Convexity and Transitions

a strict examination of the 1931 CIE inverted-U

Glenn Davis <gdavis@gluonics.com>

February 10, 2025

1 Introduction

In the fundamental paper [Log09], Logvinenko investigates the statement that a color is optimal iff it comes from a (reflectance or transmittance) spectrum that only takes the values 0 and 1, and has 0 or 2 transitions. He calls this the *two-transition assumption*. He plots chromaticity diagrams for the cone fundamentals of Govardovskii et al. (Figure 4) and of Stockman et Sharpe (Figure 7) and remarks that they are not convex. By a theorem in [Wes83], there are optimal colors for these two sets of cone fundamentals whose transmittance spectra have more than 2 transitions. So the two-transition assumption is false in these two cases. He also plots the standard 1931 CIE chromaticity diagram (Figure 5) and remarks:

> However, the completed spectral contour (in the unit plane) derived from the color matching functions adopted by the CIE as the standard colorimetric observer (Figure 5) is convex. This indicates that for this observer the two-transition assumption holds true. *[page 5]*

The goal of this vignette is to show that, from an extremely strict viewpoint, the standard 1931 CIE inverted-U is not convex either, and the two-transition assumption does not hold.

To state this all precisely requires a lot of tedious mathematics, which is then followed by an analysis at both 5nm and 1nm.

The featured functions from **colorSpec** used in this vignette are `responsivityMetrics()`, `canonicalOptimalColors()`, and `bandRepresentation()`.

```
library( colorSpec )
```

2 Wavelengths and Subintervals

Suppose we are given N wavelengths: $\lambda_1 < \lambda_2 < \ldots < \lambda_N$. Define N intervals $I_i := [\beta_{i-1}, \beta_i]$ where

$$\beta_0 := \tfrac{3}{2}\lambda_1 - \tfrac{1}{2}\lambda_2 \qquad \beta_i := (\lambda_i + \lambda_{i+1})/2,\ i{=}1,\ldots,N-1 \qquad \beta_N := \tfrac{3}{2}\lambda_N - \tfrac{1}{2}\lambda_{N-1} \tag{2.1}$$

The intervals I_i are a partition of $[\beta_0, \beta_N]$. Note that $[\beta_0, \beta_N]$ is slightly bigger than $[\lambda_1, \lambda_N]$ because the endpoints are extended. Define the $i'th$ step $\mu_i := \text{length}(I_i),\ i{=}1,\ldots,N$. If the sequence $\{\lambda_i\}$ is *regular* (μ_i is constant), then $\{\beta_i\}$ is regular with the same step, and each λ_i is the center of I_i.

3 Band Functions

Let B be the set of all functions on $[\beta_0, \beta_N]$ that take the values 0 or 1 and have finitely many transitions (jumps). As in [Cen13], we identify the endpoints β_0 and β_N to form a circle, so if the values at β_0 and β_N are different, then this is considered to be a transition. Equivalently B is the set of all indicator functions $\mathbf{1}_S$ where S is a disjoint unit of finitely many arcs in the circle. We call these arcs *bands*. For a given function $f \in B$, twice the number of the bands is the number of transitions, unless S is the entire circle when there is 1 band and 0 transitions. In any case the number of transitions is even. We think of such an $f(\lambda)$ as a transmittance function of a filter, and a superposition of bandpass and bandstop filters. If the endpoints are in the interior of a band, then the band corresponds to a bandstop filter, and otherwise it corresponds to a bandpass filter. It is clear that a given f has either 0 or 1 bandstop filters.

Let $[0,1]^N$ denote the N-cube and define a function $p()$

$$p : B \twoheadrightarrow [0,1]^N \quad \text{by} \quad p(f) := \mathbf{y} \equiv (y_1, \ldots, y_N) \quad \text{where} \quad y_i = {\mu_i}^{-1} \int_{I_i} f(\lambda)\, d\lambda \tag{3.1}$$

Note that y_i is the mean of f on I_i. It is straightforward to show that $p()$ is surjective and it follows that $p()$ has a right-inverse (or *section*), i.e. a function $p^+ : [0,1]^N \to B$ so that $p \circ p^+$ is the identity on $[0,1]^N$. Such a section is fairly easy to construct, but $p^+(\mathbf{y})$ is certainly not unique, except in special cases. If $\mathbf{v} \in [0,1]^N$ is a vertex of the cube, then $p^+(\mathbf{v})$ *is* unique. Another important case is $\mathbf{y}_{ij} = (0, \ldots, 0, y_i, 1, \ldots, 1, y_j, 0, \ldots, 0)$ and $y_i, y_j \in (0,1)$. There is a unique $f \in p^{-1}(\mathbf{y}_{ij})$ with 2 transitions (1 passband), but an arbitrarily large number of bands of f in the intervals I_i and I_j can be created without changing the value of $p()$. In the extreme case where $\mathbf{y}$ is in the interior of the cube (all $y_i \in (0,1)$), there is a band function $f \in p^{-1}(\mathbf{y})$ with $\lceil N/2 \rceil$ bands.

In **colorSpec** software, the function $p()$ is implemented as `bandMaterial()`, and $p^+()$ is implemented as `bandRepresentation()`. In the latter case, the function tries to find a function with the minimum number of bands; see the corresponding man page for details.

4 Responsivity Function

Let $\mathbf{w} : [\beta_0, \beta_N] \to \mathbb{R}^3$ be a step function that take the constant value $\mathbf{w}_i$ on I_i. Define a function

$$\Gamma : B \to \mathbb{R}^3 \quad \text{by} \quad \Gamma(f) := \int_{\beta_0}^{\beta_N} f(\lambda)\mathbf{w}(\lambda)\, d\lambda \;\; = \;\; \sum_i^N \left(\int_{I_i} f(\lambda)\, d\lambda \right) \mathbf{w}_i \tag{4.1}$$

And define a similar function

$$\Gamma^N : [0,1]^N \to \mathbb{R}^3 \quad \text{by} \quad \Gamma^N(y) \;=\; \Gamma^N(y_1, \ldots, y_N) \;\; := \;\; \sum_i^N y_i \mu_i \mathbf{w}_i \tag{4.2}$$

By 3.1 it follows that $\Gamma^N(p(f)) = \Gamma(f)$. Define $Z := \Gamma^N([0,1]^N)$; since Z is the linear image of a cube, Z is a *zonohedron*, see [Cen13]. We now have a commutative diagram in which all 3 maps are surjective:

$$\begin{tikzcd}
B \arrow[d, two heads, "p"'] \arrow[dr, two heads, "\Gamma"] & \\
{[0,1]^N} \arrow[r, two heads, "\Gamma^N"'] & Z
\end{tikzcd}$$

If $f \in B$ has 0 or 2 transitions, then $\Gamma(f)$ is called a *Schrödinger color*, see [Wes83].

In **colorSpec** software, the function $\Gamma^N()$ is implemented in `product()`, and is a simple matrix multiplication, see the corresponding man page for details.

5 Chromaticity Polygons

From this point on, we require that all $\mathbf{w}_i$, $i = 1, \ldots, N$ lie in some linear open halfspace in $\mathbb{R}^3$, except if $\mathbf{w}_i$=0. This means that there is a vector $\mathbf{u}$ so that all $\langle \mathbf{u}, \mathbf{w}_i \rangle > 0$, except if $\mathbf{w}_i$=0. If all responsivities are non-negative, which is the usual case, then we can take $\mathbf{u}$=$(1, 1, 1)$. We now define the vertices $\mathbf{v}_i := \mathbf{w}_i / \langle \mathbf{u}, \mathbf{w}_i \rangle$ which are in the plane $\{\mathbf{v} | \langle \mathbf{v}, \mathbf{u} \rangle = 1\}$. These are the vertices of what we call the *chromaticity polygon* P in the previously mentioned plane. The CIE inverted-U is the classical example; where $\mathbf{w}_i$ is $(\bar{x}, \bar{y}, \bar{z})$ at λ_i, and $\mathbf{v}_i$ is the CIE chromaticity (x, y) at λ_i (after the final coordinate z of $\mathbf{v}_i$ is dropped).

We also consider the central projection of P onto the unit sphere S^2, and call this the *spherical chromaticity polygon* P_S. It is clearly contained in the hemisphere centered at $u/|u|$. The internal angles of P and P_S may be different, but whether an internal angle θ is convex ($\theta < \pi$), straight (θ=π), or concave|reflex ($\theta > \pi$), is the same in P and P_S.

If for all distinct indexes i, j, k, the vectors $\mathbf{w}_i, \mathbf{w}_j, \mathbf{w}_k$ are linearly independent we say that the responsivities are in *general position.* If they are *not* in general position, then $\mathbf{w}_i, \mathbf{w}_j, \mathbf{w}_k$ are linearly dependent for some distinct i, j, k, which means that one of these 3 is a linear combination of the other 2. By re-indexing assume the one is $\mathbf{w}_i$ and the others are $\mathbf{w}_j$ and $\mathbf{w}_k$. There are 3 ways such a degeneracy can happen:

1. $\mathbf{w}_i = 0$
2. $\mathbf{w}_i = \alpha \mathbf{w}_j$, where $\alpha \neq 0$ and $\mathbf{w}_j \neq 0$
3. $\mathbf{w}_i = \alpha \mathbf{w}_j + \beta \mathbf{w}_k$, where $\alpha \neq 0$, $\beta \neq 0$, and $\mathbf{w}_j, \mathbf{w}_k$ are linearly independent

For the chromaticity polygon P, with 2D vertices $\mathbf{v}_i$, these translate to 3 polygon degeneracies:

$1'$. vertex $\mathbf{v}_i$ is undefined

$2'$. vertices $\mathbf{v}_i$ and $\mathbf{v}_j$ are identical

$3'$. vertices $\mathbf{v}_i$, $\mathbf{v}_j$, and $\mathbf{v}_k$ are distinct but collinear, with $\mathbf{v}_i$ between $\mathbf{v}_j$ and $\mathbf{v}_k$

The chromaticity polygon P is not simple in general; it is just a closed polygonal path. In the next section we discuss the case where P is *convex*, which means that all internal angles are $\leq \pi$. For convex P we allow all 3 of these degeneracies. However, each group of identical vertices and each group of distinct collinear vertices must have contiguous indexes. A subset of $\{1, \ldots, N\}$ is *contiguous* iff the indexes are consecutive, with wraparound from N to 1 allowed. So for this vignette, a convex P is simple, except possibly for contiguous identical vertices.

6 The Optimal Color Theorem

The preliminaries are done and we can finally state the main result from [Wes83]:

Theorem 6.1 *With Γ, Z, and P as defined above, the following are equivalent:*

1. *for any $z \in Z$, $z \in \partial Z$ iff there is an $f \in \Gamma^{-1}(z)$ with 0 or 2 transitions*
2. *the chromaticity polygon P is convex*

Moreover, in part 1, $p(f)$ is unique for all z iff all vertices of P are defined.

A point $z \in \partial Z$ is called an *optimal color*. A corollary of the theorem is that if P is not convex, then there are optimal colors that are not Schrödinger colors. We explore examples of this in the next two sections.

7 The CIE xyz Responsivities with 5nm step

In **colorSpec** software, the CIE responsivities with 5nm step are stored in the object `xyz1931.5nm`; whose values are taken from Table 1 in [AST01]. The wavelengths range from 380 to 780 nm.

Analyze the responsivities, and print the degeneracies.

```
mets = responsivityMetrics( xyz1931.5nm )
mets$zeros

[1] 780
```

So the responsivity at λ=780 nm is 0. This is not a violation of the convexity of P.

```
mets$multiples

[[1]]
[1] 735 745

[[2]]
[1] 755 760

[[3]]
[1] 765 770 775
```

There are 3 groups of multiples: 735 745 nm (not contiguous), 755 760 nm (contiguous), and 765 770 775 nm (contiguous). The non-contiguous group is a violation of the convexity of P. Now print the actual concavities in P.

```
mets$concavities

   wavelength      extangle
2         385 -2.478207e+00
3         390 -2.208011e+00
7         410 -2.282285e-01
13        440 -2.993477e-02
14        445 -1.140541e-02
41        580 -2.248750e-03
42        585 -4.388215e-05
44        595 -1.760213e-03
45        600 -4.783000e-04
46        605 -3.809528e-03
49        620 -6.852491e-03
50        625 -3.987921e-03
```

These are all violations too. The column `extangle` is the external angle at the vertex (in radians) of the spherical chromaticity polygon P_S. The sum of internal and external angles is π, so when the external angle is negative, as these are, the internal angle is greater than π. In the vicinity of these wavelengths, we can find optimal colors with more than 2 transitions. As an example, we choose the canonical optimal color with wavelengths 580 and 585 nm.

```
wave  = wavelength(xyz1931.5nm)
E.eye = product( illuminantE(1,wave=wave), '*', xyz1931.5nm )
spec = canonicalOptimalColors( E.eye, c(580,585), spectral=TRUE )
bandRepresentation( spec )[[1]]

    lambda1 lambda2
BP1   567.5   580.0
BP2   585.0   592.5
BP3   752.5   762.5
```

So this spectrum is a superposition of 3 bandpass filters, and has 6 transitions.

```
par( omi=c(0,0,0,0), mai=c(0.5,0.6,0.2,0) )
plot( spec, main=FALSE, legend=FALSE, type='step', lwd=c(3,0.25) )
```

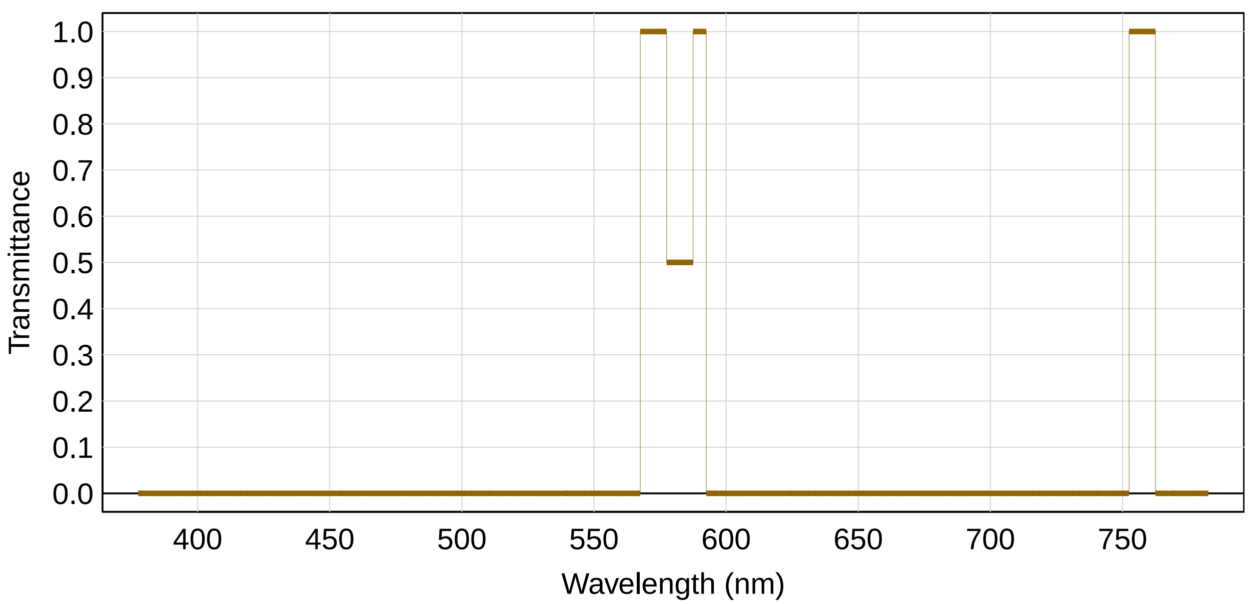

Figure 7.1: An example of a transmittance spectrum that is optimal, but has more than 2 transitions

8 The CIE xyz Responsivities with 1nm step

In **colorSpec** software, the CIE responsivities with 1nm step are stored in the object `xyz1931.1nm`; whose values are taken from Table 1 in [WS00]. The wavelengths range from 360 to 830 nm.

Analyze the responsivities, and print the degeneracies.

```
mets = responsivityMetrics( xyz1931.1nm )
mets$zeros

numeric(0)


mets$multiples
```

```
[[1]]
  [1] 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719
 [22] 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740
 [43] 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761
 [64] 762 763 764 765 766 767 768 769 770 771 772 773 774 775 776 777 778 779 780 781 782
 [85] 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800 801 802 803
[106] 804 805 806 807 808 809 810 811 812 813 814 815 816 817 818 819 820 821 822 823 824
[127] 825 826 827 828 829 830
```

So there are no wavelengths where the responsivity is 0. But all responsivities from 699 to 830 are multiples of each other (with angular tolerance of about 10^{-6} radian). It is fairly obvious that they were extrapolated in this way intentionally. Since these wavelengths are contiguous, there are no convexity violations so far. Now examine the concavities in P.

```
nrow( mets$concavities )

[1] 73
```

This is too many concave vertices to print, so look at the first quartile of external angles instead.

```
fivenum( mets$concavities$extangle )

[1] -3.611408e-01 -1.606363e-02 -8.717753e-04 -3.584991e-04 -3.014621e-06


mets$concavities[ mets$concavities$extangle <= -0.01606, ]

   wavelength    extangle
6         365 -0.02804187
7         366 -0.02013442
13        372 -0.17972851
14        373 -0.18254734
15        374 -0.12337220
16        375 -0.07284921
17        376 -0.02397177
24        383 -0.36114083
25        384 -0.31964638
26        385 -0.18907165
27        386 -0.03526279
33        392 -0.10245669
34        393 -0.18150335
35        394 -0.21463185
36        395 -0.15618045
37        396 -0.03375087
48        407 -0.01606363
49        408 -0.04297267
50        409 -0.05929131
```

```
wave  = wavelength(xyz1931.1nm)
E.eye = product( illuminantE(1,wave=wave), '*', xyz1931.1nm )
spec = canonicalOptimalColors( E.eye, c(407,409), spectral=TRUE )
bandRepresentation( spec )[[1]]

    lambda1 lambda2
BP1   403.5   407.0
BP2   409.0   415.5
```

So this spectrum is a superposition of 2 bandpass filters, and has 4 transitions.

```
par( omi=c(0,0,0,0), mai=c(0.5,0.6,0.2,0) )
plot( spec, main=FALSE, legend=FALSE, type='step', lwd=c(3,0.25) )
```

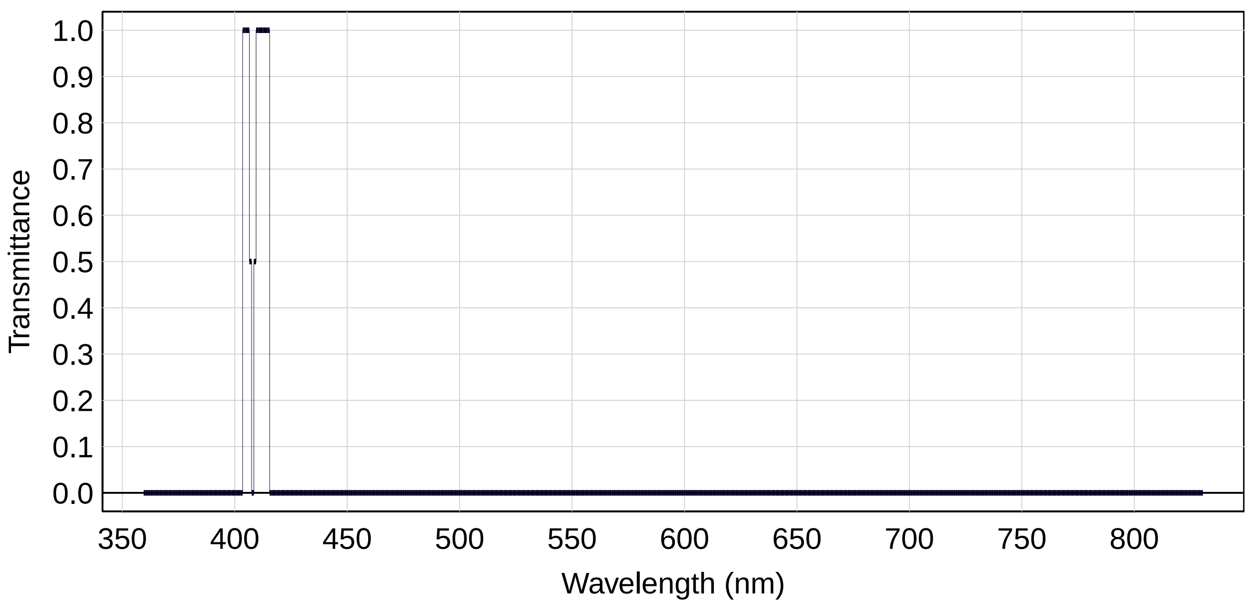

Figure 8.1: An example of a transmittance spectrum that is optimal, but has more than 2 transitions

References

[AST01] ASTM E308-01. Standard Practice for Computing the Colors of Objects by Using the CIE System. Technical report, West Conshohocken, PA, 2001.

[Cen13] Paul Centore. A zonohedral approach to optimal colours. *Color Research & Application*, 38(2):110–119, April 2013.

[Log09] Alexander D. Logvinenko. An object-color space. *Journal of Vision*, 9(11):5, 2009. `https://jov.arvojournals.org/article.aspx?articleid=2203976`.

[Wes83] West, G. and Brill, M. H. Conditions under which Schrödinger object colors are optimal. *Journal of the Optical Society of America*, 73:1223–1225, 1983.

[WS00] G. Wyszecki and W.S. Stiles. *Color Science: Concepts and Methods, Quantitative Data and Formulae*. Wiley Series in Pure and Applied Optics. Wiley, 2000.

Session Information

This document was prepared February 10, 2025 with the following configuration:

- R version 4.4.2 (2024-10-31 ucrt), `x86_64-w64-mingw32`
- Running under: `Windows 11 x64 (build 26100)`
- Matrix products: default
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: colorSpec 1.7-0, knitr 1.49, spacesRGB 1.7-0
- Loaded via a namespace (and not attached): R6 2.5.1, bslib 0.8.0, cachem 1.1.0, cli 3.6.3, compiler 4.4.2, digest 0.6.37, evaluate 1.0.1, fastmap 1.2.0, glue 1.8.0, highr 0.11, htmltools 0.5.8.1, jquerylib 0.1.4, jsonlite 1.8.9, lifecycle 1.0.4, logger 0.4.0, microbenchmark 1.5.0, rlang 1.1.4, rmarkdown 2.29, sass 0.4.9, tools 4.4.2, xfun 0.49, yaml 2.3.10, zonohedra 0.4-0