathbf{v}_i, \tag{2.26}$$

where the $\mathbf{v}_i$'s are the generating vectors for $\mathcal{Z}$. To prove the uniqueness of Equation (2.26), we will show that the maximizing interpretation implies that each coefficient α_i must be 0 or 1, depending on which side of $\mathcal{H}_\mathbf{v}$ (when translated to the origin) the i^{th} generator is on. Begin with α_1, and assume, by way of contradiction, that α_1 is neither 0 nor 1. Calculate $F_\mathbf{v}(\mathbf{v}_1)$. We will look at all three possible cases: $F_\mathbf{v}(\mathbf{v}_1)$ is positive, negative, or zero.

If $F_\mathbf{v}(\mathbf{v}_1)$ is positive, then construct the vector

$$\begin{aligned} \mathbf{w} &= \mathbf{v} + (1 - \alpha_1)\mathbf{v}_1 && (2.27)\\ &= \mathbf{v}_1 + \sum_{i=2}^{m} \alpha_i \mathbf{v}_i. && (2.28) \end{aligned}$$

$\mathbf{w}$ is the same as $\mathbf{v}$, except that the first coefficient has been increased to 1. Both vectors are in $\mathcal{Z}$, but

$$\begin{aligned} F_\mathbf{v}(\mathbf{w}) &= F_\mathbf{v}(\mathbf{v}) + (1 - \alpha_1)F_\mathbf{v}(\mathbf{v}_1) && (2.29)\\ &> F_\mathbf{v}(\mathbf{v}). && (2.30) \end{aligned}$$

Thus $F_\mathbf{v}$ takes on a higher value at $\mathbf{w}$ than at $\mathbf{v}$, contradicting the maximality of $\mathbf{v}$. This contradiction can only be avoided if α_1 is actually 1, which eliminates the second term in (2.29). Less formally, when $F_\mathbf{v}(\mathbf{v}_1)$ is positive, then linearity implies that using $\mathbf{v}_1$ as much as possible only increases $F_\mathbf{v}$, so the coefficient of $\mathbf{v}_1$ should be set to its maximum, which is 1.

If $F_\mathbf{v}(\mathbf{v}_1)$ is negative, then construct the vector

$$\begin{aligned} \mathbf{w} &= \mathbf{v} - \alpha_1 \mathbf{v}_1 && (2.31)\\ &= \sum_{i=2}^{m} \alpha_i \mathbf{v}_i. && (2.32) \end{aligned}$$

$\mathbf{w}$ is the same as $\mathbf{v}$, except that the first coefficient has been decreased to 0. Both vectors are in $\mathcal{Z}$, but

$$\begin{aligned} F_\mathbf{v}(\mathbf{w}) &= F_\mathbf{v}(\mathbf{v}) - \alpha_1 F_\mathbf{v}(\mathbf{v}_1) && (2.33)\\ &> F_\mathbf{v}(\mathbf{v}). && (2.34) \end{aligned}$$

Again, $F_{\mathbf{v}}$ takes on a higher value at $\mathbf{w}$ than at $\mathbf{v}$, contradicting the maximality of $\mathbf{v}$. The contradiction is avoided if α_1 is 0. In this case, using more of $\mathbf{v}_1$ only decreases $F_{\mathbf{v}}$, so the coefficient of $\mathbf{v}_1$ should be set to its minimum, which is 0.

The positive and negative cases require that the maximizing coefficient α_1 must be either 0 or 1; furthermore, the choice of 0 or 1 is unique. It will be shown that the third case, $F_{\mathbf{v}}(\mathbf{v}_1) = 0$, cannot occur, because it violates the assumption that $\mathcal{H}_{\mathbf{v}}$ intersects $\mathcal{Z}$ only at $\mathbf{v}$. The reason is that $F_{\mathbf{v}}(\mathbf{v}_1) = 0$ implies that

$$F_{\mathbf{v}}(\mathbf{v} + \alpha_1 \mathbf{v}_1) = F_{\mathbf{v}}(\mathbf{v}), \tag{2.35}$$

even when α_1 is positive (but no greater than 1). $\mathbf{v} + \alpha_1 \mathbf{v}_1$ is contained in the zonotope, which implies that $\mathbf{v} + \alpha_1 \mathbf{v}_1$ is also on the supporting hyperplane $\mathcal{H}_{\mathbf{v}}$, whereas we had chosen $\mathcal{H}_{\mathbf{v}}$ to intersect the zonotope *only* at $\mathbf{v}$. Therefore $F_{\mathbf{v}}(\mathbf{v}_1)$ is not zero, implying that α_1 is uniquely either 0 or 1.

Of course, the same argument could be repeated with α_2, α_3, and so on, implying that each coefficient α_1 is uniquely either 0 or 1. Generators with coefficient 0 can be disregarded, and the vertex $\mathbf{v}$ written uniquely as the sum of the subset $\mathcal{G}'$ of generating vectors whose coefficients are 1, as was to be shown.

This subset can be interpreted geometrically. Since $F_{\mathbf{v}}$ is positive on each vector in $\mathcal{G}'$, and negative on any vector outside $\mathcal{G}'$, $\mathcal{G}'$ can be seen as the set of generators that are on one side of the hyperplane given by $\ker F_{\mathbf{v}}$. This kernel, however, is just a translation of the hyperplane $\mathcal{H}_{\mathbf{v}}$. Given a vertex $\mathbf{v}$, then, and a hyperplane which intersects the zonotope only at that vertex, just shift that hyperplane to the origin. The vertex $\mathbf{v}$ is then the sum of all the generating vectors which are on the same side of the shifted hyperplane as $\mathbf{v}$.

A two-dimensional example will make the geometric interpretation clear. Figure 2.16 shows a zonogon with five generators. A vertex $\mathbf{v}$ has been chosen. In two dimensions, a hyperplane is just a line. A supporting hyperplane has been drawn through $\mathbf{v}$, that intersects the zonogon only at $\mathbf{v}$. A translated copy of that hyperplane has been drawn through the origin. The three generators $\mathbf{v}_1$, $\mathbf{v}_3$, and $\mathbf{v}_4$ are on the same side of the translated hyperplane as $\mathbf{v}$. It can be seen geometrically that $\mathbf{v}$ is the sum of these three vectors:

$$\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_3 + \mathbf{v}_4, \tag{2.36}$$

because a path from the origin to $\mathbf{v}$ can be traced along the boundary, and copies of those three vectors appear as boundary segments along

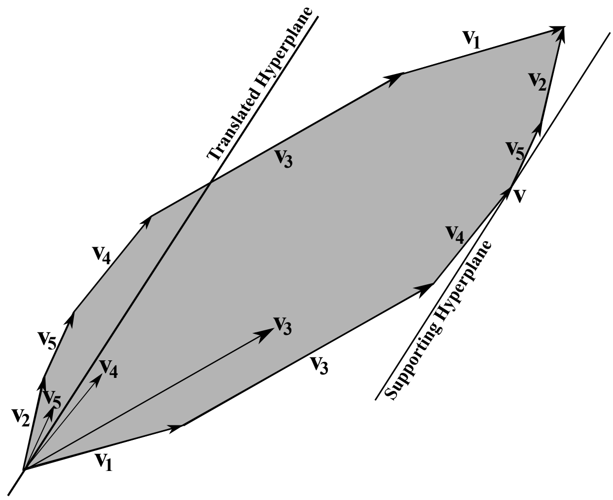

Figure 2.16: A Vertex as the Sum of Generators on One Side of a Hyperplane

that path. It can be seen from the figure that the supporting hyperplane is not unique. Any hyperplane that intersects the zonogon only at $\mathbf{v}$ will, when translated to the origin, lie between $\mathbf{v}_4$ and $\mathbf{v}_5$, and produce the same set of generators.

A corollary to this construction is that a zonotope is centrally symmetric: it has a unique center, and any point $\mathbf{p}$ on the zonotope has an antipodal point $\mathbf{p}_a$, which is diametrically opposite $\mathbf{p}$, across the center. Algebraically, the center of the zonotope is given by

$$\mathbf{v}_c = \sum_{1=1}^{m} \frac{1}{2}\mathbf{v}_i, \tag{2.37}$$

the point whose coefficients are all $1/2$. Geometrically, the center is halfway between the origin and the terminal point.

Suppose we have the vertex $\mathbf{v}$, shown in Figure 2.17. A line from $\mathbf{v}$ through $\mathbf{v}_c$ will intersect the boundary at an antipodal vertex $\mathbf{v}_a$, that is diametrically opposite to $\mathbf{v}$. It is not hard to see that $\mathbf{v}_a$ is the sum of the generating vectors $\mathbf{v}_2$ and $\mathbf{v}_5$, which make up the complement of the generating vectors that sum to $\mathbf{v}$. A simple algebraic way to

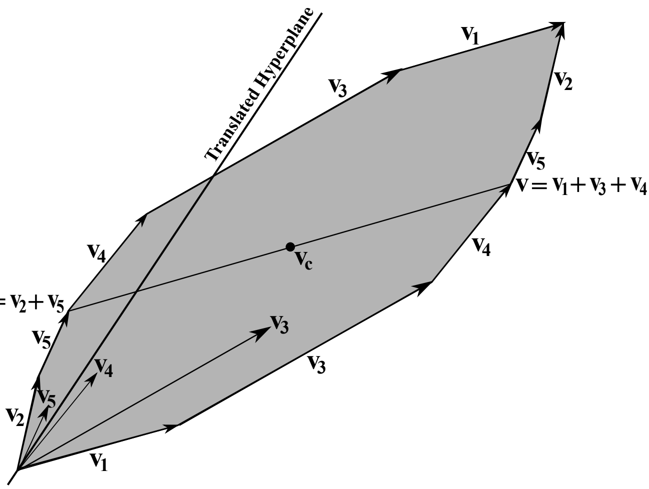

Figure 2.17: Centrally Symmetric Vertices on a Zonogon

write this relationship is

$$\mathbf{v}_a = \sum_{1=1}^{m} (1 - \alpha_i)\mathbf{v}_i. \tag{2.38}$$

Equation (2.38) actually gives the diametrically opposite point for any point $\mathbf{v}$ in the zonogon, whether or not it is a vertex.

While this example has been in two dimensions, generalizations to three or more dimensions are straightforward. Not only the proof of unique coefficient sequences for vertices, but also the algebraic development of central symmetry, works in any dimension. In the three-dimensional case of interest in this book, a zonotope is a zonohedron, and hyperplanes are just planes in $\mathbb{R}^3$. To find a vertex, then, draw a plane through the origin. All the generating vectors that are on one side of the plane will sum to one vertex, and the vectors on the other side will sum to the antipodal vertex. Furthermore, the corresponding coordinate expressions for the vertices are unique. In later chapters, some sets of colours will be shown to have a zonohedral form; this section's results will then imply that a colour represented by a vertex can be produced in only one way from the colours used as generating vectors.

2.8.2 Edges and Faces

The previous section showed that each vertex of a zonotope has a unique expression as a linear combination of the generating vectors. This section proves similar results for edges and faces, except that uniqueness is no longer automatic; rather, it now requires the generating vectors to satisfy some linear independence relationships. Fortunately, the requirements are mild: as long as no two generators are scalar multiples of each other, each point on an edge has a unique coefficient sequence, and as long as no three generating vectors are linearly dependent, each point on a face has a unique coefficient sequence. For zonohedra, both these conditions are satisfied when the generators are in general position, which is the generic case. Geometrically, general position implies that all a zonohedron's faces are parallelograms, and that each edge is a translated copy of a single generator (rather than multiple generators laid end to end).

The result about edges will be proven first. Assume, for a given zonohedron $\mathcal{Z}(\mathcal{G})$, that no two vectors in $\mathcal{G}$ are linearly dependent, i.e. that no two vectors are scalar multiples of each other.. Let $\mathcal{E}$ be an edge of $\mathcal{Z}$, and let $\mathbf{v}_{\mathcal{E}}$ be the generator of which $\mathcal{E}$ is a copy. The assumption that no two generators are linearly dependent implies that $\mathbf{v}_{\mathcal{E}}$ is a unique, single vector. (The only other option is that $\mathcal{E}$ consists of multiple generators laid end to end, but that configuration requires two or more generators to be linearly dependent.)

As an edge, $\mathcal{E}$ is bounded by two vertices, $\mathbf{w}_1$ and $\mathbf{w}_2$. Reorder the vertices if necessary so that

$$\mathbf{w}_1 + \mathbf{v}_{\mathcal{E}} = \mathbf{w}_2. \tag{2.39}$$

To show that any point $\mathbf{w}_{\mathcal{E}}$ on $\mathcal{E}$ has a unique coordinate sequence, begin by choosing a supporting hyperplane $\mathcal{H}_{\mathcal{E}}$ that contains the edge $\mathcal{E}$, but that does not otherwise intersect $\mathcal{Z}$; in particular, $\mathcal{H}_{\mathcal{E}}$ contains no face of $\mathcal{Z}$. The existence of such an $\mathcal{H}_{\mathcal{E}}$ follows from the fact that $\mathcal{Z}$ is a convex polytope.

Translate $\mathcal{H}_{\mathcal{E}}$ to the origin, and consider its relationship to the generating vectors. First, since $\mathcal{E}$ is a translated copy of $\mathbf{v}_E$, it follows that $\mathbf{v}_{\mathcal{E}}$ is contained in the translated $\mathcal{H}_{\mathcal{E}}$. Second, since $\mathbf{w}_1$ and $\mathbf{w}_2$ are bounding vertices, it follows from previous discussions that $\mathcal{H}_{\mathcal{E}}$ corresponds to a linear functional $f_{\mathcal{E}}$ that is maximized by $\mathbf{w}_1$ and $\mathbf{w}_2$. The translated $\mathcal{H}_{\mathcal{E}}$ divides $\mathcal{G}$ into three subsets: $\mathcal{G}_+$ (the generators on which $F_{\mathcal{E}}$ is positive), $\mathcal{G}_-$ (the generators on which $F_{\mathcal{E}}$ is negative), and $\mathbf{v}_{\mathcal{E}}$ (the single generator on which $F_{\mathcal{E}}$ is zero).

We can write $\mathbf{w}_1$ and $\mathbf{w}_2$ in terms of these subsets:

$$\mathbf{w}_1 = \sum_{\mathbf{v}_i \in \mathcal{G}_+} \mathbf{v}_i, \tag{2.40}$$

$$\mathbf{w}_2 = \sum_{\mathbf{v}_i \in \mathcal{G}_+ \cup \mathbf{v}_\mathcal{E}} \mathbf{v}_i. \tag{2.41}$$

Since $\mathbf{w}_1$ and $\mathbf{w}_2$ are vertices, these expressions are unique. Every point $\mathbf{w}_\mathcal{E}$ on $\mathcal{E}$ can be written as

$$\mathbf{w}_\mathcal{E} = \alpha \mathbf{w}_1 + (1-\alpha)\mathbf{w}_2, \tag{2.42}$$

for some unique α between 0 and 1. Then, since $\mathbf{w}_1 + \mathbf{v}_\mathcal{E} = \mathbf{w}_2$, we can write

$$\mathbf{w}_\mathcal{E} = \mathbf{w}_1 + \alpha \mathbf{v}_\mathcal{E} \tag{2.43}$$

$$= \sum_{\mathbf{v}_i \in \mathcal{G}_+} \mathbf{v}_i + \alpha \mathbf{v}_\mathcal{E}. \tag{2.44}$$

Equation (2.44) provides one coefficient sequence for $\mathbf{w}_\mathcal{E}$; uniqueness follows from the maximizing behavior of $F_\mathcal{E}$. Since $\mathbf{w}_1$ and $\mathbf{w}_2$ both maximize the linear functional $F_\mathcal{E}$, therefore $\mathbf{w}_\mathcal{E}$, which is a convex combination of $\mathbf{w}_1$ and $\mathbf{w}_2$, also maximizes $F_\mathcal{E}$. From Equation (2.44) we can write

$$F_\mathcal{E}(\mathbf{w}_\mathcal{E}) = F_\mathcal{E}\left(\sum_{\mathbf{v}_i \in \mathcal{G}_+} \mathbf{v}_i\right) + F_\mathcal{E}(\alpha \mathbf{v}_\mathcal{E}). \tag{2.45}$$

Eliminating any vector in $\mathcal{G}_+$, or reducing its coefficient to less than 1, in the first argument on the right of Equation (2.45), would only decrease the value of $F_\mathcal{E}$. Similarly, introducing any vector in $\mathcal{G}_-$, or increasing its coefficient beyond 0, would only cause a decrease. Therefore, all coefficients for $\mathcal{G}_+$ must remain at 1, and all coefficients for $\mathcal{G}_-$ must remain at 0. Increasing or decreasing α would not change the value of $F_\mathcal{E}$, but would invalidate Equation (2.43), so α must also stay the same. The expression in Equation (2.44) is therefore the only one possible, so every point on the edge of a zonotope, under the assumption that no two generators are linearly dependent, must have a unique coefficient sequence, as was to be shown.

Uniqueness will now be proven similarly for points on a zonotope's face, under the assumption that no three generating vectors are linearly dependent. The edge $\mathcal{E}$ is replaced with a face $\mathcal{F}$, and the hyperplane $\mathcal{H}_\mathcal{F}$ makes full contact with that face, but with no other part of the zonotope. Again, translate $\mathcal{H}_{\mathcal{F}}$ to the origin. Define $F_{\mathcal{F}}$, $\mathcal{G}_+$, and $\mathcal{G}_-$ similarly to the edge case. Instead of just containing one generator $\mathbf{v}_{\mathcal{E}}$, the translated hyperplane now contains two generators, $\mathbf{v}_{\mathcal{F}1}$ and $\mathbf{v}_{\mathcal{F}2}$. Any other generator that might be contained in the translated hyperplane would be a linear combination of $\mathbf{v}_{\mathcal{F}1}$ and $\mathbf{v}_{\mathcal{F}2}$, which would violate the assumption; therefore $\mathbf{v}_{\mathcal{F}1}$ and $\mathbf{v}_{\mathcal{F}2}$ are the only generators in the translated hyperplane.

Since the face $\mathcal{F}$ consists, after translation, of zonal combinations of two vectors, it is a parallelogram with four vertices. One of the vertices, which we'll call $\mathbf{w}_1$, is an "initial" vertex, and the opposite vertex, which we'll call $\mathbf{w}_2$, is a "terminal" vertex. These two vertices satisfy the relationship

$$\mathbf{w}_2 = \mathbf{w}_1 + \mathbf{v}_{\mathcal{F}1} + \mathbf{v}_{\mathcal{F}2}. \tag{2.46}$$

The remaining two vertices of the parallelogram are $\mathbf{w}_1 + \mathbf{v}_{\mathcal{F}1}$ and $\mathbf{w}_1 + \mathbf{v}_{\mathcal{F}2}$. The linear independence of $\mathbf{v}_{\mathcal{F}1}$ and $\mathbf{v}_{\mathcal{F}2}$ imply that, for any point $\mathbf{w}_{\mathcal{F}}$ on the face, there exist two unique constants, α_1 and α_2, both between 0 and 1, such that

$$\mathbf{w}_{\mathcal{F}} = \mathbf{w}_1 + \alpha_1 \mathbf{v}_{\mathcal{F}1} + \alpha_2 \mathbf{v}_{\mathcal{F}2}. \tag{2.47}$$

As before, the vertex $\mathbf{w}_1$ is a unique linear combination of generators, whose indices are taken from some subset $\mathcal{G}_+$, so we have

$$\mathbf{w}_{\mathcal{F}} = \sum_{\mathbf{v}_i \in \mathcal{G}_+} \mathbf{v}_i + \alpha_1 \mathbf{v}_{\mathcal{F}1} + \alpha_2 \mathbf{v}_{\mathcal{F}2}. \tag{2.48}$$

Again, apply $F_{\mathcal{F}}$ to both sides of Equation (2.48). Argue as before that the coefficients for generators in $\mathcal{G}_+$ and $\mathcal{G}_-$ cannot be changed. The two remaining coefficients, α_1 and α_2, are unique because $\mathbf{v}_{\mathcal{F}1}$ and $\mathbf{v}_{\mathcal{F}2}$ are linearly independent; this reasoning is a two-dimensional version of the reasoning after Equation (2.45), where there was only coefficient α, instead of two. Since all the coefficients in Equation (2.48) are unique, any point on the face of a zonotope, or in particular a zonohedron, provided that no three generators are linearly dependent, has a unique coefficient sequence, as was to be shown.

A simple corollary of these two proofs can also be seen. The unique coefficient sequence for a point on the edge of a zonotope consists entirely of zeros and ones, except for at most one generator. Similarly, the unique coefficient sequence for a point on the face of a zonotope consists entirely of zeros and ones, except for at most two generators. Both these corollaries, of course, make an appropriate general position assumption.

2.9 Cyclic Zonohedra

Cyclic zonohedra are non-negative zonohedra whose generators form a minimal set for their convex cone. They occur several times in colour science, and the cyclic property is vital to some results. This section describes cyclic zonohedra, presents a convenient construction for them, and shows that they possess some regularities that make them easy to work with.

Formally, suppose we have a set $\mathcal{G}$ of m non-negative generating vectors. They can be pictured as line segments emerging from the origin, in the non-negative octant of $\mathbb{R}^3$. Each generating vector can be extended to a semi-infinite ray, and the convex hull of those rays is the convex cone, $\text{cone}(\mathcal{G})$, of the generating set. Although the term "cone" suggests a regular, usually circular, profile, a cone in this context could have a very irregular profile; since there are only finitely many generating rays, the profile must be polygonal.

Regardless of its profile, any finitely-generated convex cone still has an inside, a surface, and an outside. Typically, one pictures a cone's generating rays as lying on its surface. Possibly, however, one or more of the rays is inside the cone. In that case, any rays that are inside are not needed to generate the cone. Geometrically, the inside rays are convex combinations of the rays on the surface and so are unnecessary; the generators as a whole, then, do not form a minimal set for the cone.

While this discussion has been in three dimensions, Figure 2.18 shows that the issues can be expressed more clearly in two dimensions. The left side of the figure shows five non-negative generators, their extension to rays, and the convex cone that results. A simple way to view the profile of the cone, and to determine whether any rays are inside, is to slice it with the plane $\mathcal{H}$ given by $x + y + z = 1$. The result appears on the right of the figure. Each generator produces one point $\mathbf{p}_i$ on the plane $x + y + z = 1$. Denote the set of points $\mathbf{p}_i$ by $\mathcal{P}$.

In that plane, the convex hull, $\text{hull}(\mathcal{P})$, of the $\mathbf{p}_i$'s is a polygon. It is easy to see that a ray created from a generating vector is inside the convex cone if and only the corresponding point is inside the polygon; in the figure, for instance, the ray through $\mathbf{v}_3$ can be written as a convex combination of the rays through the other four generators, and $\mathbf{p}_3$ can be written as a convex combination of the other four points. (Another interpretation is that a ray is inside the convex cone if and only if the corresponding generator is a non-negative linear combination of the other generators.) It is also easy to see that the choice of sec-

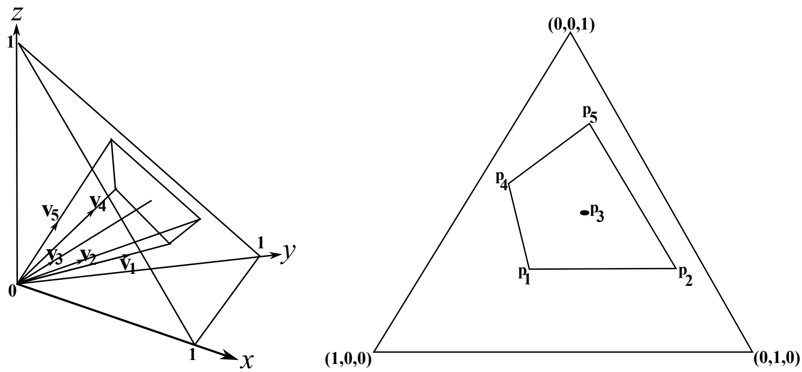

Figure 2.18: Two-Dimensional Section of a Convex Cone

tional plane $x + y + z = 1$ is not very important: any plane that cuts the x-, y-, and z-axes on their positive halves will capture the needed convexity information. For simplicity, then, we will largely work in a two-dimensional planar section.

The convex hull of $\mathcal{P}$ might also be the convex hull of a subset of $\mathcal{P}$. The smallest such subset is called the *minimal set*, and can be shown to be unique. If the minimal set consists exactly of the points in $\mathcal{P}$, then the set of generating vectors, as well as the zonohedron they generate, is said to be *cyclic*. To motivate this terminology, Figure 2.19 shows two possible configurations for $\mathcal{P}$, each with six points. In the example on the left, the points fall approximately on a hexagon, and are a minimal set for their convex hull: if any point were removed, then the resulting convex hull would be smaller. The lefthand example is therefore cyclic, and the points can be naturally ordered in a cicle, either clockwise or counterclockwise. In the example on the right, one of the points falls inside the convex hull. Removing the inside point would not affect the convex hull. Algebraically, the point on the inside can be written as a convex combination of the other points. The set of points on the right is not cyclic—no circular ordering is possible, because it is unclear where the inside point would fit in the sequence.

An "inside" point does not have to be completely inside. Suppose instead that it was located on an edge of a polygon in Figure 2.19, between two vertices. Then the point would be a convex combination of those vertices, so it would not be needed in the minimal set. By our definition, such a set would not be called cyclic, even though there

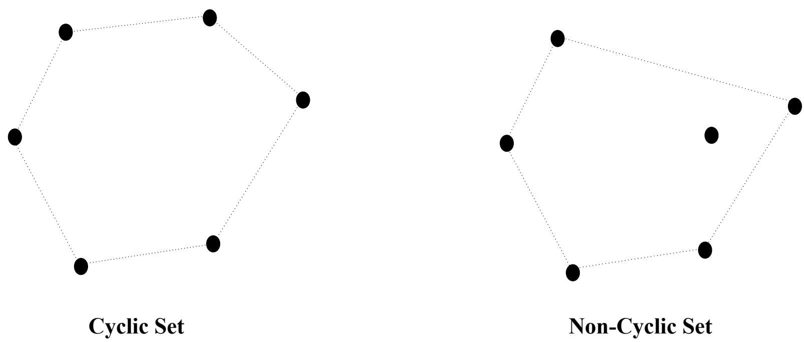

Figure 2.19: Examples of Cyclic and Non-Cyclic Sets

might be a clear circular ordering of the points. Another anomalous case would occur when two generators were scalar multiples of each other, so they produced the same point in $\mathcal{H}$. Then either of the points would trivially be a convex combination of the other point, so one of them is not necessary; this case would also be considered not cyclic. In light of these cases, it can be seen that a cyclic set of generators is necessarily in general position: three linearly dependent vectors would be coplanar, and their corresponding points in $\mathcal{H}$ would be collinear, so the middle point would be not be needed. Similarly, if two vectors were linearly dependent, then they would produce two copies of the same point in $\mathcal{H}$, only one of which would be needed for the minimal set.

A cyclic zonohedron results from a cyclic set of generating vectors, and has some distinctive structure. For instance, the fact that a cyclic set of generators is in general position implies that each face of the zonohedron is a parallelogram, and each edge is a translated copy of exactly one generator.

More surprisingly, on the zonohedron itself, each generating vector appears as an edge that starts at the origin. Figure 2.20 demonstrates this fact by construction. Suppose that a generating vector of interest produces the point $\mathbf{p}_1$ in $\mathcal{H}$. Since the vectors are cyclic, their corresponding points in $\mathcal{H}$ are a minimal set; therefore the convex hull of points $\mathbf{p}_2$ through $\mathbf{p}_6$, that is, the set of all the points except $\mathbf{p}_1$, does not contain $\mathbf{p}_1$. It can be seen that in such a case there exists a line $\mathcal{L}$ in $\mathcal{H}$ that separates $\mathbf{p}_1$ from that convex hull.

$\mathcal{H}$ is a two-dimensional section of $\mathbb{R}^3$, like the configuration shown

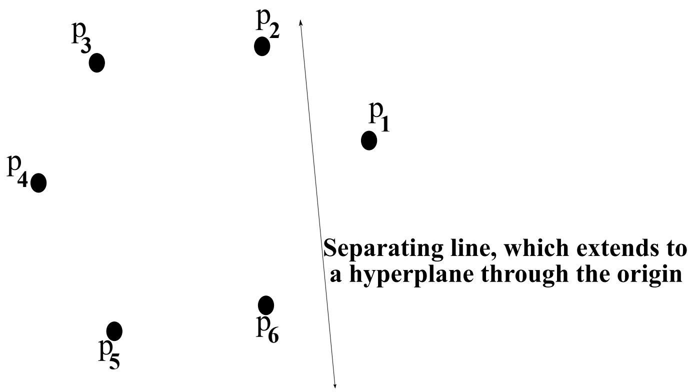

Figure 2.20: The Structure of a Cyclic Zonohedron at the Origin

in Figure 2.18. In that figure, the line $\mathcal{L}$ would appear on the triangular intersection of $x + y + z = 1$ with the positive octant. It is then clear that $\mathcal{L}$ could be extended to a plane $\mathcal{P}_{\mathcal{L}}$ that includes the origin. $\mathcal{P}_{\mathcal{L}}$ is then a hyperplane in $\mathbb{R}^3$ that separates the first generating vector $\mathbf{v}_1$ from the rest of the vectors. The previous results about coefficient sequences for vertices imply that the first $\mathbf{v}_1$ itself appears as a vertex. By convexity, the zonohedron contains the line joining $\mathbf{v}_1$ to the origin, which itself is always a vertex. Since the difference between the vertex $\mathbf{v}_1$ and the vertex at the origin consists of a single vector, the vertices must be adjacent, so the line joining them, which is identical with $\mathbf{v}_1$ seen as a line segment, must be an exterior edge.

By relabeling the generators, the argument above can be seen to apply to any generator, so each generating vector similarly appears as an exterior edge that starts at the origin, as was to be shown. Furthermore, since each edge is a translated copy of a generator, all the edges at the origin must be generators, so we have a complete description of the zonohedron's structure at the origin. Since the zonohedron is centrally symmetric, the structure at the terminal point is identical, but reversed: a copy of each generating vector appears as an edge that ends at the terminal point.

Similar constructions allow us to identify *all* the vertices of a cyclic zonohedron, and write them in a convenient sequential form. Any vertex is the sum of all the generators on one side of a hyperplane in $\mathbb{R}^3$; equivalently, it is the sum of the vectors corresponding to all the points $\mathbf{p}_i$ that appear on one side of a line $\mathcal{L}$ in $\mathcal{H}$. Figure 2.21 shows some examples of these lines. Line 1 already appeared in Figure 2.20, and gives the vertex $\mathbf{v}_1$. Line 1-2 gives the vertex $\mathbf{v}_1 + \mathbf{v}_2$. These two vertices are adjacent, and the edge between them is a translated copy of $\mathbf{v}_2$. Continuing, Line 1-2-3 gives the vertex $\mathbf{v}_1 + \mathbf{v}_2 + \mathbf{v}_3$. This vertex is adjacent to the vertex $\mathbf{v}_1 + \mathbf{v}_2$, and the edge between them is a translated copy of $\mathbf{v}_3$. This construction can continue around the set until all but one of the vectors are included; including the final vector would result in the terminal point.

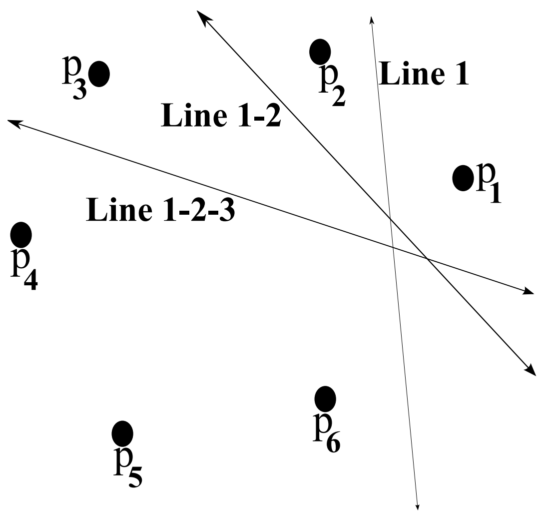

Figure 2.21: The Form of All Vertices of a Cyclic Zonohedron

Each vertex is the sum of a set of generating vectors whose indices form a consecutive sequence. The construction could have proceeded clockwise rather than counterclockwise, in which case $\mathbf{v}_1 + \mathbf{v}_6$, $\mathbf{v}_1 + \mathbf{v}_6 + \mathbf{v}_5$, and so on, would have appeared as vertices. The set 1-6-5 could also be interpreted as a consecutive sequence, if one allows a modular wraparound, so that 1 is the next number after 6. The starting point of the sequence is arbitrary. In general, any consecutive modular sequence of up to m vectors, regardless of its starting vector, sums to a vertex of a cyclic zonohedron, and all vertices have that form.

The set of consecutive sequences, and thus the set of vertices of a cyclic zonohedron, can be partitioned by the number of summands in the sequence. Geometrically and combinatorially, all vertices with the same number of summands can be thought as appearing on one level. The vertices adjacent to the origin are simply the generating vectors; they form the first level of vertices and are "sums" of one generator. The second level of vertices is above and adjacent to the first level, and they consist of sums of two generators. These levels continue until the terminal vertex, which is the sum of all the generators.

The following matrix scheme formalizes these calculations for the zonohedron's vertices. In the first row of the matrix, list the zero vector m times. In the second row of the matrix, list the m generators, numbered as a clockwise or counterclockwise sequence. The matrix's third row consists of sums of two vectors. Entry $(3, j)$, for $j = 1 \ldots m$, is the sum of entry $(2, j)$ and the vector immediately to the right of, i.e. after, the vector in $(2, j)$, in the numbering sequence chosen. The adjacent vectors should be selected modularly: the first vector comes immediately after the mth vector. The fourth row consists of sums of three vectors. Entry $(4, j)$, for $j = 1 \ldots m$, is the sum of entry $(3, j)$, which is already the sum of two vectors, and the vector adjacent to those two vectors, in the chosen order. Continue this process until the $(m + 1)^{\text{st}}$ row, which will contain m copies of the sum of all the vectors. Each row of the matrix contains all the vertices at one level. Expression (2.49) shows the matrix when m is 4:

$$\begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{v_1} & \mathbf{v_2} & \mathbf{v_3} & \mathbf{v_4} \\ \mathbf{v_1} + \mathbf{v_2} & \mathbf{v_2} + \mathbf{v_3} & \mathbf{v_3} + \mathbf{v_4} & \mathbf{v_4} + \mathbf{v_1} \\ \mathbf{v_1} + \mathbf{v_2} + \mathbf{v_3} & \mathbf{v_2} + \mathbf{v_3} + \mathbf{v_4} & \mathbf{v_3} + \mathbf{v_4} + \mathbf{v_1} & \mathbf{v_4} + \mathbf{v_1} + \mathbf{v_2} \\ \sum_{i=1}^{4} \mathbf{v_i} & \sum_{i=1}^{4} \mathbf{v_i} & \sum_{i=1}^{4} \mathbf{v_i} & \sum_{i=1}^{4} \mathbf{v_i} \end{bmatrix} \tag{2.49}$$

Every entry in Expression (2.49) is a vertex of the zonohedron, although with some duplication: the entries in the first row are all the origin, and the entries in the last row are all the terminal vertex. In all, there are $m^2 - m + 2$ vertices. Each vertex is the sum of adjacent vectors in the cyclic sequence. Two adjacent vectors are summed for the third row, three for the fourth row, and so on.

Not only can the zonohedron's vertex structure be inferred from the matrix constructed, but so can its face structure. Two vertices are adjacent, i.e. joined by an edge in the zonohedron, if and only if their difference is a generating vector. Expression (2.50) draws lines to represent edges between the adjacent vertices in Expression (2.49). The first column has been repeated at the right side of the matrix, in order to show all the adjacencies. Some edges appear twice in

Expression (2.50). For example $\mathbf{v_2}$, in entry $(2, 2)$ is joined to $\mathbf{0}$ by two lines. Since all the entries in the first row represent the same point, the origin, the two lines actually represent the same edge in the zonohedron. Similar comments apply to the last row. Expression (2.50) visually displays $m(m-1)$ parallelograms, each of which is a face of the cyclic zonohedron.

$$
\begin{tikzcd}
 & \mathbf{0} \arrow[d, dash] & \mathbf{0} \arrow[d, dash] & \mathbf{0} \arrow[d, dash] & \mathbf{0} \arrow[d, dash] \\
\mathbf{v_1} \arrow[ur, dash] \arrow[d, dash] & \mathbf{v_2} \arrow[ur, dash] \arrow[d, dash] & \mathbf{v_3} \arrow[ur, dash] \arrow[d, dash] & \mathbf{v_4} \arrow[ur, dash] \arrow[d, dash] & \mathbf{v_1} \arrow[d, dash] \\
\mathbf{v_1}+\mathbf{v_2} \arrow[ur, dash] \arrow[d, dash] & \mathbf{v_2}+\mathbf{v_3} \arrow[ur, dash] \arrow[d, dash] & \mathbf{v_3}+\mathbf{v_4} \arrow[ur, dash] \arrow[d, dash] & \mathbf{v_4}+\mathbf{v_1} \arrow[ur, dash] \arrow[d, dash] & \mathbf{v_1}+\mathbf{v_2} \arrow[d, dash] \\
\mathbf{v_1}+\mathbf{v_2}+\mathbf{v_3} \arrow[ur, dash] \arrow[d, dash] & \mathbf{v_2}+\mathbf{v_3}+\mathbf{v_4} \arrow[ur, dash] \arrow[d, dash] & \mathbf{v_3}+\mathbf{v_4}+\mathbf{v_1} \arrow[ur, dash] \arrow[d, dash] & \mathbf{v_4}+\mathbf{v_1}+\mathbf{v_2} \arrow[ur, dash] \arrow[d, dash] & \mathbf{v_1}+\mathbf{v_2}+\mathbf{v_3} \\
\sum_{i=1}^{4} \mathbf{v_i} \arrow[ur, dash] & \sum_{i=1}^{4} \mathbf{v_i} \arrow[ur, dash] & \sum_{i=1}^{4} \mathbf{v_i} \arrow[ur, dash] & \sum_{i=1}^{4} \mathbf{v_i} \arrow[ur, dash] &
\end{tikzcd}
\tag{2.50}
$$

Another interesting feature of a cyclic zonohedron is that each zone contains both the origin and the terminal point. To see this result, assume we want to find the i^{th} zone. Since a copy of $\mathbf{v}_i$ emerges from the origin, and thus contains it, the two parallelogram faces on either side of that copy also contain the origin, and therefore the zone as a whole contains the origin. A similar argument applies to the terminal point, which is shaped like the origin, but after a reflection through the zonohedron's center. Note that a non-cyclic zonohedron, such as the one shown in Figure 2.14, has zones, such as the fourth one, that do not contain the origin.

In the context of colour science, later chapters will show that the set of all surface colours, when viewed under a given illuminant, form a cyclic zonohedron. Different illuminants will lead to different zonohedra, but all such zonohedra are cyclic and we will see that they share a common structure at the origin.

2.10 Chapter Summary

This chapter has presented two important constructions, Minkowski sums and zonohedra (more generally, zonotopes), and derived some useful properties about them. This summary lists the definitions and results.

Definition of the Minkowski Sum. The *Minkowski sum*, denoted $\oplus$, of two non-empty subsets, $\mathcal{A}$ and $\mathcal{B}$, of a vector space $\mathbb{R}^n$, is defined as

$$\mathcal{A} \oplus \mathcal{B} = \left\{ \mathbf{v}_{\mathcal{A}} + \mathbf{v}_{\mathcal{B}} \middle| \mathbf{v}_{\mathcal{A}} \in \mathcal{A} \text{ and } \mathbf{v}_{\mathcal{B}} \in \mathcal{B} \right\}. \tag{2.51}$$

Properties of the Minkowski Sum:

1. The Minkowski sum of an arbitrary number of summands is well-defined, and is independent of the order of those summands.
2. The Minkowski sum of an arbitrary number of convex sets is again a convex set.

Definitions for Zonotopes and Zonohedra:

1. Suppose that we have a set of vectors $\mathcal{G} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_m\}$, in $\mathbb{R}^n$. The *zonotope* $\mathcal{Z}$ generated by $\mathcal{G}$ is the Minkowski sum of the line segments corresponding to those vectors:

   $$\mathcal{Z}(\mathcal{G}) = \mathbf{v}_1 \oplus \mathbf{v}_2 \oplus ... \oplus \mathbf{v}_m. \tag{2.52}$$

   Equivalently, $\mathcal{Z}(\mathcal{G})$ can be written as

   $$\mathcal{Z}(\mathcal{G}) = \left\{ \sum_{i=1}^{m} \alpha_i \mathbf{v}_i \middle| 0 \le \alpha_i \le 1 \; \forall i \right\}. \tag{2.53}$$

   If $n = 2$, then $\mathcal{Z}$ is a *zonogon.* If $n = 3$, then $\mathcal{Z}$ is a *zonohedron.*
2. A zonotope is *non-negative* if there exists a basis of $\mathbb{R}^n$ in which all the co-ordinates, of every vector in $\mathcal{G}$, are non-negative. A non-negative zonotope has a *terminal point*, given by

   $$\mathbf{v}_1 + \mathbf{v}_2 + ... + \mathbf{v}_m, \tag{2.54}$$

   and a *center*, given by

   $$\sum_{1=1}^{m} \frac{1}{2} \mathbf{v}_i. \tag{2.55}$$

3. Suppose that $\mathbf{v}$ is one of the generators for a zonohedron $\mathcal{Z}$. Then the *zone* of $\mathcal{Z}$ corresponding to $\mathbf{v}$ consists of all the faces of $\mathcal{Z}$ for which one edge is a translated copy of $\mathbf{v}$.
4. A set of vectors $\mathcal{G}$ in $\mathbb{R}^n$ is in *general position* if every subset of n vectors in $\mathcal{G}$ is linearly independent.
5. A set of non-negative vectors $\mathcal{G}$ in $\mathbb{R}^3$ is *cyclic* if no vector in $\mathcal{G}$ is contained within the convex cone of the other vectors. A zonohedron $\mathcal{Z}(\mathcal{G})$ that results from a cyclic generating set is also called *cyclic*.

Properties of Zonotopes and Zonohedra:

1. A zonotope is a convex polytope, with a polytope's usual structure of vertices, edges, k-dimensional faces, and so on.
2. If a zonohedron's generators are in general position, then
   (a) Each edge of the zonohedron is a translated copy of one and only one of the generators, and
   (b) Each face of the zonohedron is a parallelogram; furthermore, each edge of any such parallelogram is a translated copy of one of the generators.
3. The converse to Result 2 holds: if each edge of a zonohedron is a translated copy of a generator, and each face of the zonohedron is a parallelogram, then the zonohedron's generators are in general position.
4. A zonotope is centrally symmetric: the points

$$\sum_{1=1}^{m} \alpha_i \mathbf{v}_i \text{ and } \sum_{1=1}^{m} (1-\alpha_i)\mathbf{v}_i \tag{2.56}$$

   on the zonotope are the endpoints of a line segment whose halfway mark is the zonotope's center.
5. If $\mathbf{v}$ is a vertex of a zonotope, then there is a unique sequence of α_i's in Equation (2.53), which produce $\mathbf{v}$. Each α_i in that sequence is either a 0 or a 1. Furthermore, a hyperplane $\mathcal{H}$ through the origin can be found, such that any generator with coefficient 1 is on one side of $\mathcal{H}$, and any generator with coefficient 0 is on the other side.
6. The converse to Result 5 holds. Suppose the hyperplane $\mathcal{H}$ through the origin contains no vector in the generating set $\mathcal{G}$. Then the sum of all the generators on either side of $\mathcal{H}$ is a vertex of $Z(\mathcal{G})$.

7. Suppose no two generating vectors are linearly dependent, and that $\mathbf{v}$ is on an edge of the zonotope produced by the generating vectors. Then there is a unique sequence of α_i's in Equation (2.53), which produce $\mathbf{v}$. At most one α_i is strictly between 0 and 1; every other coefficient is either 0 or 1.
8. Suppose the generating vectors for a zonohedron in $\mathbb{R}^3$ are in general position (i.e no subset of three is linearly dependent), and that $\mathbf{v}$ is on the boundary of the zonohedron produced by the generating vectors. Then there is a unique sequence of α_i's in Equation (2.53), which produce $\mathbf{v}$. At most two α_i's are strictly between 0 and 1; every other coefficient is either 0 or 1.
9. Suppose a zonohedron is cyclic. Then each generating vector occurs as an edge that starts at the origin, and these are the only edges that emerge from the origin.
10. Similarly, each generating vector occurs as an edge that ends at the terminal point, and these are the only edges that end at the terminal point.
11. Each zone of a cyclic zonohedron contains both the origin and the terminal point.
12. The vertices of a cyclic zonohedron have a special form. Let the generating vectors be numbered cyclically (either clockwise or counterclockwise) from 1 to m. Then a sum of generating vectors is a vertex if and only if the indices of those generating vectors form a consecutive sequence. This sequence can be modular; for instance, $\{m-1, m, 1, 2\}$ would be considered a consecutive sequence.
13. The matrix pattern in Expression (2.50), when generalized to m entries rather than just four, gives the complete structure of vertices, edges, and faces for a cyclic zonohedron.

Chapter 3

Physical Factors in Colour Matching

3.1 Introduction

Both physical and perceptual factors contribute to human colour sensations. The perceptual factors result from the human visual system (HVS), including the eyes, the brain, and the optic nerve, and possibly involve psychological properties too. The physical factors behind colour, on the other hand, are independent of a human subject. They include the light sources which make vision possible, and the optical properties of perceived objects. In both theory and practice, these two factors are distinct. This chapter deals with physical factors; the next chapter deals with perceptual factors.

In an everyday viewing situation, a human's field of view consists of a scene with many objects, at various different locations. Some distribution of light enters the eyes from each direction. The distribution of light from a particular direction will be referred to as a *colour stimulus* or *visual stimulus*. A visual scene contains a multitude of colour stimuli, each of which, when interpreted by the human visual system, produces a colour sensation. Equivalently, a scene is a set of colour sensations, where each direction provides one stimulus.

A basic question is how these stimuli relate to the colours they produce. Colour science approaches this question by starting with simple cases, and moving to progressively more complicated cases—we will see, however, that even uncomplicated cases require elaborate analysis. In the simplest case, a viewer sees only a single stimulus, from a uniform patch of colour, against a "'black" background, where there is no light at all. In such experiments, observers can assign a hue (red, yellow, green, etc.) and a saturation (dull, vivid, etc.) to a stimulus. Even the hues, though, are limited: for example, no browns or greys can be produced. Furthermore, since the eye adapts to the ambient illumination level, one cannot usually say whether the stimulus is bright or dim.

The simplest case after an isolated stimulus—and the case of interest for this book—consists of two uniform patches of colour, viewed against a black background. The patches appear side by side. Typically they are both semicircles that together form a complete circle, subtending about 2° of the field of view, about the same angle subtended by a thumb held out at arm's length. The two colour patches can now be compared visually. In some cases, the colours look identical, or *match*, even though the two stimuli are physically different. Viewers in these colour-matching experiment adjust one of the stimuli until the two colours agree. By systematically varying the stimuli, and testing many human subjects, a substantial body of colour-matching data has been accumulated, which was codified in the 1931 *Standard Observer*, by the Commission Internationale de l'Éclairage (CIE). Despite their simplicity, colour-matching experiments lead to a rich vein of perceptual and geometric analysis, which this book works out in detail. Colour matching is basic to many important, and more advanced, topics in colour science, which this book will not treat, but for which it will lay a solid foundation.

This chapter begins the analysis of colour matching by examining colour stimuli: the physical light that enters the eye and produces a colour perception. While some colour stimuli travel directly from a physical light source to a perceiving eye, a more important case occurs when a stimulus results from light that leaves a source, and then reflects off an intermediate object, before finally reaching a human perceiver. The elements of this case are handled separately. First, light is considered solely as a radiometric power density over the visible spectrum. The shape of a density, independent of its magnitude, is described by an *illuminant*. An object's reflectance properties are given by its *reflectance spectrum*, which is a function that takes on values between 0 and 100% over the visible spectrum, and describes how much light of each wavelength is reflected. Radiometric densities, illuminants, and reflectance spectra all have natural vector space structures, which this chapter formalizes. Later chapters will find transformations from these physical vector spaces to perceptual and sensor vector spaces. The images of the transformations will lead to geometric constructions, including zonohedra, in the destination spaces.

3.2 Colour Stimuli

3.2.1 Radiometric Functions

The human visual system responds to light, which is electromagnetic radiation in the visible spectrum, where wavelengths are between about 400 and 700 nm; a visual or colour stimulus consists of light (provided its composition is constant for some reasonable time period) that reaches a human eye. A colour stimulus can be adequately described for colour science by the amount of power it contains at a limited set of wavelengths λ, or in a set of narrow wavelength bands. The unit of power is the Watt (W), so a colour stimulus will be modeled as a function that assigns a certain number of Watts to each wavelength, or wavelength band, in the visible spectrum. Such a function is called a *spectral power density* (SPD). (Technically, the power of a stimulus in a certain wavelength band is given by integrating the SPD over that band, and a stimulus is a *measure*, in the mathematical sense of measure theory. For our purposes, however, the common conception of an SPD as a function will suffice.) Figure 3.1 graphs an example of a colour stimulus given by an SPD. The horizontal axis is the wavelength, and the vertical axis is power.

In practical situations, the power or intensity of a stimulus is considered not just on its own, but in context, and the units of the vertical axis in Figure 3.1 change accordingly, although the change is typically unstated. An SPD, for instance, might be impinging on a piece of paper. Since the SPD delivers some amount of power to each area of the paper, natural units are Watts per square meter (W/m^2). On the other hand, an SPD might occur as the output of a light bulb, which is modeled as a point source. Then a solid angle centered at the bulb contains a certain amount of light, so natural units are Watts per steradian (W/sr). In many contexts, a stimulus has multiple simultaneous interpretations. For instance, a light bulb emits an SPD, which then illuminates a piece of paper. Even though they have different units, the emitted and the illuminating SPD have the same shape. Rather than make multiple SPD graphs with different units, only one graph is made, with no units given. Radiometry deals with the interaction of SPDs and objects, and the appropriate units, so the terms *radiometric*

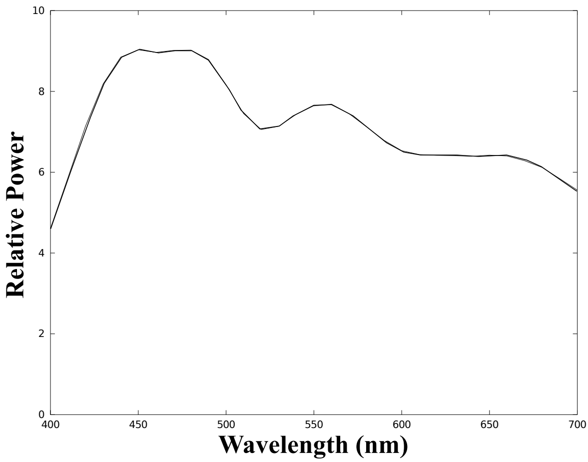

Figure 3.1: An Example of a Colour Stimulus, Graphed as an SPD

λ(nm)	SPD(λ)	λ(nm)	SPD(λ)
400	4.6098	560	7.6684
410	5.8696	570	7.4499
420	7.1440	580	7.1222
430	8.1854	590	6.7872
440	8.8481	600	6.5323
450	9.0302	610	6.4376
460	8.9646	620	6.4158
470	9.0156	630	6.4085
480	9.0229	640	6.3939
490	8.7898	650	6.4231
500	8.1636	660	6.4012
510	7.4499	670	6.2847
520	7.0566	680	6.1172
530	7.1367	690	5.8405
540	7.4353	700	5.5565
550	7.6611		

Table 3.1: Tabulated SPD Values for Figure 3.1

function, *SPD*, *colour stimulus*, and *visual stimulus* (or just *stimulus*) will be used as synonyms.

Though they are often drawn as continuous functions, colour stimuli are typically given as discrete tabulations. Table 3.1, for example contains the same data as Figure 3.1, but listed at 31 wavelengths, running from 400 to 700 nm in intervals of 10 nm. A 10 nm resolution is fine enough to account for most colour phenomena, so this book will treat colour stimuli as discrete functions over those 31 wavelengths. Adjustments can easily be made for finer or coarser resolutions.

3.2.2 Colour Stimuli as a Subset of a Vector Space

Because of the relatively low wavelength resolution needed, a vector space formulation of colour stimuli is both sufficient for colour science and computationally convenient. Throughout this book, therefore, the 31 power levels for a colour stimulus will be interpreted as the entries of a 31-dimensional vector. Each colour stimulus can equivalently be seen as a non-negative discrete function of the visible spectrum, over the wavelengths 400 nm, 410 nm, ..., 700 nm. The set $\mathcal{S}$ of colour stimuli can be viewed as a subset of the 31-dimensional real vector space $\mathbf{S}$ of discrete functions which take on arbitrary values at those wavelengths. This underlying vector space structure will be used extensively later, for further definitions and for geometric constructions.

Addition and scalar multiplication of the radiometric functions in $\mathcal{S}$ are consistent with the vector space operations in $\mathbf{S}$, and have natural physical interpretations. Addition corresponds to superposition. Suppose for instance that a piece of paper is illuminated simultaneously by two SPDs. Then the total illumination is a third SPD, which is the mathematical sum of the first two; physically, two stimuli have been superposed to produce a new stimulus. Scalar multiplying an SPD by a constant k corresponds to changing the SPD's absolute power levels while retaining its relative power levels; the graph will change its size, but not its shape. An obvious example is a dimmer switch for a light bulb, which increases or decreases brightness without changing the relative power outputs at different wavelengths.

The set $\mathcal{S}$ of SPDs is only a subset, and not a subspace, of the vector space $\mathbf{S}$, because functions in $\mathbf{S}$ can take on negative values. A vector in $\mathbf{S}$ is a physically possible SPD if and only if all its coordinates—which represent physical power levels—are non-negative. While a vector with some negative coordinates has no physical meaning, it is still useful mathematically, and we will later define some linear transformations on **S**, with the understanding that they are physically meaningful only when restricted to $\mathcal{S}$. Working with the larger space **S** allows us to apply the machinery of linear algebra, which will later prove essential for derivations.

While the coordinates for a vector space are arbitrary, an obvious choice of coordinates for **S** is the power levels at the 31 wavelengths 400 nm, 410 nm, ..., 700 nm. The corresponding basis vector for each coordinate is an indicator function, which take on the value 1 at that wavelength, and the value 0 everywhere else. Formally, define

$$\sigma_{400}(\lambda) = \begin{cases} 1, & \text{if } \lambda = 400, \\ 0, & \text{otherwise,} \end{cases} \tag{3.1}$$

to be the indicator function for 400 nm, and define other basis vectors similarly. The units of σ_{400} are chosen appropriately for the context. Such indicator functions are called *monochromatic*. As a whole, the set of indicator functions serve as a basis for **S**.

The total power of a radiometric function is an important physical concept, with a natural interpretation as a linear functional P on the vector space **S**. Define P first on basis vectors and then extend it to all vectors by linearity:

$$P\left(\sigma_\lambda\right) = 1, \tag{3.2}$$

where λ is any of the 31 basis wavelengths. The linearity of P is actually a physical statement, rather than a mathematical statement, relying on the fact that SPDs can be superposed without interacting. Later, another important functional on **S**, the luminance functional Y, will be seen to be perceptual rather than physical.

3.3 Illuminants and Light Sources

Some kind of light source or illumination is needed to produce a colour stimulus. A light source can produce a stimulus either directly, for example when a viewer looks right at a lit light bulb, or indirectly, for example after the light from a source reflects off some physical object. In either case, the SPD of the illuminating light affects the SPD of the stimulus reaching a human observer, and thus affects the resulting colour perception. This section discusses *illuminants* and *light sources*, two colour science terms that are relevant to illumination.

In typical everyday viewing, a light source or sources illuminate a scene containing multiple objects. Suppose for simplicity that there is a single light source, or, if there are multiple light sources, then they all have SPDs of the same shape. Obvious examples are an outdoor scene which is illuminated solely by the sun, or an indoors room that is lit by a multitude of identical fluorescent tubes. In either of these arrangements, the intensity of the lighting will vary greatly at different points and in different directions, due to the spreading of the light as it leaves the source, superposition of light from multiple sources, and shadows cast by some objects on other objects. In addition, a strongly directional source, such as the sun on a clear day, sheds much more light on the parts of objects that are facing the sun, and much less on the parts that are turned away from the sun.

In this example, the SPDs of the sunlight that reaches different parts of the scene vary greatly in intensity, but their shapes, that is, the relative power at various wavelengths, would not vary much. In a distinction that is not always observed, colour science applies the term *illuminant* to the relative SPD of illuminating light and the term *light source* to a physical object or phenomenon that produces that light. Any source, regardless of its intensity, whose SPD has the shape of a given illuminant, is said to be consistent with that illuminant. Two sources, such as a 100 W and a 200 W incandescent bulb, can thus have different magnitudes, yet, since their SPDs have the same shape, be consistent with the same illuminant. Another common example occurs as emitted light gets progressively farther from a point source such as a candle flame: the light's power decreases as the inverse square of the distance from the flame, but its SPD keeps the same shape.

Over the years, a series of standard illuminants has been defined. Illuminant A, shown in Figure 3.2 with two other illuminants, models light from incandescent sources. Illuminant C is an average of indirect daylight. Illuminant E has equal power at every wavelength. The vertical units in Figure 3.2 are arbitrary: they could be multiplied by any scalar constant without changing the illuminant. As with SPDs themselves, practical experience has shown that a 10 nm resolution is adequate to model illuminants for colour science. Thus an illuminant I will be treated as a 31-dimensional vector, over the same 31 wavelengths used previously. More precisely, an illuminant is an equivalence class of 31-dimensional vectors that are scalar multiples of one another, and which thus produce the same shape when graphed; this level of precision, however, is usually not needed. As 31-dimensional vectors, illuminants can be seen as elements of $\mathbf{S}$. Since an illuminant's power is non-negative at any wavelength, illuminants can also be written as elements of the set $\mathcal{S}$ of physically possible SPDs, and inherit

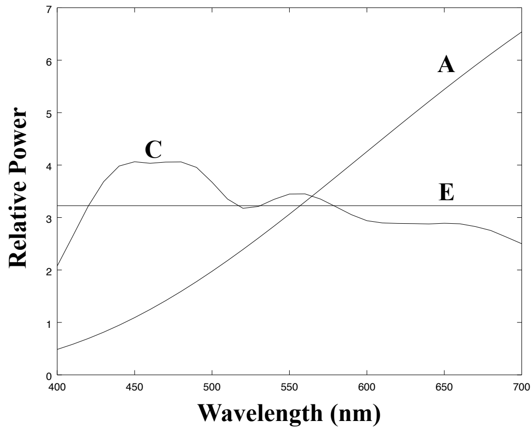

Figure 3.2: Some Standard Illuminants

the same natural vector space operations.

At various points in this book, an illuminant will be required to be *positive*, meaning that it has positive power at every wavelength; its graph over the visible spectrum never takes on the value 0. This mainly technical condition avoids degenerate cases. In practice, positiveness is the norm, because most light sources, especially natural ones, do have some power at every wavelength. Exceptions occur of course in artificial situations, such as a monochromator that outputs light of a very narrow bandwidth. Geometric constructions might also use theoretical illuminants given by monochromatic SPDs.

3.4 Reflectance Spectra

While some colour stimuli reach the eye when humans view a light source directly, for example by looking at the sun or a burning light bulb, stimuli more commonly result when lighting interacts with the objects in a scene before arriving at a human eye. Often, in fact, the light source is not even within the field of view. The most frequent form of interaction is *Lambertian* reflection, also called *matte* or *diffuse* reflection, which works on the various wavelengths independently. A Lambertian object absorbs a certain percentage of the incoming light at a particular wavelength, and reflects the rest in a Lambertian distribution: the amount reflected in a particular direction is not uniform, but is rather proportional to the cosine of that direction's angle from the normal. This section derives an expression for Lambertian reflection, and formulates mathematically how it modifies an incoming SPD, that is consistent with a certain illuminant, into another SPD that serves as a colour stimulus for a human viewer.

Physical objects, of course, can modify an SPD in many other ways than Lambertian reflection. For example, stained glass transmits some of an SPD and blocks the rest. A diamond or prism can separate one SPD into multiple SPDs. Oil films and diffraction gratings similarly decompose an SPD, weakening some wavelengths and strengthening others through interference. While this plethora of possible interactions has been thoroughly investigated, matte reflection is typical of everyday viewing, so this book will be largely restricted to this case. The term *physical object* should also be understand in a broad sense to encompass any material or matter with a surface that reflects light. A field of grass, for instance, would be an object in this context.

3.4.1 Lambertian Reflection

A surface is said to exhibit *Lambertian*, also called *matte* or *diffuse*, reflection, if it is opaque, shows no specular behavior, and reflects light in accordance with the Lambertian distribution. An opaque surface or material transmits no light; all incoming light is either reflected or absorbed. A purely specular surface reflects any incoming ray as a single outgoing ray; the two rays make equal angles with the surface normal, with which they are also coplanar. In diffuse reflection, any impinging ray, regardless of the incoming angle, exits the surface in all directions simultaneously, and the power in any direction is proportional to the cosine of that direction's angle from the normal. Such a power distribution is called *Lambertian*, after Johann Lambert, who established it in 1762. This section derives the form of the Lambertian distribution.

Surprisingly, the derivation relies only on common experience, rather than specialized measuring equipment. Suppose a viewer observes a wall that is covered with a coat of some matte paint. A spot on the wall can be viewed from multiple directions. For instance, one could view the spot along the surface normal, or at some angle θ off the normal. The wall is illuminated by some incoming light, likely from a variety of directions, and our eyes detect whatever part of that light reflects off the surface in the direction in which we are looking. Common experience tells us that the paint colour is constant, and does not appear any brighter or dimmer, regardless of the viewing angle. From this simple observation we will derive the Lambertian distribution.

Suppose the spot is viewed along the normal. To isolate the SPDs that travel into the eye along the normal, place a thin drinking straw, of radius ε, at the eye, and align it along the normal (see Figure 3.3). Then only the exiting rays that are parallel to the normal will reach the eye. All such rays exit the wall from a circle of radius ε, centered on the spot being viewed. The area of this circle is $\pi\varepsilon^2$, so the total power of the SPDs traveling through the straw is some power factor p times $\pi\varepsilon^2$.

Now view the same spot through the straw at some angle θ off the normal. Then the rays that are perceived through the straw all pass through the circular cross-section of the straw perpendicularly; they exit the wall from an ellipse, whose semi-axes are ε and $\varepsilon / \cos\theta$. The area of the ellipse is $(\pi / \cos\theta)\varepsilon^2$, which is considerably larger than the area of the circle in the normal direction. Apart from comparing the area, the total power in the two directions can also be compared. In general, the total power through a cross-section of the straw in direction θ is some power flux $f(\theta)$ across the originating area of the wall, times that area:

$$f(\theta)(\pi / \cos\theta)\varepsilon^2. \tag{3.3}$$

To determine $f(\theta)$, consider the common observation that the wall has the same colour, and “constant brightness,” no matter what direction it is viewed from. The total power of the SPDs reaching the eye from any direction must therefore be the same. Therefore

$$f(\theta)(\pi / \cos\theta)\varepsilon^2 = p\pi\varepsilon^2, \tag{3.4}$$

and so

$$f(\theta) = p\cos\theta. \tag{3.5}$$

$\theta = 0$ along the normal, so we get $p = f(0)$, implying

$$f(\theta) = f(0)\cos\theta. \tag{3.6}$$

Equation (3.6) states, counterintuitively, that the flux exiting the surface at an angle θ off the normal is less than the flux exiting along

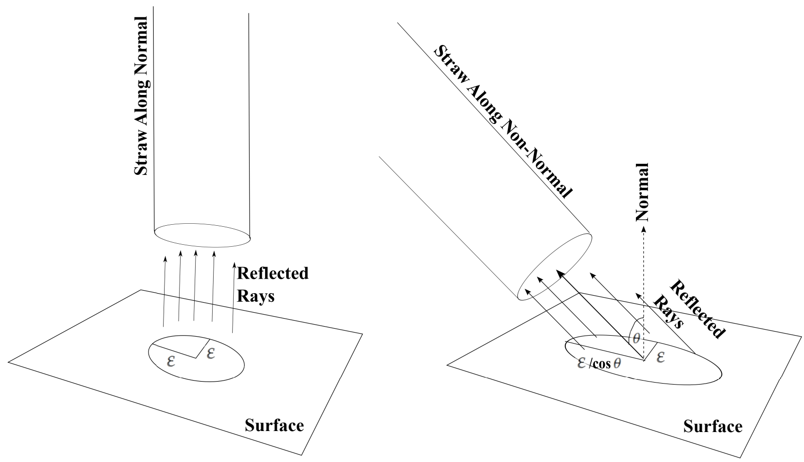

Figure 3.3: Reflected Rays in Different Directions

the normal, by a factor of $\cos\theta$. The difference can be significant: almost no power exits in directions that are nearly parallel to the surface. Figure 3.4 shows the relative amount of power flux in one plane through the normal. The length of the arrow in the direction given by the angle θ from the normal is proportional to the flux that exits in that direction. From Equation (3.6), it can be calculated that the tips of the arrows lie on a circle that is tangent to the surface. This circle just gives the results for one plane; to be extended to three-dimensional space, rotate the circle around the normal to make a sphere. In radiometry, the definition of *radiance*, referring to the light emitted by a surface such as a computer monitor, includes the factor $\cos\theta$; our derivations motivate this often unexplained factor.

The physical mechanism behind the Lambertian distribution requires another distinction, between *first-surface reflectance* and *bulk reflectance*. In first-surface reflectance, a ray reflects off the surface of an object without entering the object. In bulk reflectance, a ray enters the object before it is reflected. Once inside the object, it interacts with a variety of atoms or molecules, and can be scattered in many directions. The various photons of the ray take randomly determined paths, that can be wildly different. Some paths will cause photons to exit the object through the surface. The exiting photons can be traveling in many different directions, with no relation to their incoming

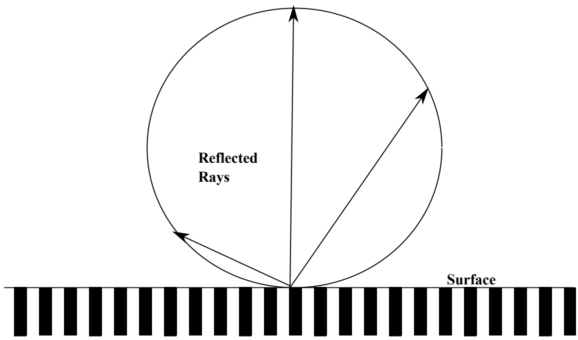

Figure 3.4: Lambertian Reflectance Distribution in a Plane Containing the Normal

direction. The observed Lambertian exiting distribution likely has a statistical explanation, but the details seem not to be well understood.

In practice, many surfaces that are mostly matte exhibit a slight gloss component, in which light that arrives at a certain angle to the normal, such as 30°, is reflected specularly, but light at any other angle is reflected diffusely. Gloss causes a highlight, sometimes sharp and sometimes blurry, that reflects the light source and moves on the surface as the viewer changes position. Both matte and gloss components can be encompassed in a *bi-directional reflectance distribution function* (BRDF). Suppose we fix an incoming direction and an outgoing direction, and shine an SPD towards the surface along the incoming direction. Then some percentage (perhaps 0) of the incoming power will be reflected in the outgoing direction. The BRDF gives that percentage, for every pair of incoming and outgoing directions. While they are the most comprehensive description of surface reflections, BRDFs are difficult to measure and cumbersome to calculate with. Though comprehensive, they are rarely necessary, because the gloss component of most surfaces is often small enough to be neglected.

This book restricts itself to opaque objects whose reflectance is solely diffuse. Gloss, transparency, iridescence, fluorescence, and so on, will not be treated. Though some interesting phenomena are eliminated, the case that is treated is in some sense the norm. A prototypical example of diffuse reflectance is acrylic paint, for either artists or housepainters, with a matte finish; these paints alone fill out the universe of object colours fairly well.

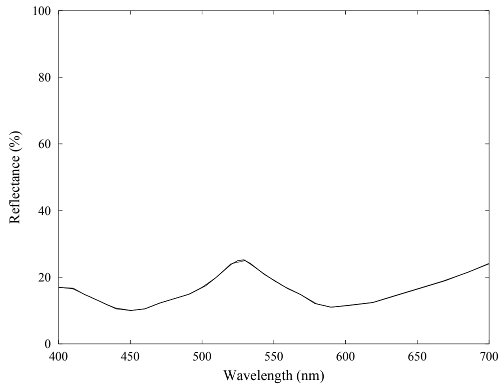

Figure 3.5: Reflectance Spectrum for Chromeoxide Green

3.4.2 Absorption and Reflectance

While a matte object reflects some incoming light diffusely, it typically also absorbs some incoming light, and only reflects the non-absorbed part. After being modified by absorption and reflectance, the SPD of the reflected light differs from the SPD of the illuminating light. For a person viewing an object, the reflected SPD is the colour stimulus associated with that object.

Absorption and reflectance are physical properties of a material, and, like colour stimuli, vary with wavelength. A material's *reflectance spectrum* gives the percentage of light that a material reflects, as a function of wavelength. Figure 3.5 shows such a spectrum for chrome-oxide green, an artist's pigment. The horizontal axis indicates wavelengths while the vertical axis indicates the percentage of light that is reflected. This material, for example, reflects 18 percent of incoming light of wavelength 500 nm, and 22 percent of incoming light of wavelength 700 nm. The reflectance percentage must, of course, be between 0 and 100. For an opaque matte material with no gloss component, any light that is not reflected is presumed to be absorbed, so the absorption percentage is just 100 minus the reflection percentage. A reflectance spectrum therefore fully describes how light interacts with a particular material.

Like an SPD, a reflectance spectrum can equally well be presented as a table. Also like an SPD, 31 wavelengths at a spacing of 10 nm has been found to be an adequate discretization for colour science.

As before, we can construct a vector space, which we will call $\mathbf{R}$, of all the functions at the 31 wavelengths, with no restrictions on those functions' values. The reflectance spectra make up a subset $\mathcal{R}$ of $\mathbf{R}$, where $\mathcal{R}$ consists of all functions with values between 0 and 100%. Equivalently, $\mathcal{R}$ consists of all functions with values between 0 and 1. A material's reflectance spectrum will be typically denoted as a function $\rho(\lambda)$, where λ represents wavelength.

The set of reflectance functions has two natural limits, described as *ideal black* and *ideal white*. An ideal black reflectance function reflects 0% of the incoming light at every wavelength, while ideal white reflects 100%. Ideal black would be the darkest surface that could possibly exist, while an ideal white surface, also called an *ideal diffuse reflector* would be the brightest. These two extremes have been closely approximated in practice, but not quite attained. Nevertheless, they are useful for theoretical and mathematical analysis.

The vector space $\mathbf{R}$ has a natural basis, analogous to the monochromatic basis for $\mathbf{S}$. The *maximal monochromatic reflectance (MMR) spectrum* at wavelength λ, denoted ρ_λ, is defined to take on the value 100% at λ, and the value 0 elsewhere. For example, the maximal monochromatic spectrum at 600 nm is given by

$$\rho_{600}(\lambda) = \begin{cases} 1, & \text{if } \lambda = 600, \\ 0, & \text{otherwise.} \end{cases} \tag{3.7}$$

Physically, such a material would reflect light only at the wavelength 600 nm (or at the wavelength band containing 600 nm), and absorb light of any other wavelength. We will use $\mathcal{M}$ to denote the set of the 31 MMR spectra, which form a vector space basis for $\mathbf{R}$. Unlike monochromatic SPDs, which are easily produced by a monochromator or lasers, no materials with MMR spectra are currently known to exist. Any one that did exist would be very dark, because it would only reflect a small portion ($1/31^{\text{st}}$) of the incoming light.

Since reflectance produces a modified SPD from an incoming SPD, a spectrum ρ can also be seen as a transformation from the vector space $\mathbf{S}$ to itself. No confusion will result if ρ denotes both the spectrum and the transformation. Of course, this transformation is physically meaningful only when restricted to the set $\mathcal{S}$ of SPDs, which ρ preserves. It is easy to see that ρ is a diagonal linear operator. The application of ρ to an incoming SPD σ can be written in matrix form:

$$\rho(\sigma) = \begin{bmatrix} \rho(400) & 0 & \dots & 0 & 0 \\ 0 & \rho(410) & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \rho(690) & 0 \\ 0 & 0 & \dots & 0 & \rho(700) \end{bmatrix} \begin{bmatrix} \sigma(400) \\ \sigma(410) \\ \dots \\ \sigma(690) \\ \sigma(700) \end{bmatrix} \tag{3.8}$$

$$= \begin{bmatrix} \rho(400)\sigma(400) \\ \rho(410)\sigma(410) \\ \dots \\ \rho(690)\sigma(690) \\ \rho(700)\sigma(700) \end{bmatrix}. \tag{3.9}$$

The last line gives the power levels, at each wavelength, of the colour stimulus resulting from an object of reflectance spectrum ρ, when illuminated by an SPD σ. The expression shows that it is produced by element-wise multiplication of ρ and σ's vector representations in $\mathbf{R}$ and $\mathbf{S}$.

3.5 Chapter Summary

This chapter has formalized mathematically the physical aspects of colour that are needed for analyzing colour matches. The main objects introduced were colour stimuli (or equivalently, radiometric functions, or SPDs), illuminants, and reflectance spectra. All these objects were cast as elements of appropriate vector spaces. To define any of these objects with enough accuracy for colour science, it is sufficient to write them as discrete functions over the set of 31 wavelengths, obtained by going from 400 nm to 700 nm (the boundaries of the visible spectrum) in steps of 10 nm.

More formally, the following terms were defined:

1. The 31-dimensional real vector space $\mathbf{S}$ of power level functions on the 31 wavelengths,
2. The monochromatic SPDs σ_λ, which take on the value 1 at wavelength λ and the value 0 elsewhere, form a basis for $\mathbf{S}$,
3. The set $\mathcal{S}$ of colour stimuli. $\mathcal{S}$ is the subset, though not the subspace, of $\mathbf{S}$ that consists of all non-negative functions on the 31 wavelengths. Physically, the non-negative functions correspond to SPDs, which are emitted by light sources and modified by material objects. An SPD that enters an observer's eye is thought of as a colour stimulus, because it can produce a colour sensation,
4. An SPD is said to be positive if it takes on only positive values, and no negative or zero values, over the entire visible spectrum,

5. An illuminant I is a relative SPD; it defines the shape of an SPD, but not its magnitude. Any SPD sharing that shape is said to be consistent with the illuminant. An illuminant could also be thought of as an equivalence class of SPDs, where two SPDs are equivalent if they are multiples of each other,
6. The 31-dimensional real vector space $\mathbf{R}$ of reflectance functions on the 31 wavelengths,
7. The set $\mathcal{M}$ of 31 maximal monochromatic reflectance (MMR) spectra ρ_λ, which take on the value 100% at wavelength λ and the value 0 elsewhere; they form a basis for $\mathbf{R}$,
8. The set $\mathcal{R}$ of reflectance spectra. $\mathcal{R}$ is the subset, though not the subspace, of $\mathbf{R}$ that consists of all functions on the 31 wavelengths that take on values between 0 and 100%. Each element ρ of $\mathcal{R}$ can also be seen as a diagonal linear operator on $\mathbf{S}$. Mathematically, a material with reflectance spectrum ρ will modify an SPD that strikes it, in accordance with this operator. Physically, the material absorbs some percentage of each wavelength of the incoming SPD and diffusely reflects the remaining percentage. The materials of interest for this book are assumed to be opaque and to interact with light only through absorption and diffuse reflectance.

Index

Colour science draws on a variety of disciplines, including physics, biology, human perception, mathematics, and art. This book shows the part that geometry plays in reaching some important conclusions in colour science. Seemingly disparate mathematical objects that arise in human vision, machine vision, and electronic displays, are shown to share a common form as zonohedra. Their internal structures all arise as Minkowski sums of vectors that correspond to individual wavelengths in the visible spectrum. The processes of light production, reflection, and response provide the relationships that define those structures.

The first two chapters lay the geometric foundation for the colour science introduced in the rest of the book. Chapter 2 introduces Minkowski sums and zonohedra from first principles, in more detail than has appeared previously. The next two chapters deal with physical and perceptual aspects of colour, deriving the 1931 Standard Observer from empirical data. The final three chapters build on the first four to construct geometric objects for colour science, and to derive conclusions from them.

The book assumes no knowledge of colour science. Some linear algebra is assumed, at about the second-year undergraduate level. Even readers without this background, however, will be able to follow the book's concrete, intuitive presentation, which emphasizes spatio-visual understanding.