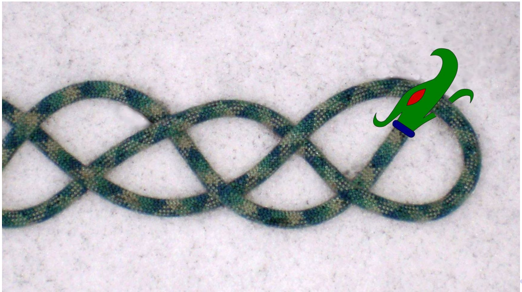
Information Theory on Convex sets

In celebration of Prof. Shun'ichi Amari's 80 years birthday

Peter Harremoës

Copenhagen Business College

June 2016



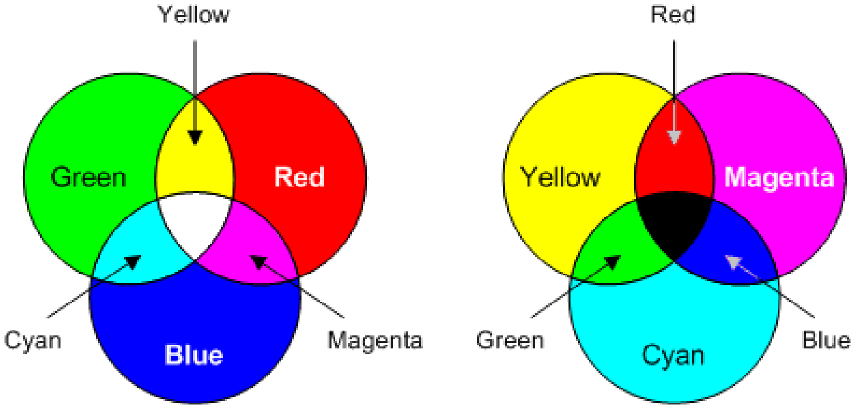
Outline

- Introduction.
- Convex sets and decompositions into extreme points.
- Spectral convex sets.
- Bregman divergences for convex optimization.
- Sufficiency and locality.
- Reversibility.

Some major questions

- Is information theory mainly a theory about sequenses?
- Is it possible to apply thermodynamic ideas to systems without conservation of energy?
- Why do information theoretic concepts appear in statistics, physics and finance?
- How important is the notion of reversibility to our theories?
- Why are complex Hilbert spaces so useful for representations of quantum systems?

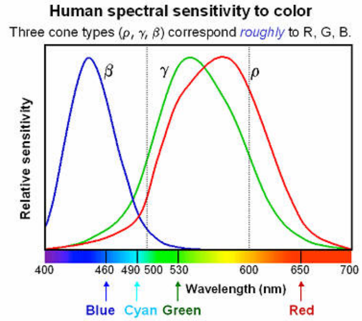
Color diagram



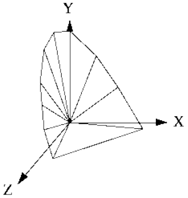
Nice but wrong!

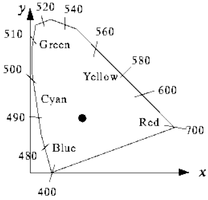
Color vision

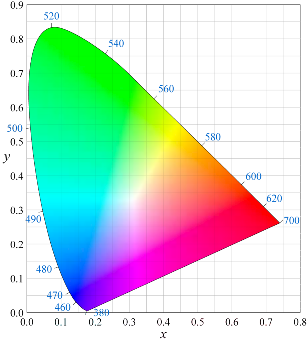
The human eye senses color using the cones. Rods are not used for color but for periferical vision and night vision.



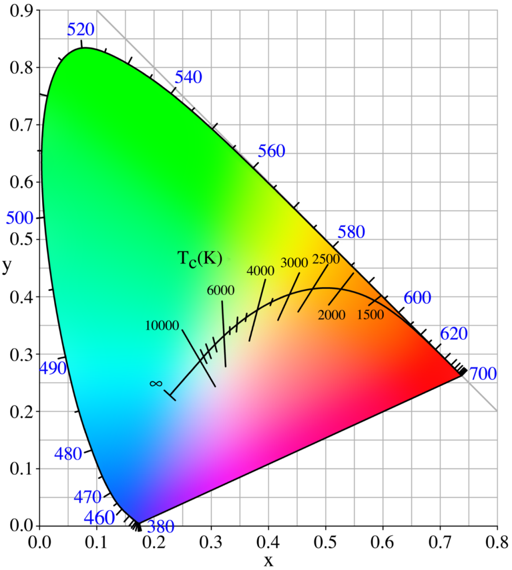
Primates have three 3 receptors.
Most mammels have 2 color receptors.
Birds and reptiles have 4 color receptors.


Y
X
Z

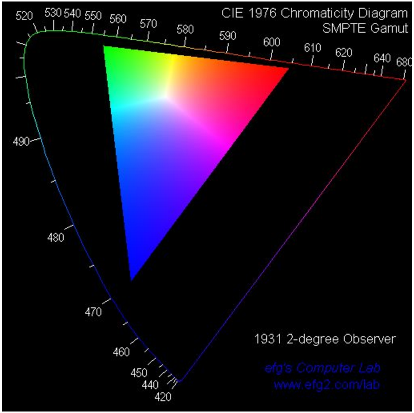

y
520
540
510
560
Green
580
500
Yellow
600
Cyan
490
Red
700
Blue
480
x
400


520
540
560
580
600
620
700
500
490
480
470
460
380
y
x
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

Black body radiation



VGA screen


CIE 1976 Chromaticity Diagram
SMPTE Gamut
520
530 540 550
560
570
580
590
600
610
620
640
680
490
480
470
460
450
440
420
1931 2-degree Observer
efg's Computer Lab
www.efg2.com/lab

The state space

Before we do anyting we prepare our system. Let $\mathcal{P}$ denote the set of preparations.

Let p_0 and p_1 denote two preparations. For $t \in [0, 1]$ we define $(1 - t) \cdot p_0 + t \cdot p_1$ as the preparation obtained by preparing p_0 with probability $1 - t$ and t with probability t.

A measurement m is defined as an affine mapping of the set of preparations into a set of probability measures on a measurable space.

Let $\mathcal{M}$ denote a set of feasible measurements.

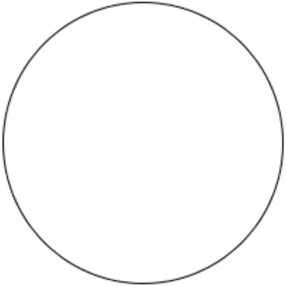
The state space $\mathcal{S}$ is defined as the set of preparations modulo measurements. Thus, if p_1 and p_2 are preparations then they represent the same state if

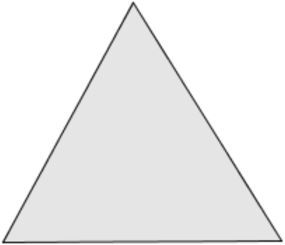
$$m(p_1) = m(p_2)$$

for any $m \in \mathcal{M}$.

The state space

Often the state space equals the set of preparations and has the shape of a simplex.
In quantum theory the state space has the shape of the density matrices on a complex Hilbert space.



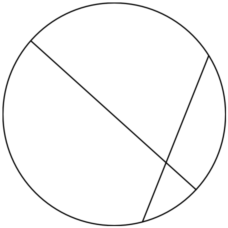


Example: Bloch sphere

- A qubit can be described by a density matrix of the form

$$\begin{pmatrix} \frac{1}{2}+x & y+iz \\ y-iz & \frac{1}{2}-x \end{pmatrix}$$

where $x^2+y^2+z^2 \leq 1/4$.
- The pure states are the states on the boundary.
- The mixed states are all interior points of the ball.



Orthogonal states

We say that two states s_0 and s_1 are *mutually singular* if there exists a measurement m with values in $[0, 1]$ such that $m(s_0) = 0$ and $m(s_1) = 1$. We say that s_0 and s_1 are *orthogonal* if there exists a face $F \subseteq S$ such that s_0 and s_1 are mutually singular as elements of F.

Lemma Any state that is algebraically interior in the state space can be written as a mixture of two mutually singular states.

Proof Use Borsuk–Ulam theorem from topology.

Improved Caratheodory Theorem In a state space of dimension d any state can be written as a mixture of at most $d + 1$ orthogonal states.

Entropy of a state

Let s denote a state. Then the entropy of s cen be defined as

$$H(s) = \inf \left\{ -\sum_i p_i \cdot \ln p_i \right\}$$

where the infimum is taken over all probability vectors $(p_{1,} p_2, \dots)$ such that there exists states $s_1, s_2, \dots$ that are extreme points such that

$$s = \sum_i p_i \cdot s_i .$$

According to Caratheodory's theorem $H(s) \leq \ln(d+1)$ when the state space has dimension d. We define the entropy of a state space S as $\sup_{s \in S} H(s)$ where the supremum is taken over all states in the state space. We define the spectral dimension of the state space S as

$$\exp(H(S)) - 1.$$

Entropic proof

$$
\begin{aligned}
H(s) &= -\sum_{i=0}^{d} p_i \cdot \ln(p_i) \\
&= (p_0 + p_1)\left(-\frac{p_0}{p_0+p_1}\ln\left(\frac{p_0}{p_0+p_1}\right) - \frac{p_1}{p_0+p_1}\ln\left(\frac{p_1}{p_0+p_1}\right)\right) \\
&\quad - (p_0 + p_1)\ln(p_0 + p_1) - \sum_{i=2}^{d} p_i \cdot \ln(p_i)
\end{aligned}
$$

and

$$
\begin{aligned}
s &= \sum_{i=0}^{d} p_i \cdot s_i \\
&= (p_0 + p_1)\left(\frac{p_0}{p_0+p_1}\cdot s_0 + \frac{p_1}{p_0+p_1}\cdot s_2\right) + \sum_{i=2}^{d} p_i \cdot s_i \,.
\end{aligned}
$$

Spectral sets

Definition

If $p_0 \leq p_1 \leq p_2 \cdots \leq p_d$ and $s = \sum_{i=0}^{d} p_i \cdot s_i$ where s_i are orthogonal we say that the vector p_0^d is a *spectrum* of s. We say that s is a *spectral state* if s has a unique spectrum. We say that the convex compact set C is *spectral* if all states in C are spectral.

Theorem

For a spectral set the entropic dimension equals the maximal number of orthogonal states minus one.

Proof.

Assume that the maximal number of orthogonal states is n. Any state can be written as a mixture of n states, and a mixture of at n states has entropy at most $\ln(n)$. The uniform distribution on n states has entropy $\ln(n)$. □

Examples of spectral sets

- A simplex.
- A d-dimensional ball.
- Density matrices over the real numbers.
- Density matrices over the complex numbers.
- Density matrices over the quaternions.
- Density matrices in Von Neuman algebras.

Actions

Let $\mathcal{A}$ denote a subset of the feasiable measurements $\mathcal{M}$ such that $a \in A$ maps S into a distribution on $\mathbb{R}$ i.e. a random variable.
The elements of A should represent actions like
* The score of a statistical decision.
* The energy extracted by a certain interaction with the system.
* (Minus) the lenth of a codeword of the next encoded input letter using a specific code book.
* The revenue of using a certain portfolio.

Optimization

For each $s \in \mathcal{S}$ we define

$$\langle a, s\rangle = E\left[a\left(s\right)\right].$$

and

$$F\left(s\right) = \sup_{a \in \mathcal{A}} \langle a, s\rangle .$$

Without loss of generality we may assume that the set of actions $\mathcal{A}$ is closed so that we may assume that there exists $a \in \mathcal{A}$ such that $F\left(s\right) = \langle a, s\rangle$ and in this case we say that a is optimal for s. We note that F is convex but F need not be strictly convex.

Regret

Definition

If $F(s)$ is finite *the regret* of the action a is defined by

$$D_F(s, a) = F(s) - \langle a, s \rangle$$

The regret D_F has the following properties:

- $D_F(s, a) \geq 0$ with equality if a is optimal for s.
- If $\bar{a}$ is optimal for the state $\bar{s} = \sum t_i \cdot s_i$ where $(t_1, t_2, \ldots, t_\ell)$ is a probability vector then

$$\sum t_i \cdot D_F(s_i, a) = \sum t_i \cdot D_F(s_i, \bar{a}) + D_F(\bar{s}, a).$$

- $\sum t_i \cdot D_F(s_i, a)$ is minimal if a is optimal for $\bar{s} = \sum t_i \cdot s_i$.

Bregman divergence

Definition

If $F(s_1)$ is finite *the regret of the state* s_2 is defined as

$$D_F(s_1, s_2) = \inf_a D_F(s_1, a) \tag{1}$$

where the infimum is taken over actions a that are optimal for s_2.

If the state s_2 has the unique optimal action a_2 then

$$F(s_1) = D_F(s_1, s_2) + \langle a_2, s_1 \rangle$$

so the function F can be reconstructed from D_F except for an affine function of s_1. The closure of the convex hull of the set of functions $s \to \langle a, s \rangle$ is uniquely determined by the convex function F.
The regret is called a *Bregman divergence* if it can be written in the following form

$$D_F(s_1, s_2) = F(s_1) - (F(s_2) + (s_1 - s_2) \cdot \nabla F(s_2)).$$

Properties of Bregman divergences

The Bregman divergence has the following properties:

- $d(s_1, s_2) \geq 0$
- $d(s_1, s_2) = a_2(s_1) - a_2(s_2)$ where a_2 denotes the action for which $F(s_2) = a(s_2)$.
- $\sum t_i \cdot d(s_i, \tilde{s}) = \sum t_i \cdot d(s_i, \hat{s}) + d(\tilde{s}, \hat{s})$ where $\hat{s} = \sum t_i \cdot s_i$.
- $\sum t_i \cdot d(s_i, \tilde{s})$ is minimal when $\hat{s} = \sum t_i \cdot s_i$.

Sufficiency

- Let (P_θ) denote a family of probability measures or a set of quantum states.
- A transformation Φ is said to be sufficient for the family (P_θ) if there exists a transformation Ψ such that

$$\Psi\left(\Phi\left(P_\theta\right)\right)=P_\theta.$$

- For probability measures the transformations should be given by Markov kernels.
- A divergence d satisfies the sufficiency condition if $d\left(\Phi\left(P_1\right),\Phi\left(P_2\right)\right)=d\left(P_1,P_2\right)$ when Φ is sufficient for P_1,P_2.
- f-divergences are the typical examples of divergences that satisfy the sufficiency condition.
- A Bregman divergence that satisfies sufficiency is proportional to information divergence (Jiao et al. 2014).

Locality

- A Bregman divergence on a convex set is said to local if the following condition is fulfilled.
- For any three states s_0, s_1 and s_2 such that s_1 is mutually singular with both s_1 and s_2 and for any $t \in [0, 1[$ we have that

$$d\left((1-t)\cdot s_0 + t\cdot s_1\right) = d\left((1-t)\cdot s_0 + t\cdot s_2\right).$$

- Sufficiency on a set of probability measures implies locality.

Locality (example)

- Sunny weater is predicted with probability p_0.
- Cloudy weater is predicted with probability p_1.
- Rain is predicted with probability p_2.
- The becomes sunny weather.
- The score should only depend on p_0 and not on p_1 and p_2.

Bregman divergence on spectral sets

Theorem

Let C denote a spectral convex set. If the entropy function has gradients parellel to convex hulls of embedded simplices, then the Bregman divergence generated by the (minus) entropy is local.

Proof.

Assume that $s = (1 - p)\, s_0 + p s_1$ where s_0 and s_1 are orthogonal. Then one can make orthogonal decompositions

$$s_0 = \sum p_{0i} \cdot s_{0i} \text{ and } s_1 = \sum p_{1j} \cdot s_{1j}$$

Then

$$d_H\left(s_0, s\right) = \sum p_{0i} \cdot \ln \frac{p_{0i}}{(1 - p)\, p_{0i}}$$

$$= \sum p_{0i} \cdot \ln \frac{1}{1 \cdots} = \ln \frac{1}{1 \cdots}$$

Entropic dimension 1

Theorem

Let C denote a spectral convex set where any state can be decomposed into two orthogonal states. Then the convex set is a balanced set without one dimensional faces and any Bregman divergence is local.

Theorem

Let C be a spectral convex set with at least three orthogonal states. If a Bregman divergence d defined on C is local then the Bregman divergence is generated by the entropy times some constant.

Proof Assume that the Bregman divergence is generated by the convex function $f : C \to \mathbb{R}$. Let K denote the convex hull of a set $s_0, s_1, \ldots s_n$ of singular states. For each s_i there exists a simple measurement ψ_i on C such that $\psi_i(s_j) = \delta_{i,j}$. For $Q \in K$ weak sufficiency implies that

$$d(s_i, Q) = d(s_i, \psi_i(Q)\, s_i + (1 - \psi_i(Q))\, s_{i+1}).$$

Proof cont.

Let f_i denote the function $f_i(x) = d(s_i, xs_i + (1-x)s_{i+1})$ so that $d(s_i, Q) = f_i(\psi_i(Q))$. Let $P = \sum p_i s_i$and $Q = \sum q_i P_i$. Then

$$\begin{aligned} d(P,Q) &= \sum p_i d(s_i, Q) - \sum p_i d(s_i, P) \\ &= \sum p_i f_i(q_i) - \sum p_i f_i(p_i) \end{aligned}$$

As a function of Q it has minimum when $Q = P$. Assume the f is differentiable.

$$\frac{\partial}{\partial q_i} d(P,Q) = p_i f_i'(q_i)$$

and

$$\frac{\partial}{\partial q_i} d(P,Q)_{|Q=P} = p_i \cdot f_i'(p_i).$$

Using Lagrange multipliers we get that there exist a constant c_K such that $p_i \cdot f_i'(p_i) = c_K$.

Proof cont.

Hence $f_i'(p_i) = \frac{c_K}{p_i}$ so that $f_i(p_i) = c_k \cdot \ln(p_i) + m_i$ for some constant m_i. Therefore

$$\begin{aligned} d(P,Q) &= \sum p_i \left(f_i(q_i) - f_i(p_i)\right) \\ &= \sum p_i \left((c_K \cdot \ln(q_i) + m_i) - (c_K \cdot \ln(p_i) + m_i)\right) \\ &= -c_K \cdot \sum p_i \ln \frac{p_i}{q_i} \\ &= -c_K \cdot d_H(P,Q) . \end{aligned}$$

Faces of entropic dimension 1

Theorem

Assume that a spectral set has entropic dimension at least 2 and has a local Bregman divergence. Then any face of entropic dimension 1 is isometric to a ball.

Proof.

The Bregman divergence restricted the the face is given by the entropy of the orthogonal decomposition. The gradient is only radial if the face is a ball. □

Some applications

- In portfolio theory we want to maximize the revenue. The corresponding Bregman divergence is local if and only if all portfolios are dominated by portfolios corresponding to gambling in the sense of Kelly.
- In thermodynamics the locality condition is satisfied near thermodynamic equilibrium and the amount of extracable energy equals

$$kT \cdot D\left(P \| P_{eq}\right)$$

where P_{eq} is the state of the corresponding equilibrium state.

Conclusion

- Caratheodory's theorem can be improved.
- Information divergence is the only local Bregman divergence on spectral set.
- Information theory only works for spectral sets.
- A complete classification of spectral sets is needed.