Introduction to Classical Information Theory

https://iuk.one/103-0002

Clemens H. **Cap**
ORCID: 0000-0003-3958-6136

Department of Computer Science
University of **Rostock**
Rostock,Germany
clemens.cap@uni-rostock.de

23

Overview (1/2)

Overview (2/2)

1. Motivation

Why do we want to study information theory?

Information and Physics

Norbert Wiener

Information is information not matter or energy. [?], p132

Carl-Friedrich von Weizsäcker

Jede Alternative von Möglichkeiten [...] läßt sich entscheiden
indem man sukzessive Ja/Nein Entscheidungen macht. [?], [?], [?]

Rolf Landauer

Information is Physical. [?]

John Archibald Wheeler

It from a bit: Every physical quantity, every it,
derives its ultimate significance from bits, binary yes-or-no indications. [?] [?]

David Deutsch

It from qubit. [?]

Attempts to Define Information

Information is a concept of resolving uncertainty. (bad: just another word)

Information as a **means for constructing objects** (will talk a bit on this)

- **Algorithmic information theory, complexity theory**
  Chaitin, Solomonov, Kolmogorov, Martin-Löf, Blum

Information as **choice of the actual among the potential** (will talk a lot on this)

- **Probabilistic information theory:** Wiener, Shannon, Nyquist, Hartley

Information as a human cognitive construct (will not talk about this)

- **Belief:** Calculus of human belief: Bayes, Pearl. [?], [?].
- **Frequentist:** Analysis of empirical outcomes. [?]
- **Propensity:** Tendency of favoring an outcome: Peirce, Popper. [?], [?].
- **Economy:** Readyness to engage in a bet. Ramsey [?], [?]

2. (Non-)Determinism

Information has something to do with **uncertainty**

- how to build something
- what to expect in the next experiment

Uncertainty is related to **non-determinism**.

What are these two concepts:

- determinism
- non-determinism

Related Concept: Determinism

Hypothesis of Determinism

We can describe the state of a system at a specific moment in time.
Given suitable initial conditions, we can predict the state in the future.

Problem:

- There is no concept of (global) time.
- Thus there is no concept of state.
- The definition of state and of determinism fails.

Debate on Determinism

Aspect 1: Physics: SRT & ART
Idea 1: Invent notions of local state and local determinism.
Idea 2: Glue local states together to an artificial event or spacetime manifold.

Aspect 2: Distributed computing
Aspect 2a: Computing is a subset of physics, so aspect 1 applies.
Aspect 2b: Even without this (i.e. computing in Newtonian space$\times$time) there is a problem.

- Set of nodes
- Communicate about their local states
- Communication incurs a delay (in contrast to physics we do not know how much)
- During delay remote state can change (and computation turns wrong)
- **Idea 1:** Causal models of distributed computation (aka Petri-nets)
- **Idea 2:** Virtually synchronous and virtually serialized computations
  Use models which (incorrectly) assume synchronous or serialized computation.
  Problem: Incorrect assumptions may cause incorrect results.
  If a shift in time does not change the computed result – the programmer does not care.
  Thus: Restrict model to computations that are equivalent in result to serialized computation.

Against Determinism

Arguments:

1. There is no global concept of time and thus of state (local state workarounds exist)
2. Measuring an object disturbs the object.
3. We cannot know the state of the measurement device
   and thus we cannot determine the disturbance produced by measurement.
4. Measured state is established only *after* the measurement.
5. The environment affects the measurement process (Zurek: einselection)
6. Most interpretations of QM postulate non-determinism (von Neumann measurement)
7. State and state change cannot *both* be determined at the same moment in time (Heisenberg)
8. State and state change cannot, each at a time, be precisely determined.

Epistemological Paradox:

1. We *never* can do the *same* experiment twice.
2. The second experiment always is different: We know the result of the first.
3. Determinism is not accessible to experimentation.
4. Determinism is not a reasonable notion in (at least: empirical) science.

Related Concept: Non-Determinism

Hypothesis of Non-Determinism and Disorder "Regellosigkeit"

There is no rule telling "nature" what to do next.

Laplacian Principle of Indifference:
What happens if *"there are no reasons"* to prefer a specific outcome over all possible outcomes?

Interpretations of *"there are no reasons"*:

1. **Practical** limit: We could know but will not: Universe is too complex.
2. **Systematic** limit: We cannot access the reasons: We are somehow limited.
3. **Conceptual** limit: Determinism is the wrong concept.

3. Where are the Difficulties?

Important differences between mathematical and physical models.

Einstein (Vortrag "Geometrie und Erfahrung", 27. 1. 1921, Preussische Akademie der Wissenschaften)

Insofern sich die Sätze der Mathematik auf die Wirklichkeit beziehen, sind sie nicht sicher, und insofern sie sicher sind, beziehen sie sich nicht auf die Wirklichkeit.

Physics and Mathematics

Physics: The experiment says different.

- Theory dismissed as wrong.
- Theory may remain as useful approximation. (eg: Thermodynamics, classical mechanics)

Mathematics: There is no experiment.

- What does this mean?
- Isn't mathematics restricted by the laws of logic?
- **No!**
- Mathematics is only restricted by the decisions of the designer of the mental model.

Question 1: Was god restricted by the laws of logic?

Question 2: Is logic empirical? [?], [?].

What is Logic?

Symbols (aka formulae) describe things in my **mind**.

Reasoning about things in my mind is replaced by operations on symbols. $x^2 \to 2x$

Mind: May have states `true`, `false` but also `unknown`, `unsure`, `not-determined`, `highly-probable`, `improbable` and more.

Important: `true` has no magic meaning, it just is an *(arbitrary)* state of mind the designer of the formalism *wants* to model (at least in modern logic).

Assume a framework for this as in $\phi, \vartheta, \ldots \vdash \gamma, \alpha, \ldots$

Sequence of formulae $\vdash$ **sequence** of formulae

$\vdash$ means **deduce**. Not necessarily connected with a notion of truth.

Could also be set, multiset, boolean algebra (classical logic), lattice (quantum logic!)

A First Example

$S \vdash W$ If (the **S**un shines) we can deduce that (it is **W**arm outside).

$S \vdash H$ If (the **S**un shines) we can deduce that (everybody is **H**appy).

$S \vdash W \wedge S$ If (the **S**un shines) we can deduce that (it is **W**arm outside) and (everybody is **H**appy).

Let us introduce the following rule into our logic:

$$\frac{\alpha \vdash \varphi \quad \alpha \vdash \psi}{\alpha \vdash \varphi \wedge \psi} \; (1)$$

A Second Example

$\$ \vdash W$ — If (I have one \$) we can deduce that (I can buy a glass of **W**hiskey).

$\$ \vdash H$ — If (I have one \$) we can deduce that (I can buy a **H**amburger).

Let us apply our rule:

$$\frac{\alpha \vdash \varphi \quad \alpha \vdash \psi}{\alpha \vdash \varphi \land \psi} \ (1)$$

$\$ \vdash W \land H$ — If (I have one \$) we can deduce that
(I can buy a glass of **W**hiskey) and (I can buy a **H**amburger).

I just love logic!

The Second Example Revisited

$\$ \vdash W$	If (I have one \$) we can deduce that (I can buy a glass of **W**hiskey).
$\$ \vdash H$	If (I have one \$) we can deduce that (I can buy a **H**amburger).
$\$ \wedge \$ \vdash W \wedge H$	If (I have one \$) and (I have one \$) we can deduce that (I can buy a glass of **W**hiskey) and (I can buy a **H**amburger).

We rather need a different rule:

$$\frac{\alpha \vdash \varphi \quad \beta \vdash \psi}{\alpha \wedge \beta \vdash \varphi \wedge \psi} \; (2) \qquad \text{The old rule was:} \frac{\alpha \vdash \varphi \quad \alpha \vdash \psi}{\alpha \vdash \varphi \wedge \psi} \; (1)$$

After some more analysis: We even need a different conjunction operator.

There are Several Brands of Propositional Logic

	Classical	Linear Logic	
		Multiplicative	Additive
Conjunction	$\wedge$	$\star$	$\sqcap$
Disjunction	$\vee$	$+$	$\sqcup$
True	T	1	$\top$
False	F	0	$\bot$
Implication	$\Rightarrow$	$\multimap$	$\multimap$
Negation	$\neg$	$\sim$	$\sim$

Overview

1. **Multiplicative** linear logic: Implication consumes resources.
2. **Additive** linear logic: No conservation of resources.
3. Classical propositional logic: Employs the conjunction $\wedge$

Compare:

Quantum mechanics: Measurment destroys an (assumed preexisting) status and generates an eigenvector as postmeasurement status.

Why Did We Do All This?

1. There is no generic truth and *no generic logic*.
2. We *always* have to check with the goals of our modeling domain.
3. Often, we see paradoxic consequences of modeling decisions only *much* later after the axiomatization.
4. The paradoxes do not point to peculiar properties of the studied objects but to *bad choices* of our axiomatization.

Application:

1. Wrong: "Information does not have certain properties."
2. Correct: "Our axiomatization of information has certain properties."

Here:

Which concept of information is the best description of our modeling domain.

4. Algorithmic Information Theory

Information as
means for constructing objects.

Problem Statement

Let A be a finite set, whose elements are called **symbols**.

Let $A^* := \{a_1 a_2 \ldots a_n \mid a_j \in A, n \in \mathbb{N}\} \cup \{\varepsilon\}$ be the **freely generated monoid**
i.e.: The set of (finite) strings together with the operation of concatenation.

$A^\infty := \{f : \mathbb{N} \to A \mid f \quad \text{function}\}$ is the set of infinite strings.

Question: How do we **want** to **define**
the **amount of information contained** in a **single** string $w \in A^*$ or $w \in A^* \cup A^\infty$?

1. It is a matter of **choice** (i.e.: a definition)
2. It is about a **single** string, not n strings or even $\lim_{n\to\infty}$ of n strings.

Example 1: Naïve Repetition

Let A be the set of ASCII symbols and w be the following word:

yyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy

Question: What are the *shortest* means of *describing or constructing* this?

```
print("yyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy");

for (var num=0; num < 80; num++) {print("y");}    // shorter program

for(var i=0;i<80;i++)print("y")                   // still shorter

i=80;while(i--)print("y")                         // even still shorter
```

Src. 1: Four programs for printing 80 copies of "y".

Example 2: More Advanced

Question: What are the *shortest* means of *describing or constructing* this:

```
,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{
```

```
print(",-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\\]^_\`
  abcdefghijklmnopqrstuvwxyz{");

for (var num=44; num <= 122; num++) {printChar(num);}

for (var n=44;n<=122;n++)printChar(n);
```

Src. 2: Two programs for printing a special ASCII string.

Example 3: Infinite Strings

`3.1415926535897932...`

Thoughts: This is π! How would I know? Maybe just first 20 digits?

And: What is π, after all?

Maybe: $\int_{-1}^{+1} \frac{1}{\sqrt{1-x^2}}\, dx$ But what is *that*?

Rather: A program, which prints out all decimal digits of π.

Note: This works for an infinite string only, if there is a program printing it.
This is *not* always the case.

Inconstructive Strings

Theorem: There are infinite strings for which there is no program, which prints them.

Proof: The programs printing a finite or infinite string can be ordered lexicographically.

Think of them as being written down as (countably infinite) sequence.

Imagine that the representations are replaced by the string they represent:

$$a_1(1)a_1(2)a_1(3)\dots$$
$$a_2(1)a_2(2)a_2(3)\dots$$
$$a_3(1)a_3(2)a_3(3)\dots$$

1. Pick a symbol different from $a_1(1)$ and call it b_1
2. Pick a symbol different from $a_2(2)$ and call it b_2
3. Pick a symbol different from $a_3(3)$ and call it b_3 ...

So there exists a string $b_1 b_2 b_3 \dots$ which is not in this list
and thus has no program printing it
and thus escapes every analysis by algorithmic information theory.

Information

Intuition: The information given by an object equals the complexity required for constructing this object.

Definition: The **information** given by a string is the length of a shortest program printing this string.

Definition: A string is called **compressible** iff there exists a program printing this string which is shorter than the string itself; otherwise it is called **random**.

Example: Naïvely: Things "such as" `aIz4TqWWeMn90-2KqLGr40iPF7D`.
Example: Strictly: Chaitin Ω and all Martin-Löf random numbers.

Chaitin Omega

Chaitin Ω:

- Use our lexicographic ordering of programs.
- Put a 0 if the program terminates.
- Put a 1 if not.
- Since the halting problem is not solvable, there is **no** algorithm printing out Ω.
- Hence there is no shortest length.
- Hence the minimum length is ∞.
- Hence we call this a truly random number.

Problems to Solve in Algorithmic Information Theory

Problem 1: We need some notion of construction.

- A Java program is fine.
- A definite integral is fine, provided we can numerically approximate its value.
- An arbitrary possibly "inconstructive" specification is **not** fine.

Problem 2: Different notions of construction concepts may lead to different lengths.

- One language has a concept of a `goto`.
- Another language has a concept of a `for loop`.
- Another language has a concept of recursion.

Problem 3: Different encoding alphabets

- Over $\{0, 1\}$ a program coding will be twice as long than over $\{a, b, c, d\}$.

Chaitin-Kolmogorov-Solomonoff Complexity (1)

Suppose: We know, what a computational concept is.

More precisely: A **computational concept** is a "mechanism", which

1. we "feed with" an element p of a language $\mathcal{L}$ ("program")
2. and a finite number of natural numbers ("input")
3. which then "stops" after a finite number of "steps" and "outputs" a string ("result")
4. or never stops ("infinite loop")
5. and which fulfills some technical conditions
   1. It provides a partial recursive function $\beta : \mathcal{L} \times \mathbb{N}^* \hookrightarrow \mathbb{N}$
   2. satisfies the **UTM** (Universal Turing Machine) property
   3. satisfies the **SMN** (Kleene parametrisation or partial evaluation) property

Even more precisely: Attend a 2 term-filling lecture series in theoretical computer science and/or read the texts [?], [?].

Chaitin-Kolmogorov-Solomonoff Complexity (2)

Let $\beta\colon \mathcal{L} \times \mathbb{N}^* \hookrightarrow \mathbb{N}$ be a computational concept.

The **Kolmogorov complexity** of a word[1] is the **length of the shortest program** which stops on the empty input and outputs the word w.

$$\gamma_\beta(w) := min(\{len(p) \mid p \in \mathcal{L},\ \beta(p,\varepsilon) = w\})$$

Problem: γ_β depends on the computational concept β.

Solution: The dependency is not very strong: [?, ?], [?], [?, ?].

The Kolmogorov complexities of two computational concepts β_1 and β_2 differ at most by an **additive constant** which holds uniformly for all words w:

$$\forall \beta_1, \beta_2\colon \exists C_{\beta_1,\beta_2} : \forall w : -C_{\beta_1,\beta_2} < \gamma_{\beta_1}(w) - \gamma_{\beta_2}(w) < C_{\beta_1,\beta_2}$$

[1] Natural numbers in some encoding.

Practical Problem

Theorem: Given a word w and a computational concept β, the Chaitin-Kolmogorov-Solomonoff complexity γ_β cannot be algorithmically determined.

Determining $\gamma_\beta(w)$ is one of the many not computable (more precisely: semi-computable) problems of computer science. [?]

Sad consequences:

- Despite its theoretical attractiveness it is **useless** for all **systematic practical** purposes.
- $\gamma_\beta(w)$ is known for only the most trivial examples so it is **useless** even for all **interesting practical** purposes.

5. Probabilistic Information Theory

Information as
choice of the actual among the potential.

What do we want to achieve?

Goal: Information as choice of the actual in the set of the potential.
We want to quantify the size of a set.

Ansatz 1: Intuition of **counting**, leads to the concept of **cardinality**.

Ansatz 2: Intuition of **contents**, leads to the concept of a **measure**.

Both approaches produce **interesting problems:**

- often ignored in *applications* (compare: Dirac δ-"*function*"/ *distribution*)
- deemed solvable by *theory* (compare: Schwartz distributions)
- point to **fascinating problems** in the non-set-theoretic *foundations* of mathematics

Categorial (topoi) foundations have recent applications in quantum physics [?, ?], [?], [?, ?, ?], [?], [?].

Concept of Cardinality

Two sets are said to be **equipotent**,
iff there exists a *bijective function* between them.
Nice and easy for the finite case.

1	2	3	4
$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$
a	b	c	d

Big problem with infinite sets:
A set may be *equipotent* to a true subset
even to its naïve *"half"*.

1	2	3	4	...
$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	
2	4	6	8	...

Even worse with the continuum:
$(-\infty, +\infty) = \mathbb{R}$, half-$\mathbb{R}$, i.e. $(-\infty, 0)$,
and arbitrarily "small" non-empty open intervals (a, b) **all are equipotent**.

Conclusion: Cardinalities are a bad approach
to model our intuition of *set size* and *information theory* in infinite sets.

Concept of Measure

Find all **functions** of all subsets of n-dim. space, $\mu\colon 2^{\mathbb{R}^n} \to [0, \infty]$, which satisfy:

(1) Scaling: Unit cubes have measure 1: $\mu([0,1]^n) = 1$
Empty set has measure zero: $\mu(\emptyset) = 0$

(2) Translation Invariance:

$$\forall A \subseteq \mathbb{R}^n,\ \vec{x} \in \mathbb{R}^n: \quad \mu(A + \vec{x}) = \mu(A)$$

(3) Rotation and Reflection Invariance:

$$\forall A \subseteq \mathbb{R}^n,\ f \in (S)O(n): \quad \mu(f(A)) = \mu(A)$$

(4) σ-Additivity: For every family $(A_j)_{j \in \mathbb{N}}$ of subsets which are pairwise non-overlapping (=disjoint), i.e. $i \neq j \Rightarrow A_i \cap A_j = \emptyset$ we have

$$\mu(\biguplus_{j\in J} A_j) = \sum_{j \in J} \mu(A_j)$$

Note: Summands non-negative, series absolute-convergent, *thus* sequence of summation irrelevant.

"No-Go Theorem" of Measure Theory

Theorem by Vitali: There are no such functions! [?].
The fundamental problem of measure theory cannot be solved.

Paradox of Banach-Tarski: [?], [?], [?].

The unit ball in $\mathbb{R}^3$, i.e. $\mathbb{B}_3 = \{\vec{x} \in \mathbb{R}^3 \mid \|\vec{x}\| = 1\}$ (with volume $4\pi/3$)

1. can be represented as union of 5 pairwise disjoint subsets $\mathbb{B}_3 = T_1 \uplus T_2 \uplus T_3 \uplus T_4 \uplus T_5$ with $i \neq j \Rightarrow T_i \cap T_j = \emptyset$,
2. onto which translations, rotations and reflections can be applied
3. such that the union of the resulting sets are a unit ball of **twice** the radius $\{\vec{x} \mid \|\vec{x}\| = 2\}$ (and **eight** times the volume).

This is in **fundamental contradiction** with our intuition of a volume!

Explanation and Solution

Explanation for Vitali:
There are sets which are not measurable in any reasonable sense.

Explanation for Banach-Tarski:

- Partition a measurable set into several non-measurable sets.
- Work on those using translations, rotations and reflections.
- Union is a measurable set of twice the volume.
- **Blow-up** happens "under the radar" on sets which are not measurable.

The set $\mathbb{R}^3$ of triples of real numbers does **not** reflect our intuition of content.
It is merely a vague approximation thereof! We need...

1. **Additional** structures: Topologies, measures, distances
2. **Restriction** of concepts: Borel σ-algebras, measurability; continuity

"Repairing" Measure Theory

Attempt 1: Remove set theory axioms allowing proof of Banach-Tarski paradox.

1. *Powerset* Axiom: Cannot remove, needed for higher order constructions.
2. *Infinity* Axiom: Cannot remove, needed for construction of natural numbers.
3. *Choice* Axiom: Removes inconstructive results, leads to intuitionistic logic.

Only choice: Remove axiom of choice.

But: Produces unpleasant mathematics and still is said to allow some variants of the Banach-Tarski paradoxon, according to[?].

Attempt 2: Restrict notion of a measurable set.
Only some subsets will be considered measurable. $\mu\colon \mathcal{A} \to [0, \infty]$ with $\mathcal{A} \subsetneq 2^{\mathbb{R}^n}$

Definition: Measurable Space

A **measurable space** is a pair $(\Omega, \mathcal{A})$ consisting of a set Ω and a set $\mathcal{A} \subseteq 2^{\Omega}$ of subsets of Ω. The elements of $\mathcal{A}$ are called $\mathcal{A}$-**measurable** sets.

The following must hold:

1. $\mathcal{A}$ contains the *set* Ω *itself*.
2. $\mathcal{A}$ is *closed under set-complement*: $\forall A \in \mathcal{A} \colon \complement A \in \mathcal{A}$
3. $\mathcal{A}$ is *closed under countable union*: $\forall (A_j \in \mathcal{A})_{j \in \mathbb{N}} \colon \bigcup_{j \in \mathbb{N}} \in \mathcal{A}$

A **measure space** is a triple $(\Omega, \mathcal{A}, \mu)$ consisting of a measurable space $(\Omega, \mathcal{A})$ and a σ-additive function $\mu \colon \mathcal{A} \to [0, +\infty] = \mathbb{R}_0^+ \cup \{+\infty\}$.

Core idea: σ-additivity is not required for all subsets of Ω but only for the measurable subsets of Ω.

Easy Examples: Finite and Countable Infinite Case

Finite case:

Note: The base set Ω is finite, not necessarily the measure!

$$\Omega = \{a_1, a_2, \ldots, a_n\} \qquad \mathcal{A} = 2^\Omega \qquad \mu(\{b_1, b_2, \ldots b_k\}) = \sum_{j=1}^{k} \mu(\{b_j\})$$

Countably infinite case:

$$\Omega = \{a_1, a_2, \ldots\} \qquad \mathcal{A} = 2^\Omega \qquad \mu(\{b_1, b_2, \ldots\}) = \sum_{j=1}^{\infty} \mu(\{b_j\})$$

In both examples:

1. all singleton sets $\{a\}$ are measurable, so μ is defined on singletons.
2. the values of μ on the singletons uniquely define all values of μ on $\mathcal{A}$.

Advanced Example: The Continuum Case

Let $\Omega = \mathbb{R}$

Let $\mathcal{A}$ be the smallest subset of $2^{\mathbb{R}}$ which contains all open intervals (a, b) and which is closed under countable union, countable intersection and set complement. (**Borel sets**).

Define μ on **intervals**: $\mu((a, b)) = b - a$.

Further results of measure theory "look good": [?], [?], [?].

- $\mathcal{A}$ is well-defined ("smallest") and μ can be uniquely extended from intervals to $\mathcal{A}$.
- The no-go theorem of Vitali does not hold any more.
- The Banach-Tarski paradox is no longer paradoxical.

  The measure μ is not defined on all 5 partitioning sets. The congruence transformations are applied to sets which are not measurable. We have no expectation of keeping a measure constant when transforming a set for which no measure exists.
- Can be extended to $\mathbb{R}^n$ using "cubes" and to topological spaces.
- Concepts of density functions may be introduced.

6. Shannon Information Theory

6.1. Probability

6.2. Conditional Probability

6.3. Information

Probabilistic Information Theory
which is based on Measure Theory.

Probability

Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means.

Bertrand Russell as cited in [?].

Finite Measures and Probability Spaces

The measure μ of a measure space $(\Omega, \mathcal{A}, \mu)$ is called **finite**,
iff the measure only has finite values: $\mu\colon \mathcal{A} \to [0, +\infty) \subsetneq [0, +\infty]$.

A **probability space** is a **measure space** $\mathcal{P} = (\Omega, \mathcal{A}, P)$ with $P(\emptyset) = 0$ and $P(\Omega) = 1$.

The measure of $\mathcal{P}$ is called a **probability measure**.

Prop: If $(\Omega, \mathcal{A}, \mu)$ is a measure space with finite measure, then $(\Omega, \mathcal{A}, P)$ with

$$P(X) := \frac{\mu(X)}{\mu(\Omega)}$$

is a probability space.

Example and Counter Example

Consider: $\Omega = [0,5]$ $\quad \mu([a,b]) = b - a$ $\quad \mu(\Omega) = 5$ as measure space.

Obtain: $P([a,b] = \frac{b-a}{5}$ as probability space: **Equi-distribution** on $[0,5]$.

Density: $\varphi(x) = \frac{1}{5}$

Distribution: $P([a,b]) = \int_a^b \varphi(x)dx = \Phi(b) - \Phi(a)$ $\qquad \Phi(x) = \int_0^x \varphi(x)dx$

Modify: $\Omega = \mathbb{R}$ $\quad \mu([a,b]) = b - a$

Problem! No longer finite: $\mu(\Omega) = \mu(\mathbb{R}) = \infty$.

Norming: $P(X) = \frac{\mu(X)}{\mu(\Omega)} = \frac{\mu(X)}{\infty}$

Finite intervals have measure zero: $P([a,b]) = \frac{b-a}{\infty} = 0$

Infinite sets have indefinite measure: $P(X) = \frac{\mu(X)}{\infty} = \frac{\infty}{\infty} =$

Definition: Conditional Probability

Idea 1: Only consider events where the validity of a set B of properties is ensured.

Idea 2: Renormalize probability to still sum up to 1 *despite* smaller summation domain.

Let $(\Omega, \mathcal{A}, P)$ be a probability space.

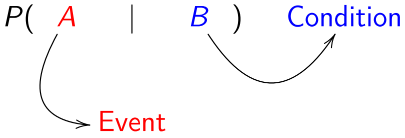
Let $B \in \mathcal{A}$ with $P(B) \neq 0$.

The **conditional probability under the condition** B is the function

$$\begin{aligned} P_{|B} = P(\,\cdot \mid B): \quad \mathcal{A} &\to [0,1] \\ A &\mapsto P_{|B}(A) = P(A \mid B) \end{aligned}$$

with

$$P(A \mid B) := \frac{P(A \cap B)}{P(B)}$$



Properties of Conditional Probability

Define the **pointwise intersection** of a σ-algebra: $\mathcal{A} \cap B := \{X \cap B \mid X \in \mathcal{A}\}$

(1) The conditional probability $p_{|B} : \mathcal{A} \to [0,1]$ is a **probability measure** on $(\Omega, \mathcal{A})$.
Proof obligation: Show that it sums up to 1.

(2) The conditional probability $p_{|B} : \mathcal{A} \to [0,1]$ induces a probability measure **on** $(B, \mathcal{A} \cap B)$.
Proof obligation: Show proper set of base sets.

p:	$\mathcal{A}$	$\to$	$[0,1]$	original probability measure		
$p_{\mid B}$:	$\mathcal{A}$	$\to$	$[0,1]$	modified measure		(1)
$p_{\mid B}$:	$\mathcal{A} \cap B$	$\to$	$[0,1]$	modified measure and algebra	$\mathcal{A} \cap B \xhookrightarrow{id} \mathcal{A} \xrightarrow{p_{\mid B}} [0,1]$	(2)

Notation of Conditional Probability

Probability is a thing $p(\cdot)$ where we can fill in sets of all kinds, A, $A \cap B$, and more.

The conventional notation of **conditional probability** breaks this.
We write $p(A|B)$ although there is no suitable set $A|B$.

Better notation: $p_{|B}$ where we can plug in set A: $p(A|B) = p_{|B}(A)$.

Theorem: Classical Bayes Rule and Bayes Chain Rule

Classical Bayes Rule: Swapping event and condition

$$P(A \mid B) = \frac{P(A)}{P(B)} P(B \mid A)$$

holds for A, B with $P(A), P(B) \neq 0$

$$\frac{P(B \mid A)}{P(B)} = \frac{P(A \mid B)}{P(A)} = \frac{P(A \cap B)}{P(A) \cdot P(B)}$$

Classical Bayes Rule, written differently

$$P(A \cap B) = P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A)$$

Bayes Chain Rule

$$P(A \cap B \cap C) = P(A \mid B \cap C) \cdot P(B \cap C) = P(A \mid B \cap C) \cdot P(B \mid C) \cdot P(C)$$

Iterated chain

Preparation: Splitting Rule

An event may be split on a single condition B

Logic: $A \Leftrightarrow (A \wedge B) \vee (A \wedge \neg B)$

Sets: $A = (A \cap B) \uplus (A \cap \complement B)$

$$\begin{aligned}
A &= A \cap (A \cup \complement B) \\
&= A \cap [(A \cup \complement B) \cap \Omega] \\
&= A \cap [(A \cup \complement B) \cap (B \cup \complement B)] \\
&= [(A \cap B) \cup A] \cap [(A \cup \complement B) \cap (B \cup \complement B)] \\
&= [(A \cap B) \cup A] \cap [(A \cap B]) \cup \complement B] && \text{now: distributive law} \\
&= (A \cap B) \cup (A \cap \complement B) && \text{even: disjoint sum} \\
&= (A \cap B) \uplus (A \cap \complement B)
\end{aligned}$$

Thus: $P(A) = P[(A \cap B) \uplus (A \cap \complement B)] = P(A \cap B) + P(A \cap \complement B)$

Now: Apply Bayes Chain Rule twice.

Special Case: Bayes Splitting Rule

Binary case: Assume: $P(B), P(\complement B) \neq 0$.

$$P(A) = P(B)P(A \mid B) + P(\complement B)P(A \mid \complement B)$$

General case: Assume: $X_1, X_2, \ldots, X_n$ is a partition of Ω with $\forall i : P(X_i) > 0$.

$$\forall X \in \mathcal{A} : P(X) = \sum_{i=1}^{n} P(X_i)P(X \mid X_i)$$

$$\forall X \in \mathcal{A}, P(X) > 0 : P(X_i \mid X) = \frac{P(X_i)P(X \mid X_i)}{\sum_{i=1}^{n} P(X_i)P(X \mid X_i)}$$

Splitting Rule and Double Slit Experiment (1)

$$P(A)=P(B)P(A \mid B)+P(\complement B)P(A \mid \complement B)$$

Experiment produces **black curve** $P(A)$.

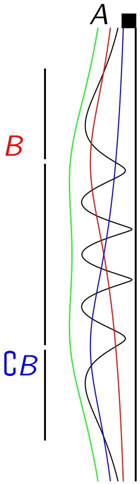


Fig. 1: Double Slit Experiment

Splitting Rule and Double Slit Experiment (2)

Nice: Splitting works in classical propositional logic (which is distributive).

Nice: Splitting works in set theory (which is distributive).

Cave: Splitting does not work in quantum mechanics – but why?

Reasons why nature behaves differently than theory suggests are *speculations*!

Nature does not meet one of our implicit assumptions leading to $P(A) = P(A)$.

1. **Particle assumption:** Electron does not pass through either B xor $\complement B$.
2. **Experiment:** Measurement of green = red + blue does not make sense. These are two different experiments, the addition of whose values does not correspond to a single physical experiment.
3. **Counterfactual definiteness:** Cannot assume that properties we did not really measure have a definite value. (Eg: Theoretizing on the value red could have while actually measuring blue.)
4. **Distributivity:** Quantum logic is not distributive but needs an *orthomodular* law. [?]

Definition and Proposition: Independence

Definition: Two events $X, Y \in \mathcal{A}$ of a probability space $(\Omega, \mathcal{A}, P)$ are called **independent**, iff their "probabilities multiply"; more formally iff:

$$P(X \cap Y) = P(X) \cdot P(Y)$$

Proposition: In case the respective conditional probabilities exist:
Two events X and Y are independent, if and only if
conditioning one event by the other *does not change* its probability.

$$P(X|Y) = P(X) \qquad P(Y|X) = P(Y)$$

Proof: Directly from the definition of conditional probability.

This criterion gives a *better intuitive understanding* of independence.
This criterion provide a *worse formal definition*, as it is less general.
(Since it only holds in cases where conditional probabilities exist).

Definition: Information

The **information content** I of a probability space $\mathcal{P} = (\Omega, \mathcal{A}, P)$ is the function

$$I : \mathcal{A} \to [0, +\infty] \quad \text{with} \quad I(A) := -\log_r(\ P(A)\ )$$

r	Name of unit
2	bit
e	nat
10	Hartley

Tab. 1: Units for measuring information content.

Core consequence: Information content of *independent* events is *additive*:

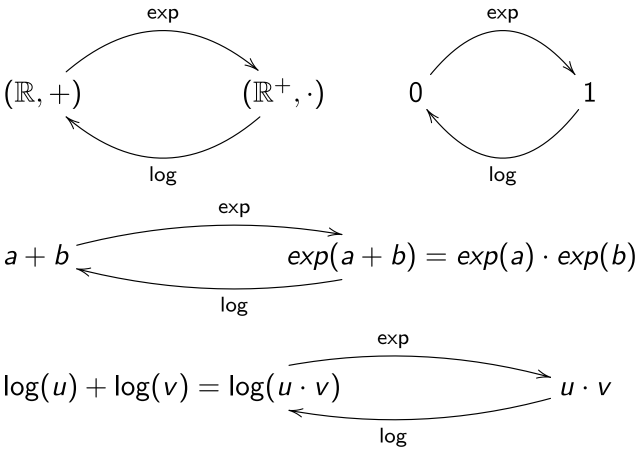
$$P(X \cap Y) = P(X) \cdot P(Y) \Rightarrow I(X \cap Y) = I(X) + I(Y)$$

Information and Probability

From an algebraic point of view information and probability are **isomorphic** (i.e. identical).

Similarly, for a slide-rule, adding and multiplying is just a matter of (logarithmic) scales.

With regard to **independence:**
Independent probability *multiplies*.
Independent information *adds*.



7. Information Sources

7.1. Basic Definitions

7.2. Entropy and Redundancy

7.3. Examples

7.4. Convexity

Describing where information comes from.

Intuition: Finite Memoryless Information Sources

Finite: From a finite number of different (digital) symbols *one* symbol is provided.

Extending probability from elements (singleton sets) to sets is trivial σ-additivity:

- Start with a function $\pi\colon A \to [0,1]$ for *symbol* probability
- Extend to $p\colon 2^A \to [0,1]$ with $p(X) := \sum_{\xi \in X} \pi(\xi)$ for *set* probability

We *could* also consider countably infinite or uncountable sets (analogue signals). Then, continuity, convergence and σ-algebras become important (technical) issues.

Memoryless: Assume a repetition of experiments and

1. probability is time-independent $\Rightarrow$ can model by one value
2. repeated experiments are pairwise independent $\Rightarrow$ probabilities multiply
3. in repeated experiments, relative symbol frequency converges to probability

Note: 3 is **not** guaranteed but a seriously restricting assumption. Law of large numbers holds only "almost surely" or in adapted notions of convergence and under (strong) conditions of independence, which cannot naturally be assumed to hold in nature. Examples see [?] and [?].

Definition: Finite Memoryless Information Sources

A finite, memoryless **information source** is a pair $\mathcal{S} = (A, p)$ consisting of

1. a finite set A, whose elements are called symbols
2. a probability measure $p\colon 2^A \to [0, 1]$

Notation: Often $p(a)$ is used for $p(\{a\})$.

Random Variables, Expectation Values and Conditions

A **random variable** is a finite, memoryless information source (A, p) together with a function $f\colon A \to \mathbb{R}$.

The **expectation value** of a random function $((A, p), f)$ is defined as the sum of the values weighted by the respective probabilities

$$\mathcal{E}_{(A,p)}(f) := \sum_{a \in A} p(a) \cdot f(a)$$

The **conditional expectation value** of random function $((A, p), f)$ (under a condition $B \subseteq A$) is the *expectation value* of f under the *conditional probability* (of said condition B).

$$\mathcal{E}_{(A,p)}(f) = \mathcal{E}_{|B}(f) = \sum_{a \in A} p(a|B) \cdot f(a) = \sum_{a \in A} \frac{p(\{a\} \cap B)}{p(B)} \cdot f(a) = \underbrace{\sum_{a \in B}}_{\text{Note different summation domain!}} \frac{p(\{a\})}{p(B)} \cdot f(a)$$

Dice as Information Source – A Beginners Toy Example (1)

$$\mathcal{Q} = (A, p) \qquad p\colon A \to [0,1] \qquad f\colon A \to \mathbb{R}$$

$$A = \{\text{⚀}, \text{⚁}, \text{⚂}, \text{⚃}, \text{⚄}, \text{⚅}\} \qquad (\text{⚀}, \text{⚁}, \text{⚂}, \text{⚃}, \text{⚄}, \text{⚅}) \overset{p}{\mapsto} \left(\tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}\right)$$

$$(\text{⚀}, \text{⚁}, \text{⚂}, \text{⚃}, \text{⚄}, \text{⚅}) \overset{f}{\mapsto} (1,2,3,4,5,6) \qquad \mathcal{E}_{\mathcal{Q}}(f) = \mathcal{E}_{(A,p)}(f) = \vec{f} \cdot \vec{p} = \textstyle\sum_{j=1}^{6} \frac{j}{6} = \frac{7}{2}$$

$$\textbf{Even} := \{\text{⚁}, \text{⚃}, \text{⚅}\} \qquad p(\textbf{Even}) = 1/2$$

$$p_{|\text{ Even}}(\ \{\text{⚀}\}\ ) = p(\ \{\text{⚀}\} \mid \textbf{Even}) = \frac{p(\ \{\text{⚀}\} \cap \textbf{Even}\ )}{p(\textbf{Even})} = \frac{p(\emptyset)}{\frac{1}{2}} = 0$$

$$p_{|\text{ Even}}(\ \{\text{⚁}\}\ ) = p(\ \{\text{⚁}\} \mid \textbf{Even}) = \frac{p(\ \{\text{⚁}\} \cap \textbf{Even}\ )}{p(\textbf{Even})} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$

Dice as Information Source – A Beginners Toy Example (2)

$$\mathcal{E}_{A,p_{|\mathbf{Even}}}(f) = \sum_{a \in A} p_{|\mathbf{Even}}(\{a\}) \cdot f(a) =$$

$$p_{|\mathbf{Even}}(\{\text{⚀}\}) \cdot f(\text{⚀}) + p_{|\mathbf{Even}}(\{\text{⚁}\}) \cdot f(\text{⚁}) + p_{|\mathbf{Even}}(\{\text{⚂}\}) \cdot f(\text{⚂}) +$$

$$p_{|\mathbf{Even}}(\{\text{⚃}\}) \cdot f(\text{⚃}) + p_{|\mathbf{Even}}(\{\text{⚄}\}) \cdot f(\text{⚄}) + p_{|\mathbf{Even}}(\{\text{⚅}\}) \cdot f(\text{⚅})$$

$$= \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 2 + \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 4 + \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 6$$

approach 1: summing over entire set with conditional probabilities

$$= \frac{1}{3} \cdot 2 + \frac{1}{3} \cdot 4 + \frac{1}{3} \cdot 6$$

approach 2: summing only over conditioned set

$$= 4$$

Small Remark

Why do I emphasize this difference so much, pointing it out with two different colors?

We can take two perspectives of conditioning:

1. Keep the original set but modify the summation.
2. Reduce the set and sum over the entire (new) set.

and the color choice points out these two perspectives.

These are **two different mathematical objects**.

They provide identical results in most cases (such as probabilities or expectations).

But there are subtle aspects which may go wrong

- when defining conditional entropy — important for us
- when dealing with cases where we need σ-algebras — not important for us

Definition: Entropy

The **entropy** $H(\mathcal{S})$ of a source $\mathcal{S} = (A, p)$ is the **expectation value of the information content**, i.e. the average information content of a symbol.

$$H(\mathcal{S}) = \mathcal{E}_{p;\forall a \in A}\left(I(a) \right) = \sum_{a \in A} p(a) \cdot I(a) = -\sum_{a \in A} p(a) \cdot \log_2(p(a))$$

Theorem: Maximal Entropy

The **maximal value** of the entropy of a source with n symbols is

$$H_{max}(n) := \log_2(n)$$

Of all sources with n symbols the (unique) source of **maximal entropy**, is the source, for which **all symbols are equally probable:** $\forall a \in A\colon\ p(a) = 1/n$.

Informally: The higher the variance, the smaller the entropy.

1. Higher variance means: Individual symbols have *higher information content* (due to their smaller probability).
2. But: These symbols also have *smaller probability* of occurring.
3. Thus: The effect of the smaller probability in the expectation value sum is stronger than the effect of having a higher information content.

Definition: Redundancy: How far below what is possible?

The **redundancy** of a source $\mathcal{Q}$ is its *deficit* to the maximally possible entropy:

$$R(\mathcal{Q}) := H_{max}(\mathcal{Q}) - H(\mathcal{Q})$$

The **relative redundancy** of a source $\mathcal{Q}$ is its *redundancy after linear scaling* to the domain $[0, 1]$:

$$r(\mathcal{Q}) := 1 - \frac{H(\mathcal{Q})}{H_{max}(\mathcal{Q})}$$

Interpretation: The redundancy measures how far a source stays under its possibilities of information generation.

Example: Binary Sources

Consider **all binary** sources.

Base set: $A = \{0, 1\}$. **One parameter:**
$P(0) =: q$.
Thus $P(1) = 1 - P(0) = (1 - q)$.

The binary sources form a 1-parameter object with parameter $q \in [0, 1]$.

Entropy is
$H(q) = -q \log_2(q) - (1 - q) \log_2(1 - q)$.

At $q = P(0) = P(1) = 1/2$
we get maximal entropy
Its *value:* $H_{max}(2) = \log_2(2) = 1$.

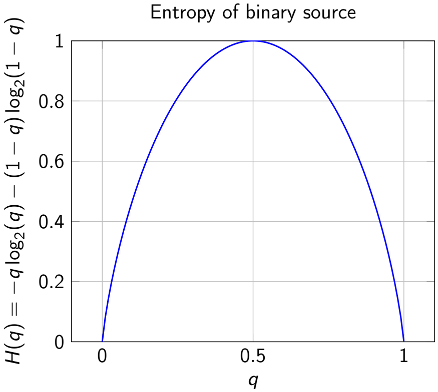


Fig. 2: Entropy of binary source as 1-parameter object.

Example: Ternary Sources: Parametrization

Consider **all ternary** sources.

A ternary source is a 2-parameter object, defined over a planar triangular domain in $\mathbb{R}^3$
$\{(x, y, z) \mid 0 \leq x, y, z \leq 1 \ \wedge \ x + y + z = 1\}$

One possible parametrization:
Base set: $A = \{0, 1, 2\}$
1. param: $x := P(0) \in [0, 1]$
2. param: $y := P(1) \in [0, 1]$

Thus: $P(2) = (1 - P(0) - P(1)) \in [0, 1]$.

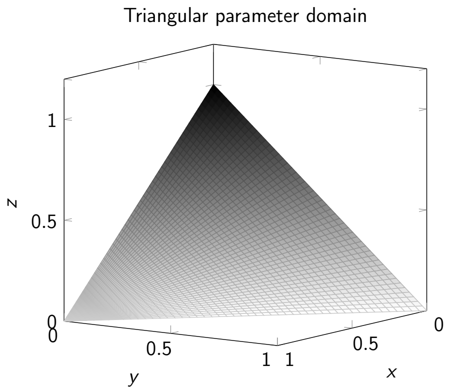


Fig. 3: Twodimensional triangular parameter domain of ternary sources as a plane in three-dimensional space.

Example: Ternary Sources: x-y Coordinates

Looking on triangular domain from above.
Using x and y as parameters.

We see a distortion due to the slant projection π_z on the parameter space.

Entropy is $H(x,y) = -x\log_2(x) - y\log_2(y) - (1-x-y)\log_2(1-x-y)$

Maximal entropy at $x = y = z = 1/3$ has *value* $H_{max}(3) = log_2(3) = 1.585\ldots$

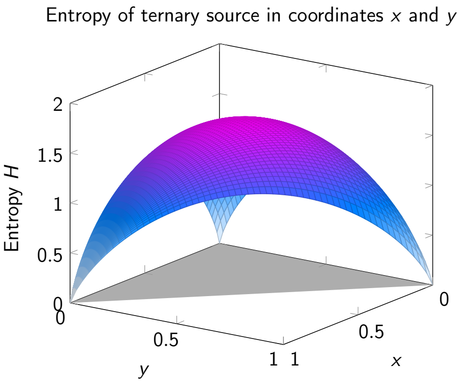


Fig. 4: Entropy of ternary source, x-y coordinates.

Example: Ternary Sources: Orthogonal Projection

Looking on triangular domain via orthogonal projection.

We see an equilateral triangle since the orthogonal projection incurs no distortion.

Note the **concave shape** of the entropy function.

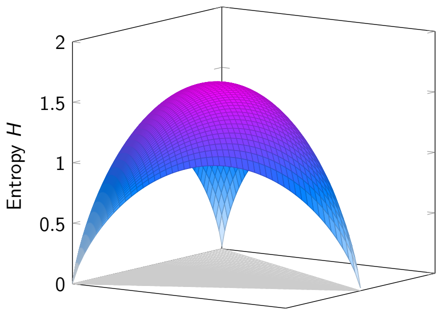


Fig. 5: Entropy of ternary source, orthogonal projection.

Example: Ternary Source as Convex Object

Observations:

1. The three corners are the extremals.
2. Their convex hull is the state space.
3. Entropy is maximal in an inner point.
4. Negentropy is maximal in the extremals.

Interpretations:

1. **High negentropy** means **high** degree of **order**.
2. **High entropy** means **high** degree of **disorder** and thus **information content**.

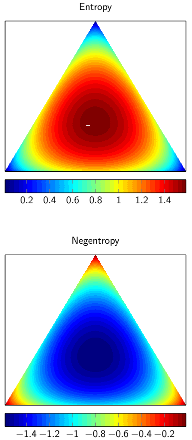


Fig. 6: Entropy and negentropy of ternary source as 2-parameter object without projective distortion.

Example: Recoding Ternary Sources (1)

Let $A = \{a, b, c\}$ represent a ternary information source.

Goal: We want to represent this source over a binary alphabet.

Goal 2: If possible, we want to recode in a more efficient way.

We try below recoding:

Symbol	Prob	Recode
a	x	00
b	y	10
c	$1 - x - y$	11

Observe: The average length of a code word is $2x + 2y + 2(1 - x - y-) = 2$.

Question: Can we do better?

Answer: Except in the case $x = y = 1/3$

Definition: Prefix-Free Coding

Definition: A coding is called **prefix-free**, iff no element of the set of codewords is a prefix of a codeword.

Proposition: A coding which is prefix-free allows a unique decoding.

Example: The coding $a \mapsto 0, \quad b \mapsto 10, \quad c \mapsto 11$ with its codeword set $\{0, 10, 11\}$ is prefix-free.

Observation: This allows a unique left-to-right linear decoding:

Example: 0001110 decodes as aaacb

Counterex: If we would encode a as 1 then 11 could decode as c or as aa.

Example: Recoding Ternary Sources (2)

Idea: Consider the following prefix-free coding:

Symbol	Prob	Recode
a	x	0
b	y	10
c	$1 - x - y$	11

Observation:

- The average length of a code word is $1x + 2y + 2(1 - x - y) = 2 - x$.
- For all cases except $x = 0$ (one-digit case is never used) this is a more efficient coding.

Convex Sets

A subset $S \subseteq V$ of a vector space V with scalars $\mathbb{K} \supset \mathbb{R}$ is called **convex**, iff for all points $\vec{x}, \vec{y}$ in S the *open line segment* $\mathcal{O}(\vec{x}, \vec{y})$ is in the set S.

$$\mathcal{O}(\vec{x}, \vec{y}) := \{\lambda\vec{x} + (1-\lambda)\vec{y} \mid \lambda \in (0,1)\}$$

This obviously equivalent definition will soon become important:

$$\mathcal{O}(\vec{x}, \vec{y}) := \{p_1\vec{x} + p_2\vec{y} \mid p_1, p_2 \geq 0 \wedge p_1 + p_2 = 1\}$$

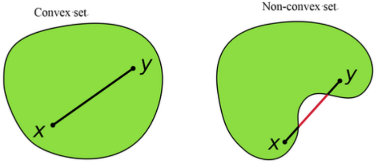


Fig. 7: Convex and non-convex set.

The concept of "*concave* = not-convex" for sets is occasionally found, but **not useful** as it produces misunderstanding.

Convex Notions

A point of a convex set S is called **extreme**, iff it is not element of an *open line segment* between two points of the set S.

The **convex hull** $< S >_c$ of a subset S of a vector space with scalars $\mathbb{K} \supset \mathbb{R}$ is the set $< S >_c := \{\lambda\vec{x} + (1-\lambda)\vec{y} \mid \vec{x}, \vec{y} \in S, \lambda \in [0,1]\}$

Two further, equivalent definitions:

1. The smallest convex superset of S.
2. The intersection of all convex supersets of S.

Convex sets are important for us due to:

- **Jensen inequality** of classical information theory.
- **Pure versus mixed states** in quantum information theory.
- **Krein-Milman Theorem:** Convex sets are (often) the *convex hull of their extreme points*.
  Thus: In math, we only need to know the extremes of convex sets.
  Thus: In physics, we only need to study pure states.
- Quantum-useful results in functional analysis (Hahn-Banach Theorem).

Convex Functions

A function f is called

- **convex** iff its *epigraph* is convex.
- **concave** iff its negative $-f$ is convex.

Classify: (1) Convex, (2) concave and (3) **others**.

Convex and concave are **dual** to each other.
Concave = not-convex is **simply wrong**.

Convex functions defined over convex sets have **important extremal** properties:

- Maxima are on the boundaries of the convex set.
- A local minimum is also a global minimum.

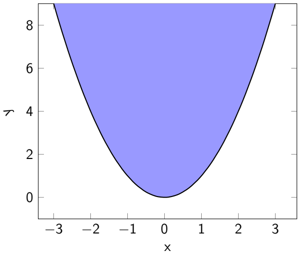


Fig. 8: The **epigraph** of a function consists of the graph and all points "above": $\text{epi}(f) := \{(x, y) \mid x \in \text{dom}(f) \wedge y \geq f(x)\}$. Obviously, this function **is convex**.

Convexity Rephrased

By definition: f is convex, iff the epigraph is convex.

By the alternative definition of the line segment this is equivalent to:

Whenever $p_1 + p_2 = 1$ for $p_i \geq 0$ then

$$p_1 \cdot f(x_1) + p_2 \cdot f(x_2) \geq f(p_1 \cdot x_1 + p_2 \cdot x_2)$$

Question: Can this be generalized? Maybe to:

$$\sum_i p_i \cdot f(x_i) \quad \geq \quad f\left(\sum_i p_i \cdot x_i\right)$$

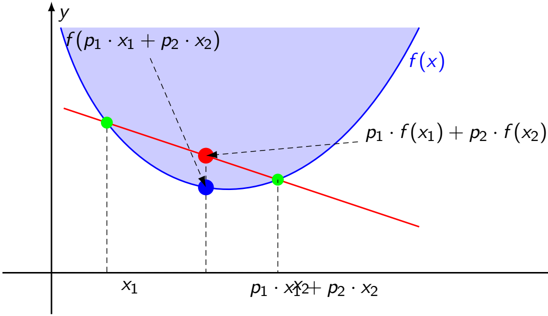


Fig. 9: Convex function and inequalities: The red dot is above the blue dot. As f is convex the epigraph (above the blue line) is convex. Thus the points on the red line between the two green dots are in the epigraph. Thus the red dot in the epigraph is above the blue dot on its boundary.

Theorem: Jensen Inequality

When f is convex, then for $p_i \geq 0$ with $\sum p_i = 1$ the **Jensen inequality** holds:

$$\sum_i p_i \cdot f(x_i) \quad \geq \quad f\left(\sum_i p_i \cdot x_i\right)$$

Note: $p_i \geq 0$ and $\sum_i p_i = 1$ **is** *exactly* probability theory.

Jensen can be interpreted as an inequality on expectation values:

$$\mathcal{E}(f(X)) \quad \geq \quad f(\mathcal{E}(X))$$

Convexity of Information Sources

A vector is called **stochastic**, iff its entries are in $[0,1]$ and their sum is 1.

n-**ary information sources** $\{a_1, \ldots, a_n\}, P$ may be (bijectively) represented by stochastic *n*-vectors $(P(a_1), P(a_2), \ldots, P(a_n))$ with $P(a_i) \geq 0$ and $\sum_i P(a_i) = 1$.

Let $\mathfrak{I} \subseteq \mathbb{R}^n$ be the set of all *n*-ary information source stochastic vectors in $\mathbb{R}^n$.

- $\mathfrak{I}$ is **convex** and an $(n-1)$-dimensional **simplex** in $\mathbb{R}^n$.
- The **entropy** function on $\mathfrak{I}$ is **concave**.
- The **negentropy**, the *neg*ative *entropy*, is a **convex** function on $\mathfrak{I}$. Negentropy is defined in physics for describing order by [**?**], [**?**].
- The negentropy is **maximal at the extremals** of $\mathfrak{I}$ and has a **local minimum** in the interior, which is **global**.
- The entropy is **minimal at the extremals** of $\mathfrak{I}$ and has a local maximum in the interior, which is **global**.
- $\mathfrak{I}$ is the **convex hull** of its corners: Knowing the corners means knowing the set.

Probability Theories as Geometries

Classical probability is (pretty much exactly) real convex geometry.

Quantum probability is complex *non-commutative* geometry.

Idea is:

1. Start with a geometric space S.
2. Define complex-valued functions $f : S \to \mathbb{C}$ and operations between them.
3. Think of operator algebras – oh, this looks like algebras of observable functions.
4. Remember that there is a $C\star$ algebra approach to measurements.
5. Fall in love with these non-commutative algebras and forget the geometric space S.
6. Can we recover geometric structures when studying only this algebra?
7. Yes! We do geometry without points, only checking function algebras.
8. Similar stuff known by the ironic name of *pointless topology*.

Conceptual Similarities of Theories

Classical Information Theory

1. **Pure** states (strings of length 1): Only the elements of $A = \{a, b, c\}$
2. **Mixed** states: (Formal) convex hull of A: Elements $\vec{x} = \alpha \cdot a + \beta \cdot b + \gamma \cdot c$.
3. **Real, positive** coefficients: $\alpha, \beta, \gamma \in \mathbb{R}_0$
4. **Normalize:** May divide by $\alpha + \beta + \gamma$ or assume this is one.
5. **Norming constraint:** $\langle \vec{1}, \vec{x} \rangle = \alpha + \beta + \gamma = 1$ is linear
6. **Orthogonality:** $\vec{a} = 1 \cdot a + 0 \cdot b + 0 \cdot c$ and $\vec{b}, \vec{c}$ form a (real) orthonormal basis.
7. **Base:** Only this base, no other bases, no base changes.

Quantum Information Theory

1. **Pure** states: Every element $\alpha \cdot a + \beta \cdot b + \gamma \cdot c \in \mathbf{span}_{\mathbb{C}}(A)$
2. **Mixed** states: (Formal) convex hull of projectors: Density operator.
3. **Complex** coefficients: $\alpha, \beta, \gamma \in \mathbb{C}$
4. **Normalize:** May divide by $\sqrt{\bar{\alpha}\alpha + \bar{\beta}\beta + \bar{\gamma}\gamma}$
5. **Invariance:** Global phase plays no role.
6. **Symmetry:** $U(3)$
7. **Norming constraint:** $\langle \vec{x}, \vec{x} \rangle_{\mathbb{C}} = \bar{\alpha} \cdot \alpha + \bar{\beta} \cdot \beta + \bar{\gamma} \cdot \gamma = 1$ is sesquilinear.
8. **Orthogonality:** $\vec{a}, \vec{b}, \vec{c}$ form a (complex) orthonormal basis.
9. **Bases:** Arbitrary base changes via $U(3)$.

Fundamental Differences in Theories

State:

- **Classical:** Does not consider $0.3 \cdot a + 0.7 \cdot b$ a state or string or character. Represents merely an abstract, stochastically mixed information source.
- **Quantum:** Arbitrary complex superpositions. $(1/\sqrt{2}) \cdot a + (i/\sqrt{2}) \cdot b$ is a physical state. Is **not a stochastic mixture** but a (pure) state.

Bases:

- **Classical:** Only one base: The elements of A are singled out.
- **Quantum:** All bases are created equal.

Superposition:

- **Classical:** Not existent.
- **Quantum:** Every state is a superposition in ∞-many ways

Quantum has two significantly different concepts of state combination.

- **Superposition:** Phase difference allows interference phenomena.
- **Mixture:** Similar as in classical theory.

8. Products and Compounds

Information and interaction &
Preparation for classical channel theory.

Intuition behind Products and Compounds

Situation: Two finite, memoryless information sources $\mathcal{S}_A = (A, \alpha)$ and $\mathcal{S}_B = (B, \beta)$

Goal: We want to study pairs of results: $(a, b) \in A \times B$.
We want to study sequences of results: $a_1 a_2 a_3 \ldots \in A^n \subseteq A^*$

Products: Symbol set is Cartesian product, *measure is direct product*.

- Information sources S_A and S_B considered independent.
- In this case we know: Probabilities multiply.

Compounds: Symbol set is Cartesian product, *measure is arbitrary*.

- Study arbitrary probabilities which happen to exist on the product set.
- Study how these probabilities deviate from the independence assumption.
- Proper setting to analyze **probabilistic dependencies** or correlations.

Why is this interesting? (1)

Note: Probabilistic dependency is different from causal dependency.

Science: *Observes* probabilistic dependencies and *searches* for causal explanation.

Example: Water the roof of your house to make it rain.

W The roof of my house is wet.
R It rains.

	W	$\neg W$
R	100	0
$\neg R$	0	200

Possible Explanations of Correlations:

1. **Causality:** (a) $R \Rightarrow_{\text{causes}} W$ xor (b) $W \Rightarrow_{\text{causes}} R$.
2. **Common Cause:** $C \Rightarrow_{\text{causes}} R$ and $C \Rightarrow_{\text{causes}} W$.
3. **Coincidence:** There is no "reason". Possible but unlikely. Need test statistics. Spurious correlations always exist in large data corpses.
4. **Mixtures:** Combination of 1, 2, 3.

Question: How can we distinguish these three cases?

Why is this interesting? (2)

Experiment:	Does an intervention on one variable change the other variable? Can I make it rain by watering the roof of my house?
Research:	Coincidence is a highly unsatisfactory explanation! Find a common cause!
Einstein:	Effects must be in the light cone of the cause. Properties are localized in time-space manifold.
Schrödinger:	Entanglement allows non-localized properties.
Bell:	Events may be correlated better than permitted by local causality mechanisms.
Aspect:	This really happens in nature.
Problem:	How can we explain correlations of space-like separated events A and B?
Idea:	The explanation is consequence of a non-localized property.

Definition: Product Source

The **product** of the finite, memoryless information sources $\mathcal{S}_A = (A, \alpha)$ and $\mathcal{S}_B = (B, \beta)$ is the information source $\mathcal{S}_A \times \mathcal{S}_B := (A \times B, p)$

where the measure $p = \alpha \otimes \beta$ on the product set is defined as follows:

1. $\alpha \otimes \beta$ is first **defined on singletons** (a_i, b_j) by $(\alpha \otimes \beta)(a, b) := \alpha(a) \cdot \beta(b)$.
2. and then **extended to sets** of singletons by σ-additivity.

Tensor notation $\otimes$:

- Initially does not indicate vector spaces but corresponds to set and category theory.
- Many formal connections to properties of the linear tensor theory!

Concept:

- Easy in the finite case: E.g.:
  $p\ (\{\ (a_2, b_3),\ (a_8, b_6)\ \}) = p(\{(a_2, b_3)\}) + p(\{(a_8, b_6)\}) = \alpha(a_2)\beta(b_3) + \alpha(a_8)\beta(b_6)$
- Much more complex in the infinite cases (for discrete and continuous scenarios).
  Need to work with σ-algebras.

Example: Product Source

$$A := \{a_1, \ldots, a_n\} \qquad B := \{b_1, \ldots, b_m\} \qquad \alpha_i := \alpha(\{a_i\}) \qquad \beta_j := \beta(\{b_j\})$$

$$p_{ij} = p\ (\ \{\ (a_i, b_j)\ \}\ ) = \alpha_i \cdot \beta_j$$

using product yields independence

$$\begin{pmatrix} \alpha_1\beta_1 & \alpha_1\beta_2 & \cdots & \alpha_1\beta_m \\ \alpha_2\beta_1 & \alpha_2\beta_2 & \cdots & \alpha_2\beta_m \\ \vdots & & & \\ \alpha_n\beta_1 & \alpha_n\beta_2 & \cdots & \alpha_n\beta_m \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \cdots \\ \alpha_n \end{pmatrix} \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_m \end{pmatrix} = \vec{\alpha} \otimes \vec{\beta}$$

Definition: Compound Source

A (binary) **compound source** is a source of the form $\mathcal{S} = (A \times B, p)$, i.e. a source where the set of symbols is a product of two sets A and B.

$$A := \{a_1, \dots, a_n\} \qquad B := \{b_1, \dots, b_m\} \qquad p_{ij} := p(\ \{(a_i, b_j)\}\ ) = p(a_i, b_j)$$

Questions:

- Can we understand a compound source as a product source?
- Can we approximate a compound source by a product source?
- Tools for analyzing the probabilistic dependencies:
  Joint, marginal and conditional probabilities.

Example: Compound Source with Joints and Marginals

$A := \{a_1, a_2, a_3\} \qquad B : = \{b_1, b_2, b_3\} \qquad p_{ij} = p(\{(a_i, b_j\}) = p(a_i, b_j)$

$$\begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \end{array} \begin{array}{c} \begin{array}{ccc} b_1 & b_2 & b_3 \end{array} \\ \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} \end{array}$$

$$p_{1\bullet} = p_{11} + p_{12} + p_{13} = p_A(a_1) = p(\{\ (a_1, b_1),\ (a_1, b_2),\ (a_1, b_3)\ \})$$
$$p_{2\bullet} = p_{21} + p_{22} + p_{23} = p_A(a_2) = p(\{\ (a_2, b_1),\ (a_2, b_2),\ (a_2, b_3)\ \})$$
$$p_{3\bullet} = p_{31} + p_{32} + p_{33} = p_A(a_3) = p(\{\ (a_3, b_1),\ (a_3, b_2),\ (a_3, b_3)\ \})$$

$$p_{\bullet 1} = p_{11} + p_{21} + p_{31} = p_B(b_1) = p(\{\ (a_1, b_1),\ (a_2, b_1),\ (a_3, b_1)\ \})$$
$$p_{\bullet 2} = p_{12} + p_{22} + p_{32} = p_B(b_2) = p(\{\ (a_1, b_2),\ (a_2, b_2),\ (a_3, b_2)\ \})$$
$$p_{\bullet 3} = p_{13} + p_{23} + p_{33} = p_B(b_3) = p(\{\ (a_1, b_3),\ (a_2, b_3),\ (a_3, b_3)\ \})$$

1

Black: **Joint** probabilities $p_{ij} \qquad p\colon A \times B \to [0, 1]$

Blue: **Marginal** probabilities $p_A\colon A \to [0, 1] \qquad p_B\colon B \to [0, 1]$

Defined by *summing up to the matrix margin*

Definition: Marginals

Let $p\colon A \times B \to [0,1]$ be a compound with A and B *finite*.

$$p_A\colon A \to [0,1] \qquad p_A(a) := \sum_{b \in B} p(a,b)$$

$$p_B\colon B \to [0,1] \qquad p_B(b) := \sum_{a \in A} p(a,b)$$

Note: Generalizes in straight-forward manner to finite products $p\colon A_1 \times \ldots \times A_n \to [0,1]$.

Notations: Abusive Conventions for Marginals

Error: We define a 2-variable function $p(a, b)$ and then write $p(a)$.

Abusive conventions:
$p(a)$ used instead of $p_A(a) = p(\ \{a\} \times B\ )$
$p(b)$ used instead of $p_B(b) = p(\ A \times \{b\}\ )$

Problem: What is $p(\xi)$ for a variable or value ξ? 👎

Set notation does not hide complexity, buys clarity at the expense of more brackets 👍.
It is always unambiguous. 👍
As in $p(\ \{a_1\} \times B\ )$ or $p(\ \{\sigma\} \times B \mid A \times \{\lambda\}\ )$.

Explicit notation for marginals provides correct typing in the index.
As in $p_A(a_1)$ or $p_B(\xi)$ 👍

Abusive convention breaks the substitution principle of Leibniz,
poses unnecessary issues for systems such as Mathematica,
destroys notational clarity and prevents reasoning by strict formula manipulation.

Notation: Special Conditionals for Compounds

Shorthand notation:

$$p(a \mid b) := p(\{a\} \times B \mid A \times \{b\})$$
$$p(a, b) := p(\{(a, b)\})$$
$$p(b) := p_B(\{b\})$$

By definition: $p(X \mid Y) = \dfrac{p(X \cap Y)}{p(Y)}$

Special conditionals in extensive notation:

$$p(\{a\} \times B \mid A \times \{b\}) = \frac{p((\{a\} \times B) \cap (A \times \{b\}))}{p(A \times \{b\})} = \frac{p(\{(a, b)\})}{p_B(\{b\})}$$

Special conditionals in **shorthand notation:**

$$p(a \mid b) = \frac{p(a, b)}{p(b)}$$

Same syntax as for single source completely *different semantics*.

Problem: What is $p(\xi|\eta)$ for concrete values ξ and η 👎
Problem: What is $p(\gamma|\gamma)$ for a concrete value γ which happens to be an element of A and of B 👎

Conditionals and Marginals

Conditionals from Joints and Marginals:

$$p(a|b) = \frac{p(a,b)}{p_B(b)} = \frac{p(a,b)}{\sum_{a\in A} p(a,b)} \qquad p(b|a) = \frac{p(a,b)}{p_A(a)} = \frac{p(a,b)}{\sum_{b\in B} p(a,b)}$$

Marginals from Conditionals via Chain-Rules:

$$p_A(a) = \sum_{b\in B} p(a|b) p_B(b) \qquad p_B(b) = \sum_{a\in A} p(b|a) p_A(a)$$

Joints recovered from Conditionals and Marginals:

$$p(a,b) = p(a|b)\cdot p_B(b) \qquad p(a,b) = p(b|a)\cdot p_A(a)$$

Why is that so?

While this looks intuitively obvious, with all the issues in $p(a|b)$ versus $p(b|a)$ notations we want to check this more formally using set notation at least in one example:

$$p(a,b) = \text{ go to set notation}$$

$$= p(\ \{(a,b)\}\ )$$

$$= p\Big(\ (\ \{a\} \times B\ )\ \cap\ (\ A \times \{b\}\ )\ \Big) =$$

use definition of conditional $p\big(X \cap Y\big) = p\big(X \mid Y\big) \cdot p\big(Y\big)$

$$= p\Big(\{a\} \times B \mid A \times \{b\}\Big) \cdot p\Big(A \times \{b\}\Big) = \text{ go back to "abusive" notation}$$

$$= p(a|b) \cdot p_B(b)$$

Technical Problems with Marginals

Problem 1: A compound is rather $p\colon 2^{A\times B} \to [0,1]$ where $U \subseteq A \times B$ and $p(U) = \sum_{u\in U} p(\{u\})$.

Problem 2: With A or B not finite, the $\sum$ is not so easy to define.

Problem 3: A compound is rather $p\colon \mathcal{S} \to [0,1]$ where $\mathcal{S} \subseteq A \times B$ is a σ-algebra.

Good News:

1. We only need the easy case.
2. All other problems can be solved nicely.
3. Even extension to compounds with an infinite number of components. Think of $\times_{\lambda\in R} A_r$ instead of $A \times B$.

Alternative Definition 1: Marginals as Compositions

Marginals are compositions:

$$p_A := p \circ \pi_A^{-1}$$

$$\begin{array}{ccc} A \times B & \xrightarrow{\pi_A} & A \\ (a,b) & \longmapsto & a \end{array} \qquad \begin{array}{ccc} A & \xrightarrow{\pi_A^{-1}} & 2^{A\times B} \\ a & \longmapsto & (\{a\} \times B) \end{array} \qquad \begin{array}{ccc} 2^A & \xrightarrow{\pi_A^{-1}} & 2^{A\times B} \\ U & \longmapsto & (U \times B) \end{array}$$

$$\begin{array}{ccccc} 2^A & \xrightarrow{\pi_A^{-1}} & 2^{A\times B} & \xrightarrow{p} & [0,1] \\ U & \longmapsto & U \times B & \longmapsto & p(U \times B) \end{array} \qquad 2^A \xrightarrow{p \circ \pi_A^{-1}} [0,1]$$

$$\underbrace{p \circ \pi_A^{-1}(\{a\})}_{\text{New def}} = p(\{a\} \times B) = \sum_{b \in B} p(\{(a,b)\}) = \sum_{b \in B} p(a,b) = \underbrace{p_A(\{a\})}_{\text{Old def.}}$$

Better definition – holds in arbitrary situations.

Note: We did not provide nor check proper σ-algebra conditions.

Expectation Values: Extension to Vector Values

We recall:

For $q\colon B \to [0,1]$ and $f\colon B \to \mathbb{R}$ we can define an expectation value:

$$\mathcal{E}_q(f) := \sum_{b \in B} q(b) \cdot f(b) \in \mathbb{R}$$

This may be generalized from $\mathbb{R}$ to arbitrary real vector spaces V.

Generalization:

For $q\colon B \to [0,1]$ and $f\colon B \to V$ we can define an expectation value:

$$\mathcal{E}_q(f) := \sum_{b \in B} q(b) \cdot f(b) \in V$$

It represents the average vector in V with weights / probabilities given by q.

Reinterpreting: Partial Conditionals as Vectors

We consider:

$$p(\cdot \mid \cdot) \colon \quad \begin{array}{ccc} A \times B & \to & [0,1] \\ (a,b) & \mapsto & p(a \mid b) \end{array}$$

Can be seen as *vector-valued* function of the *second variable*,
We supply the second variable and leave the first variable open.

Currying of the function:

$$p(\cdot_2 \mid \cdot_1) \colon \quad \begin{array}{ccc} B & \to & [A \to [0,1]] \\ b & \mapsto & p(\cdot \mid b) \colon \begin{array}{ccc} A & \to & [0,1] \\ a & \mapsto & p(a \mid b) \end{array} \end{array}$$

Observation: For fixed $b \in B$ function $p(\cdot \mid b) \colon A \to [0,1]$ is the vector $p(\cdot \mid b)$ of probabilities as given by $p(a_1 \mid b), p(a_2 \mid b), \ldots, p(a_n \mid b)$.

Alternative Definition 2: Marginals as Expect. of Conditionals

The marginal p_A is the vectorial expectation value of all vectors $p(\cdot \mid b)$.

Similar to all the $p(\cdot \mid b)$ also p_A is a vector in the sense of $A \to [0, 1]$.

Show $p_A = \mathcal{E}_{p(b)}(p(\cdot \mid b))$

We know: $p_A(a) = \sum_{b \in B} p(a|b) p_B(b)$

Products, Compounds and Factorization

Every product source **is a compound** source.

A **compound** source **can be factored into a product** of two sources,
if and only if the probability matrix of the compound source has **rank 1**.

Example: Left side shows rank 1, right side shows product factoring.

$$\begin{pmatrix} \alpha_1\beta_1 & \alpha_1\beta_2 & \alpha_1\beta_3 \\ \alpha_2\beta_1 & \alpha_2\beta_2 & \alpha_2\beta_3 \\ \alpha_3\beta_1 & \alpha_3\beta_2 & \alpha_3\beta_3 \end{pmatrix} = \left(\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} \cdot \beta_1 \quad \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} \cdot \beta_2 \quad \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} \cdot \beta_3 \right) \sim \vec{\alpha} \otimes \vec{\beta}$$

Generic: Compound sources generically have full rank.

Degenerate: Product sources are the highly degenerate case of rank 1.

Factorizables versus Compounds in Information Theory

Products: We **know** product structure; probability **is** factored.
Compounds: We **know** product structure; probability may be interdependent.

$A = \{\text{red}, \text{blue}\}$ $\quad B = \{\text{small}, \text{large}\}$
$A \times B = \{\ (\text{red}, \text{small}),\ (\text{red}, \text{large}),\ (\text{blue}, \text{small}),\ (\text{blue}, \text{large})\ \}$

Product: Probability depends only on color and size.
Compound: There is an interdependence between color and size.
Example: red is more often large than blue.

Question 1: Given a compound $(A \times B, p)$, can it be written as $(A, \alpha) \otimes (B, \beta)$?
Question 2: Given a source (X, p), can it be written as $(A, \alpha) \otimes (B, \beta)$?

Example: {a, b, c, d} (bad example, as it indicates a specific factorization)
Example: $\{a, e, i, u\}$ (better example)

Factorization

Will be part of the exercises / seminar.

Factoring

Factoring **compounds**:	*Only* a matter of **linear dimension and rank**
Factoring **sources**:	*Also* a matter of **partitioning** (much higher complexity!)
If **not factorizable**:	How close is it to a factorizable source?

We can define convex combinations (or sums) of sources:

Let $\mathcal{A}_1, \ldots, \mathcal{A}_n$ be information sources and $q_1 + \ldots + q_n = 1$ with $q_j \geq 0$.
The weighted sum or convex combination $\sum q_j \mathcal{A}_j$ works as follows:

1. With probability q_j select source $\mathcal{A}_j$.
2. Then use this source to select a symbol of this source.

Can I describe every source as a convex combination of factorizable sources? How?
When symbol sets overlap: Direct sum or various forms of "interference".

These are just random thoughts to show that some concepts of quantum information can be reformulated in classical language – despite the **big** conceptual differences in some aspects.

Factorizables versus Compounds in Physics

Note: Quantum physics has new state-space concepts.
Combine two quantum systems with state spaces A and B.
Resulting state space is not $A \times B$ but the much larger $A \otimes B$.
Need **superposition** and for the latter **Hilbert spaces** to describe this.

From space to entangled states:
Assume two spin 1/2 systems with projective state-space $\mathcal{Q} = \mathbb{C}^2/_\sim$.
State space of the compound is $\mathcal{Q} \otimes \mathcal{Q}$.
Strong correlation across space-separated system boundaries (Bell, CHSH).

Reverse question: ⚠ Can we go back from entangled states to space?
Given a holistic system, which subsystem aspects can we factor out?
How do we know the number of subsystems? And whether they are spatially separated.
What kind of separation / spatial / location properties do we find?
Is that necessarily what we plugged in (space-separation, 2x spin 1/2)
Compare: [?], [?], [?].

Bell-Type Experiment: Setup

State Base: Let $(\vec{u}, \vec{d})$ be an ON basis of $\mathbb{C}^2$.

Bell State: Let $\psi := (\ \vec{u} \otimes \vec{d} - \vec{d} \otimes \vec{u}\ )/\sqrt{2}$.

Measurement Base:: Let $(\vec{a}_1, \vec{a}_2)$, $(\vec{b}_1, \vec{b}_2)$ be two ON bases of $\mathbb{C}^2$.

2 Observables: Let $A := |\vec{a}_1\rangle\langle\vec{a}_1| - |\vec{a}_2\rangle\langle\vec{a}_2|$ $\qquad B := |\vec{b}_1\rangle\langle\vec{b}_1| - |\vec{b}_2\rangle\langle\vec{b}_2|$

Experiment: Measure $A \otimes B$ at ψ.

1. Operators commute: $A \otimes B = (A \otimes I)(I \otimes B) = (I \otimes B)(A \otimes I)$.
2. Sequential measurement: Arbitrary sequence of $A \otimes I$ and $I \otimes B$.
3. Parallel measurement: Measure $A \otimes I$ and $I \otimes B$ at space-like separated events.

Possible Results: $\vec{a}_1 \otimes \vec{b}_1$, $\vec{a}_1 \otimes \vec{b}_2$, $\vec{a}_2 \otimes \vec{b}_1$, $\vec{a}_2 \otimes \vec{b}_2$

Bell-Type Experiment: Results

The experiment yields the following probabilities:

θ is a parameter which is the angle between the real, 3-dimensional Bloch vectors belonging to A and B.

	b_1	b_2	
a_1	$\frac{1}{2}\sin^2\frac{\theta}{2}$	$\frac{1}{2}\cos^2\frac{\theta}{2}$	$\frac{1}{2}$
a_2	$\frac{1}{2}\cos^2\frac{\theta}{2}$	$\frac{1}{2}\sin^2\frac{\theta}{2}$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$	1

Tab. 2: Compound and marginal probabilitiesof the "Bell" compound source.

Special Parameter Choices

$\theta = 0$ — perfect anticorrelation

	b_1	b_2	
a_1	0	$\frac{1}{2}$	$\frac{1}{2}$
a_2	$\frac{1}{2}$	0	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$	1

$\theta = \pi/4$ — half way to center

maximal Bell violation

	b_1	b_2	
a_1	$\frac{2-\sqrt{2}}{8}$	$\frac{2+\sqrt{2}}{8}$	$\frac{1}{2}$
a_2	$\frac{2+\sqrt{2}}{8}$	$\frac{2-\sqrt{2}}{8}$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$	1

$\theta = \pi/2$ — zero coupling

in the "middle"

	b_1	b_2	
a_1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
a_2	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$	1

$\theta = \pi$ — perfect correlation

	b_1	b_2	
a_1	$\frac{1}{2}$	0	$\frac{1}{2}$
a_2	0	$\frac{1}{2}$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$	1

Tab. 3: Joint and marginal probabilities of the "Bell" compound source at particular values of θ.

Note 1: Every matrix is *symmetric* along main- & anti-diagonal. We only look at (a_1, b_1) and (a_2, b_1).
Note 2: Marginals are independent of θ and symmetric (always $1/2$)
θ only influences the **"inner" correlation**!

Marginals (Using Graphs)

Observations:

- Marginals are constant 0.5, independent of θ.
- Probabilities (0.5) and information content (1.0 [bit]) connected to each other as expected.
- Symmetries as expected.
- Pretty boring.

Marginal Probabilities and Marginal Information Contents

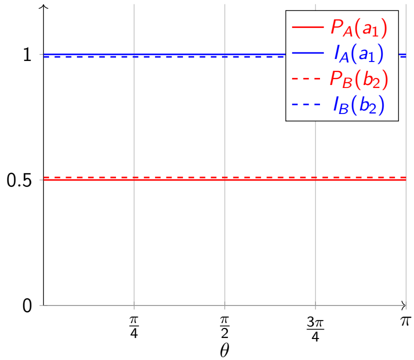


Fig. 10: Marginal probabilities (red) and marginal information contents (blue) of the "Bell" compound source are independent of the parameter θ.

Marginals (Using Formalism)

	b_1	b_2	
a_1	$\begin{bmatrix} 0 & \theta=0 \\ 1/4 & \theta=\pi/2 \\ 1/2 & \theta=\pi \end{bmatrix} = \frac{1}{2}\sin^2\frac{\theta}{2}$	$\begin{bmatrix} 1/2 & \theta=0 \\ 1/4 & \theta=\pi/2 \\ 0 & \theta=\pi \end{bmatrix} = \frac{1}{2}\cos^2\frac{\theta}{2}$	$\frac{1}{2}$
a_2	$\begin{bmatrix} 1/2 & \theta=0 \\ 1/4 & \theta=\pi/2 \\ 0 & \theta=\pi \end{bmatrix} = \frac{1}{2}\cos^2\frac{\theta}{2}$	$\frac{1}{2}\sin^2\frac{\theta}{2}$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$	1

Observation (a_1, b_1) tells us that

1. **Marginal** A: a_1 is there. $P_A(a_1) = 1/2$. Provides 1 bit at all θ. *Boring.*
2. **Marginal** B: b_1 is there. $P_B(b_1) = 1/2$. Provides 1 bit at all θ. *Boring.*
3. **Joint:** a_1 and b_1 are there. $P(a_1, b_1) = \sin^2(\theta/2)/2$.
   Interesting dependency on θ, which we want to study further.

Joints (Using Graphs, Only Probabilities)

Observations:

- Highly dependent on θ.
- The other two pairs look identical.
- How does information content look like?

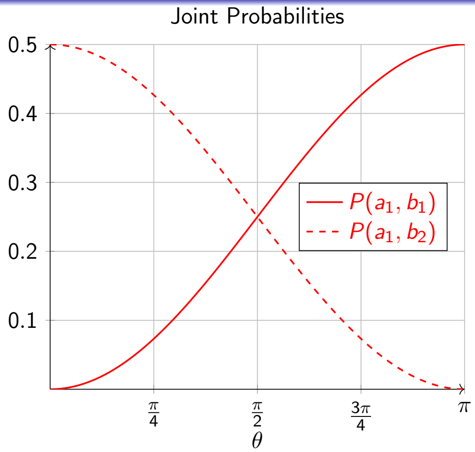


Fig. 11: Joint probabilities (red). Dashed versions shows a different pair.

Joints (Using Graphs)

Observations:

- *Low probability* leads to *high information* content.
- Logarithm produces non-linear stretching.
- *Singularity:* Information content $+\infty$ when probability is zero.

Joint Probabilities and Joint Information Contents

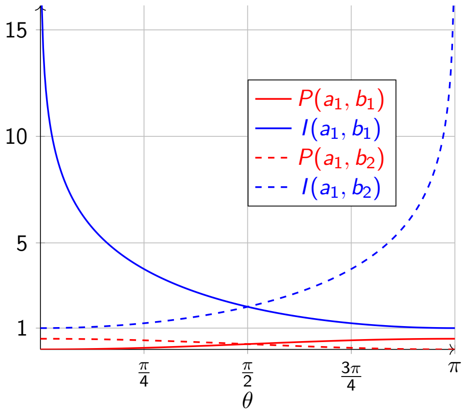


Fig. 12: Joint probabilities (red) and joint information contents (blue) of the "Bell" compound source. Dashed versions show a different pair.

Analyzing the Singularity

At $\theta = 0$ we have

- probability 0
- information content ∞

How does this affect entropy
as average information content?

$0 \cdot \infty$ is problematic.

de l'Hopital shows: $\lim_{h \to +0} h \cdot \log_2(h) = 0$

Thus: Singularity is no problem.
Contribution to entropy is zero.

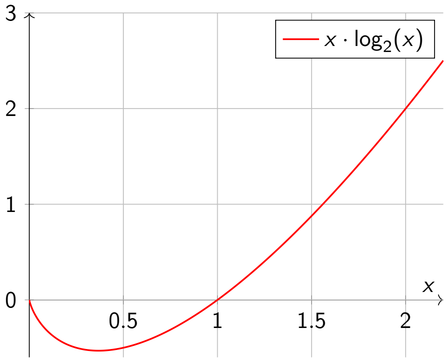
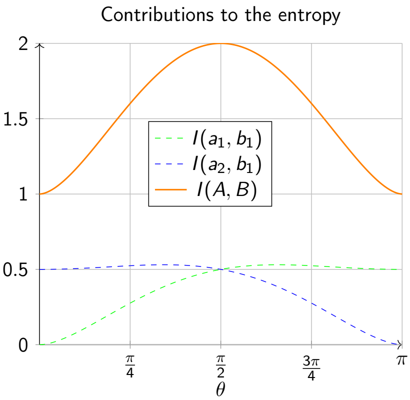


Fig. 13: Additive contribution of a symbol to the entropy.

Total Contributions of Pairs to Entropy



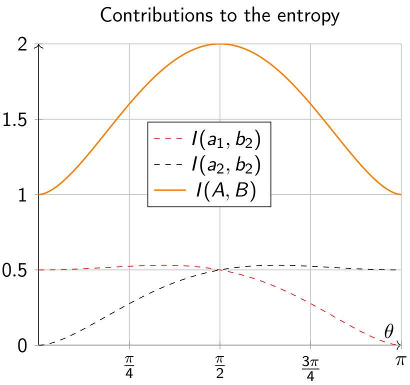


Fig. 14: Contributions of the four pairs (a_1, b_1), (a_1, b_2), (a_2, b_1) and (a_2, b_2) to the to the total entropy of the source.

Relative Contributions of Pairs to Entropy

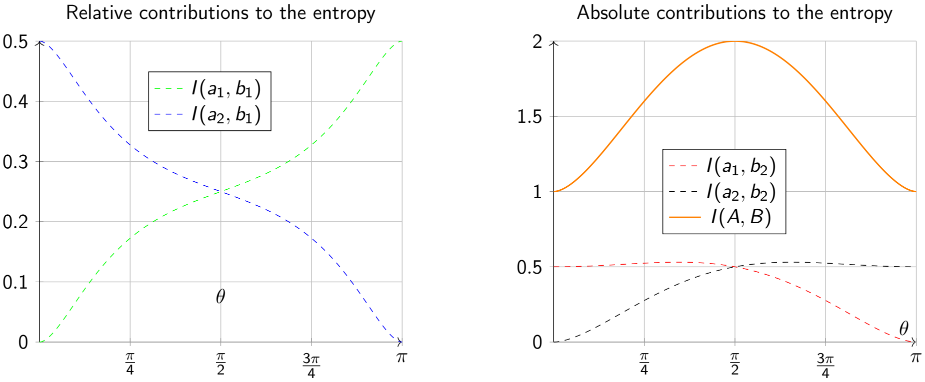


Fig. 15: Absolute and relative contributions of the pairs to the total entropy of the source.

Example: "Bell" Compound: Symbol Pairs: Fresh Look

	b_1	b_2	
a_1	$\begin{bmatrix} 0 & \theta = 0 \\ 1/4 & \theta = \pi/2 \\ 1/2 & \theta = \pi \end{bmatrix} = \frac{1}{2}\sin^2\frac{\theta}{2}$	$\begin{bmatrix} 1/2 & \theta = 0 \\ 1/4 & \theta = \pi/2 \\ 0 & \theta = \pi \end{bmatrix} = \frac{1}{2}\cos^2\frac{\theta}{2}$	$\frac{1}{2}$
a_2	$\begin{bmatrix} 1/2 & \theta = 0 \\ 1/4 & \theta = \pi/2 \\ 0 & \theta = \pi \end{bmatrix} = \frac{1}{2}\cos^2\frac{\theta}{2}$	$\frac{1}{2}\sin^2\frac{\theta}{2}$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$	1

1. $\theta = 0$: $P(a_1, b_1) = 0$. Combination is **highly unlikely, which adds** high amount of pair-information (∞) to the information by a_1 and b_1 alone.
2. $\theta = \pi/2$: $P(a_1, b_1) = 1/4$ which is the average we might expect for four pairs. No further information added by the combination, this equals the average of the alternatives.
3. $\theta = \pi$: With a_1 present we **expect** b_1 to be present and vice versa.
   a_1 and b_1 **do not contribute** their information **independently**.
   Combination yields a **loss** of information.

Per-Pair Transinformation: Ansatz and Definition

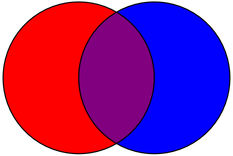


Fig. 16: Venn diagram for two sets motivates the definition of an overlap.

The overlap in the Venn diagram for sets motivates the ansatz:

$$\underbrace{I(a_i\ ,\ b_j)}_{\text{info in pair}} = \underbrace{I_A(a_i)}_{\text{contribution of } a_i} + \underbrace{I_B(b_j)}_{\text{contribution of } b_j} - \underbrace{I(a_i\ ;\ b_j)}_{\text{correction for overlap}}$$

The **per-pair transinformation** (also: **mutual information**) is defined as

$$I(a_i\ ;\ b_j) := I_A(a_i) + I_B(b_j) - I(a_i\ ,\ b_j)$$

Beware the subtle notational difference of $\boxed{;}$ versus $\boxed{,}$ (another notational abuse!).

Per-Pair Transinformation: Analysis

***Contrary* to Venn-diagram intuition but *in line* with our example the *per-pair* transinformation may be negative!**

Interpretation:

- **Negative:** Common occurrence of the two symbols is unusual.
  Thus it provides *additional* information.
- **Zero:** The two symbols in the pair are stochastically independent.
- **Positive:** One symbol in the pair can be predicted from the other with some chance.

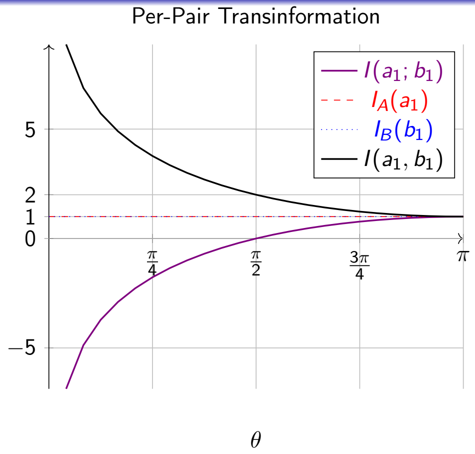


Fig. 17: Per-pair transinformation for the Bell example.
$I(a_i\,;\,b_j) := I_A(a_i) + I_B(b_j) - I(a_i\,,\,b_j)$

Expectation Value of Transinformation

The **expectation value** of the per-pair transinformation **over all pairs** of a compound $p\colon A \times B \to [0,1]$ is

$$I(A\,;\;B) = \mathcal{E}_{(a,b)\in A\times B}(I(a\,;\;b))$$

$$I(A\,;\;B) := \sum_{a\in A, b\in B} p(a,b)\cdot I(a\,;\;b)$$

Again surprising: The **expectation value** over all pairs always is non-negative. Formal proof see slide ??.

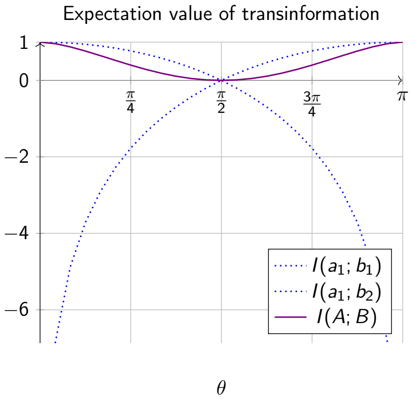


Fig. 18: The expectation value of the transinformation is non-negative, although the contribution of some individual pairs may be negative.

Expectation Vaue of Transinformation: Running Example

$\theta = 0$ $\theta = \pi$ **perfect anti correlation**

	b_1		b_2		
a_1	0	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
a_2	$\frac{1}{2}$	0	0	$\frac{1}{2}$	$\frac{1}{2}$
	$\frac{1}{2}$		$\frac{1}{2}$		1

$\theta = \pi/2$ **zero coupling**

	b_1	b_2	
a_1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
a_2	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$	1

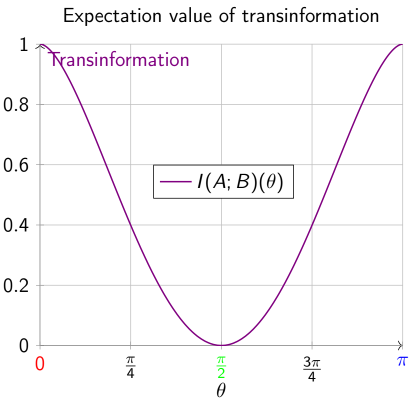


Fig. 19: The expectation value of the transinformation in a better magnified plot.

Formulae for Information and Transinformation

Information:

$$I_A(a_i) = -\log_2(P_A(a_i)) \qquad I_B(b_j) = -\log_2(P_B(b_j)) \qquad I(a_i, b_j) = -\log_2(P(a_i, b_j))$$

(Per-pair) transinformation:

$$I(a_i\,;\,b_j) = I_A(a_i) + I_B(b_j) - I(a_i, b_j) = \log_2 \frac{P(a_i, b_j)}{P_A(a_i) \cdot P_B(b_j)}$$

(Expected) transinformation:

$$I(A\,;\,B) = \sum_{a \in A\ b \in B} P(a, b) \cdot \log_2 \frac{P(a, b)}{P_A(a) \cdot P_B(b)} = -\sum_{a \in A\ b \in B} P(a, b) \cdot \log_2 \frac{P_A(a) \cdot P_B(b)}{P(a, b)}$$

Transinformation is Non-Negative

Proposition: (Expectation of) transinformation is non-negative.

Proof:

$$
\begin{aligned}
I(A;B) &= -\sum_{a\in A\ b\in B} P(a,b)\log_2 \frac{P_A(a)\cdot P_B(b)}{P(a,b)} && \text{(definition)}\\
&\geq -\log_2\left(\sum_{a\in A\ b\in B} P(a,b)\frac{P_A(a)\cdot P_B(b)}{P(a,b)}\right) && \text{(Jensen on negative log)}\\
&= -\log_2\left(\sum_{a\in A\ b\in B} P_A(a)\cdot P_B(b)\right) && \text{(reduction)}\\
&= -\log_2\left(\sum_{a\in A} P_A(a)\cdot\sum_{b\in B} P_B(b)\right) && \text{(distributivity)}\\
&= -\log_2(1\cdot 1) = 0 && \text{(probability)}
\end{aligned}
$$

Outlook

Classically modeled information leads to non-negative transinformation.

Quantum phenomena can be interpreted as

- having negative information (Feynman: 1984 & 1987 (in Hiley & Peat: Quantum implications))
- exhibiting interference (wave intuition)
- being deterministic plus guide wave (Bohmian mechanics)
- requiring an orthomodular logic (Birkhoff)
- holistically dependent on the entire universe (Zurek, Pietschmann)
- being completely described by a Fortran program

Glacier metaphora...

9. Information Channels

9.1. Transforming Information and Processing Data

9.2. Concept of a Channel

9.3. Symmetric Binary Channel

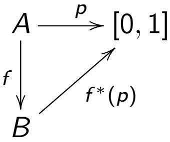
9.4. Channel Capacity and Conditional Entropy

Definition: Push Forward Measure

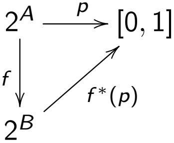
Let (A, p) be an information source and $f\colon A \to B$ an arbitrary function.

We recall: A finite, $p\colon 2^A \to [0,1]$ probability measure, p on 2^A induced by its restriction $p|_A\colon A \to [0,1]$ to A.

The **push forward measure** of p under f (“**Bildmaß**”) is the uniquely defined function $f^*\colon 2^B \to [0,1]$ which makes the following diagram commutative:



equiv.



equiv.

$$f^*(p) = p \circ f^{-1}$$

$$[f^*(p)](\{a\}) = p(f^{-1}(\{a\}))$$

$$[f^*(p)](U) = p(f^{-1}(U))$$

Data Processing Theorem: Entropy of a Transformed Source

Definition: Function $f : A \to B$ remaps the symbols and transforms information source $\mathcal{S}(A, p)$ into information source $(\mathcal{S}) := (B, f^*(p))$.

Data Processing Theorem:[2]

$$H(\ f(\mathcal{S})\ ) \leq H(\mathcal{S})$$

Special Case 1: Equality if and only if f is a bijection.

Special Case 2: Collapsing symbols $(f(a_1) = f(a_2))$ destroys information.

Special Case 3: If f is constant, then $H(\ f(\mathcal{S})\ ) = 0$

Interpretation: Deterministic processing cannot increase the entropy.

Application: Applying scrambling functions cannot be used to increase the entropy of a source of randomness. Important in cybersecurity.

[2] Weak form; there is a stronger version using Markov Chains!

Proof of Weak Data Processing Theorem (1)

Proof obligation: Make the sum

$$H(f(\mathcal{S})) = \mathcal{E}_{f^*(p);b\in B}(I_{f^*(p)}(b)) = -\sum_{b\in B} p(f^{-1}(\{b\})) \cdot \log_2(p(f^{-1}(b)))$$

larger-or-equal until we obtain

$$H(\mathcal{S}) = -\sum_{a\in A} p(a)\log_2(p(a))$$

There are three types of summands in $\sum_{b\in B}$.

Type 1: $f^{-1}(\{b\}) = \emptyset$ has no contribution and may be neglected due to a continuity argument and $0 \cdot \log_2(0) = \lim_{x\to 0+} x\log_2(x) = 0$.

Type 2: $f^{-1}(\{b\}) = \{a\}$ only produces a 1-1 relabeling.

Type 3: $f^{-1}(\{b\}) = \{a_1, \ldots, a_k\}$ with some k.

Proof of Weak Data Processing Theorem (2)

$-p(\{a_1,\ldots,a_k\})\cdot\log_2(p(\{a_1,\ldots,a_k\})) = [p(a_1)+\ldots+p(a_k)](-\log_2)[p(a_1)+\ldots+p(a_k)]$

Using Jensen inequality on the convex function $(-\log_2)$

$\leq [p(a_1)+\ldots+p(a_k)][p(a_1)(-\log_2)(p(a_1))+\ldots+p(a_k)(-\log_2)(p(a_k))]$

The sum of probabilities is less-or-equal 1

$\leq -p(a_1)\log_2(p(a_1))-\ldots-p(a_k)\log_2(p(a_k))$

The theorem follows from an application of all 3 types.

The special cases are easy to see.

Do Classical Physical Processes Destroy Information? (2)

Liouville Theorem: Phase space volumes, when transported by the flow of a Hamiltonian evolution, stays constant.

Interpretation: The phase space points move like an incompressible liquid.

In **time discrete** and **space discrete** situations this corresponds to:

Interpretation: If the information source transformation function is bijective, it does not merge or "compress" points ($f(a_1) = f(a_2)$) and the entropy remains constant.

Thus: Conservative Hamiltonian systems do not destroy or generate information.

Do Quantum Physical Processes Destroy Information?

We know: Density operator ρ evolves by conjugation with a unitary semi-group:

$$\rho(t) = U(t)\rho(0)U^*(t)$$

We know: von Neumann entropy is invariant under unitary transformation:

$$S(U\rho U^*) = S(\rho)$$

Thus: Closed quantum mechanical systems do not destroy or generate information.

Intuition for Finite Memoryless Channel

Generalize: From deterministic transformation to *probabilistic* transformation.

Channel mechanism:

- Whenever the channel sees an input symbol $a \in A$ at the input port
- it produces a *random* output symbol $b \in B$ at the output port.
- Probability may depend on input symbol $a \in A$
  For $a \in A$ we know the probability distribution of the produced output symbol.

Finite: From a finite number of different (digital) symbols *one* symbol is provided.

Memoryless: Assume a repetition of channel transmissions and

1. probability is time-independent $\Rightarrow$ can model by one value
2. repeated transmissions are pairwise independent $\Rightarrow$ probabilities multiply
3. in repeated transmissions, relative symbol frequency converges to probability

Definition for Finite Memoryless Channel

A (finite, memoryless) **information channel** is a triple $\mathcal{C} = (A, c, B)$ consisting of

1. a finite set A, whose elements are called **input** symbols
2. a finite set B, whose elements are called **output** sybols
3. a function $c\colon A \to \mathcal{M}(B)$, which **maps** every **input** symbol a **to a probability** measure $c(a)(\cdot)\colon 2^B \to [0,1]$.

$\mathcal{M}$: "set of measures"

$$c(a)\colon 2^B \to [0,1]$$

or rather: on σ — algebra

$$c(a)\colon B \to [0,1] \text{ with } \sum_{b\in B} c(a)(b) = 1$$

Most convenient form:

$$c\colon A \times B \to [0,1] \text{ with } \forall a \in A : \sum_{b\in B} c(a,b) = 1$$

Situation 1: Clamping Input to a Channel

Observation: If we clamp the input of a channel to a fixed symbol $a \in A$, we see an information source over B at the output of the channel with probability measure $c(a) : B \to [0, 1]$.

Observation: A channel is an (input symbol)-parametrized information source.

Observation: In the clamped situation, all information at the channel output is channel noise. There is no information at the input!

Situation 2: Connecting a Source to a Channel

Architecture: Information source (A, s) is connected to input of channel (A, c, B).

Independence: Channel action is independent from source action.

Consequence: Probability that we see input a and output b is given by:

$$p(a, b) = s(a) \cdot (c(a, b)$$

Thus: Conditional probability to get output b under the condition of input a is

$$p(b|a) = \frac{p(a, b)}{p(a)} = \frac{s(a) \cdot c(a, b)}{s(a)} = c(a, b)$$

matches the interpretation of $c(a, b)$ from before.

Reminder: In a non-quantum situation *measuring* the input character has no influence on its probabilities.

Situation 3: Interpreting Output of a Channel

Question: We got output symbol b. With which probability was a the input symbol?

$$p(a|b) = \frac{p(a,b)}{p_A(b)} = \frac{s(a) \cdot c(a,b)}{\sum_{\alpha \in A} s(\alpha) \cdot c(\alpha, b)}$$

Observation 1: When source is equi-distributed, it is a weighted average of the channel factors:

$$\frac{c(a,b)}{\sum_{\alpha \in A} c(\alpha,b)}$$

Observation 2: When source is skewed, it may heavily depend on the source distribution.

Example: Typical Channel (1)

$A := \{R, S\}$ $B: \ = \{\rho, \sigma, \tau\}$ $c_{a \in A;\ b \in B}$

	ρ	σ	τ	
R	0.8	0.1	0.1	1
S	0.0	0.0	1.0	1
	0.8	0.1	1.1	2

R @input becomes ρ @output with some errors made by channel

S @input reproduced as τ @output

Channel matrix:

- Rows are **stochastic**: Rows sum to 1.
- Columns are **not** stochastic.
- Overall sum is number of input symbols.
- Blue: Marginal sums, which here are *not* probability distributions.

Example: Typical Channel (2)

Clamping the $\mathcal{C}$-input to R produces an information source over $\{\rho, \sigma, \tau\}$ with
$p_{R\hookrightarrow\mathcal{C}}(\rho) = 0.8 \quad p_{R\hookrightarrow\mathcal{C}}(\sigma) \quad p_{R\hookrightarrow\mathcal{C}}(\tau) = 0.1$.

Clamping the $\mathcal{C}$-input to S produces an information source over $\{\rho, \sigma, \tau\}$ with
$p_{S\hookrightarrow\mathcal{C}}(\rho) = p_{S\hookrightarrow\mathcal{C}}(\sigma) = 0 \quad p_S(\tau) = 1$.

Connecting the source $\mathcal{S} = (\{R, S\}, s)$ with $s(R) = 0.2$ and $s(S) = 0.8$ to the $\mathcal{C}$ input port produces an information source over $\{\rho, \sigma, \tau\}$ with

$$p_{\mathcal{S}\to\mathcal{C}}(\rho) = s(R)c(R, \rho) + s(S)c(S, \rho) = 0.16$$

$$p_{\mathcal{S}\to\mathcal{C}}(\sigma) = s(R)c(R, \sigma) + s(S)c(S, \sigma) = 0.02$$

$$p_{\mathcal{S}\to\mathcal{C}}(\tau) = s(R)c(R, \tau) + s(S)c(S, \tau) = 0.82$$

Connecting Sources to Channels

Convention: Write channel matrices as above:
Rows denote **input** ports
columns denote **output** ports.

Convention: Write information sources as row vectors.
In our case: $\begin{pmatrix} p_{\mathcal{S}}(R) & p_{\mathcal{S}}(S) \end{pmatrix}$

Result: Connecting the source $\mathcal{S}$ to the channel $\mathcal{C}$.
Produces information source $\mathcal{S} \to \mathcal{C}$
Characterized by the row vector:
$\vec{p}_{\mathcal{S}\to\mathcal{C}} = \vec{\mathcal{S}} \cdot \overset{\leftrightarrow}{\mathcal{C}}$

Channels as Compound Sources, Definition

Now we model source-channel interaction as compound information source $\mathcal{S} \triangleright \mathcal{C}$.

Channel Protocol: Observe occurrence of input $a_j \in A$ and then output $b_j \in B$.

Source probability: $s\colon A \to [0,1]$ with $\sum_{a\in A} s(a) = 1$

Channel description: $c\colon A \times B \to [0,1]$ with $\forall a : \sum_{b \in B} c(a,b) = 1$

Compound probability: $p\colon A \times B \to [0,1]$ with $p(a,b) := s(a) \cdot c(a,b)$

Product warranted due to assumption of independence.

Is p **really** probability on $A \times B$? Check that $\sum_{a,b} p(a,b) = 1$.

$$\sum_{a,b} p(a,b) = \sum_a \sum_b s(a) \cdot c(a,b) = \sum_a s(a) \cdot \underbrace{\sum_b c(a,b)}_{=1} = \sum_a s(a) = 1$$

Channels as Compound Sources, Analysis

We study the channel protocol $\mathcal{S} \triangleright \mathcal{C}$ as compound $p \colon A \times B \to [0,1]$.

1. **Marginal:** $p_A \colon A \to [0,1]$ recovers the (source) **distribution** s at the **input port**.
$$p_A(a) = \sum_{b \in B} s(a) \cdot c(a,b) = s(a) \cdot \sum_{b \in B} c(a,b) = s(a)$$

2. **Marginal:** $p_B \colon B \to [0,1]$ is the symbol **distribution** at the **output port**.

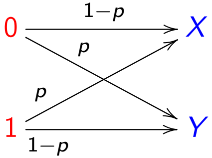
Joint: In general: $P(X \cap Y) = P(X) \cdot P(Y|X)$
Specialized: $P(i = a \ \wedge \ o = b) = P(i = a) \cdot P(o = b \mid i = a)$
Here: $p(a,b) = s(a) \cdot c(a,b)$

$p(a,b)$ probability to see pair (a,b) in the protocol

Conditional: $c(a,b)$ conditional probability that the channel outputs b under the condition that the provided input was a

Definition of Symmetric Binary Channel

Symbols for **input** $A = \{0, 1\}$, **output** $B = \{X, Y\}$, channel behavior as below.



$$\begin{array}{c c} & \begin{array}{cc} X & Y \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} \end{array}$$

with $p \in [0, 1]$.

Symmetric: Exchanging the roles of the symbols (either in input or in output) does not change anything. Matrix is bi-symmetric.

When coupled to source: **two-parameter system** in $(s, p) \in [0, 1]^2$

s Probability distribution of source; here of symbol 0

p:
- $p = 0$ Deterministic mapping input to output
- $p = 1$ Deterministic mapping input to output, dual variant
- $p \in (0, 1)$ Some room for "error"

Probabilities at the Output Ports

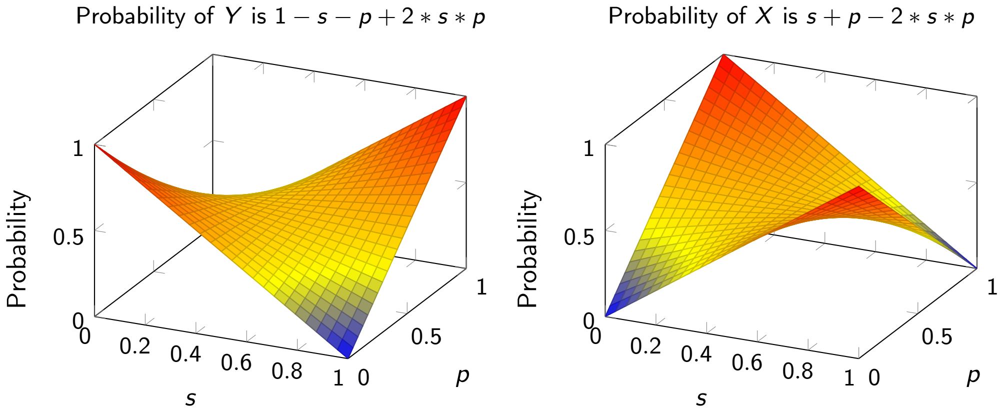


Fig. 20: Probabilities of the output symbols in the symmetric binary channel. For $p = 0$ we see an exact reproduction of the source distribution. When moving from $p = 0$ to $p = 1$ the straight line is "flipped". Probabilities of X and Y add up to 1.

Entropy Analysis (1): Input Port and Output Port

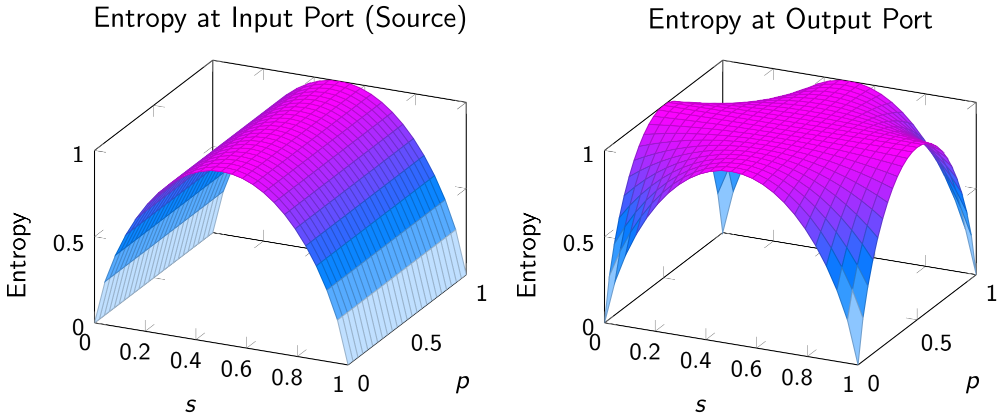


Fig. 21: The entropy at the input port depends only on the source parameter s. The entropy at the output port is larger and depends also on the channel parameter p. Idea: Clamp input to fixed value to see influence of channel alone (see Fig. ??).

Entropy Analysis (2): Output Port with Clamped Input

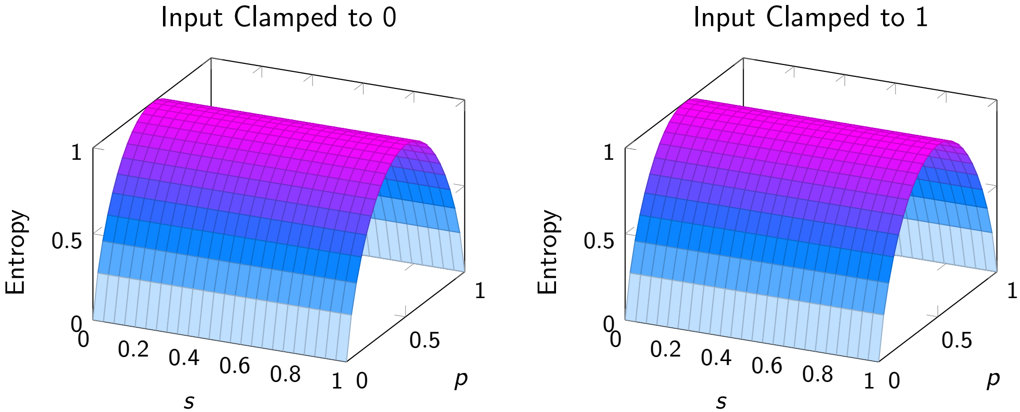


Fig. 22: Clamping the input to a fixed value reveals the entropy on the output port for constant input. It is not necessarily the same for all inputs but here, for the symmetric channel, it is. $p = 0$ and $p = 1$ is a deterministic mapping.

Entropy Analysis (3): How does Output Entropy arise?

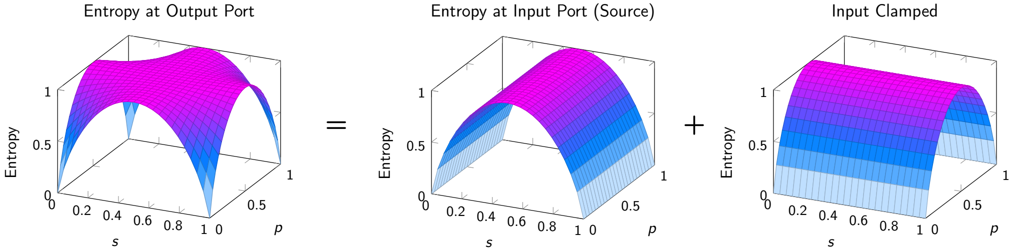


Fig. 23: Looking at the entropy from the input source and the entropy from the channel at clamped inputs gives us an idea why the shape of the output entropy is as it is.

Parameter Set of Maximum Output Entropy (1)

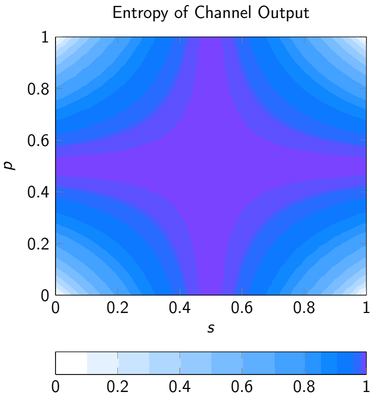


Fig. 24: Contour plot of entropy found at the channel output. It confirms that on the parameter set $\{(s,p) | s = 0.5 \vee p = 0.5\}$ we have maximum output entropy.

Parameter Set of Maximum Output Entropy (2)

Situation 1: Deterministic Mapping

$p = 0$ channel matrix $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $p = 1$ channel matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

Channel implements error free, non-random mapping of source: $0 \mapsto X \quad 1 \mapsto Y$ or $1 \mapsto X \quad 0 \mapsto Y$.

We get maximal output entropy only for $s = 0.5$ (where the source alone produces maximum entropy).

Parameter Set of Maximum Output Entropy (3)

Situation 2: No Effect from Input

$p = 0.5$ channel matrix $\begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$

Input symbol has no effect at all, for all s.

Proof: Joint probabilities are $\begin{pmatrix} 0.5 \cdot s & 0.5 \cdot s \\ 0.5 \cdot (1-s) & 0.5 \cdot (1-s) \end{pmatrix}$

We get:

$$p(o = X) = \frac{1}{2} = p(o = X | i = 0)$$

$$p(o = X) = \frac{1}{2} = p(o = X | i = 1)$$

Thus: Output $o = X$ is independent of the choice of the input $i = 0$ or $i = 1$.

Parameter Set of Maximum Output Entropy (2)

Situation 3: Maximal Source Entropy

At $s = 0.5$, source entropy is maximal.

At $p = 0$ and at $p = 1$ there is no disturbance by the channel.

At $p = 0.5$ there is maximal disturbance by the channel, so strong that even no source information can go through.

Input-Output Transinformation

Transinformation of the channel protocol models the amount of information output port has in common with input port.

Best model for channel information flow.

Graph corresponds to our analysis.

$p = 0.5$ maximal disturbance.
$p = 0$ and $p = 1$ source undisturbed.
At every p with $s = 0.5$
source most effective for channel.

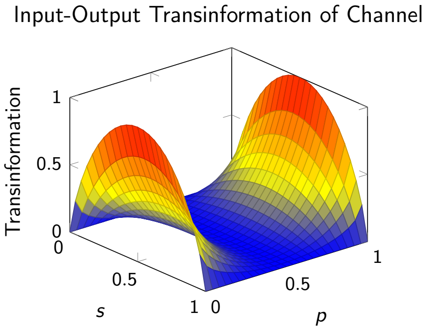
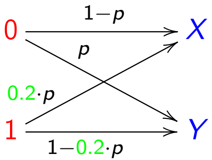


Fig. 25: Input-output transinformation of symmetric binary channel.

Example: Asymmetric Binary Channel

Symbols for **input** $A = \{0, 1\}$, **output** $B = \{X, Y\}$, channel behavior as below.



$$\begin{matrix} & X & Y \\ 0 & \multirow{2}{*}{} 1-p & p \\ 1 & 0.2 \cdot p & 1 - 0.2 \cdot p \end{matrix}$$

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} X & Y \end{array} \\ \begin{pmatrix} 1-p & p \\ 0.2 \cdot p & 1 - 0.2 \cdot p \end{pmatrix} \end{array} \quad \text{with } p \in [0, 1].$$

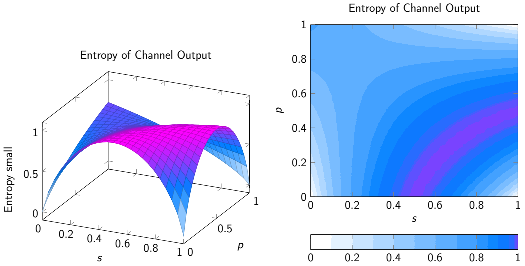


Fig. 26: Just to show that for an asymmetric binary channel, the situation is in fact more twisted.

Channel Capacity

For a channel $\mathcal{C} = (A, c, B)$ and a suitable source $\mathcal{S} = (A, s)$ the **joint probability** is $(a, b) \mapsto s(a) \cdot c(a, b)$ and the **transinformation** thereof shall be written $I(\mathcal{S}\,;\mathcal{C})$

Fact 1: Given a channel, the transinformation depends on the (probability distribution) of the source.

Fact 2: The source can be adapted to the channel to maximize the transinformation.

The **capacity** $\mathfrak{C}(\mathcal{C})$ of a channel $\mathcal{C} = (A, c, B)$ is the maximum value of the transinformation a suitable source $\mathcal{S} = (A, s)$ coupled to a channel may achieve.

$$\mathfrak{C}(\mathcal{C}) := \sup_{\mathcal{S}} I(\mathcal{S}\,;\mathcal{C})$$

Motivation: Conditional Entropy

Conditioning studies changes a condition imposes on probability or info content.
Entropy is the expectation value of the information content.

Conditional entropy is defined as the
normal **expectation value** of the *conditional* **information content.**

Recall: Condition also affects the expectation value operator.
Cave: The conditional entropy *could* be defined but **is not** defined as

- *Conditional* expectation value of the *normal* information content.
- *Conditional* expectation value of the *conditional* information content.

This makes a difference, since conditioning-induced norming affects

1. probabilities by a multiplicative factor
2. information content additively via the logarithm
3. expectation value operators via the summation range

Definition: Conditional Entropy

Let $\mathcal{S} = (A, p)$ be an information source. Let $X \subseteq A$ with $p(X) \neq 0$.

The **conditional entropy** $H_{|X}(\mathcal{Q})$ of the source $\mathcal{S}$ under the condition X is the **expectation value of the conditional information content**:

$$H_{|X}(\mathcal{S}) = H(\mathcal{S}|X) = \sum_{a \in A} p(a) \cdot I_{|X}(a)) = -\sum_{a \in A} p(a) \cdot \log_2(p_{|X}(a)) = -\sum_{a \in A} p(a) \cdot \log_2 \frac{p(a)}{p(X)}$$

Joint Entropy for Compounds

Let $\mathcal{C} = (A \times B, p)$ be a compound.

The **(joint) entropy** is the expectation value of the joint information content:

$$H(\mathcal{C}) = H(A, B) = -\sum_{a \in A, b \in B} p(a, b) \cdot \log_2 p(a, b) = \mathcal{E}(I(a, b))$$

The **marginal entropies** are the entropies of the marginal information content:

$$H_A = H(A) = -\sum_{a \in A} p_A(a) \cdot \log_2 p_A(a)$$

$$H_B = H(B) = -\sum_{b \in B} p_B(b) \cdot \log_2 p_B(b)$$

Connecting Joints, Marginals and Conditionals

For **probabilities** we had: $p(a,b) = p_A(a) \cdot p(b|a)$.

$$H(\mathcal{Q}) = H(A,B) = -\sum_{a,b} p(a,b) \cdot \log_2 p(a,b) = -\sum_{a,b} p(a,b) \cdot \log_2 \left(p_A(a) \cdot p(b|a)\right) =$$

$$-\sum_{a,b} p(a,b) \cdot \log_2 p_A(a) - \sum_{a,b} p(a,b) \cdot \log_2 p(b|a) =$$

$$-\sum_{a} \underbrace{\left(\sum_{b} p(a,b)\right)}_{=p_A(a)} \cdot \log_2 p_A(a) - \sum_{a,b} p(a,b) \cdot \log_2 p(b|a) =$$

$$-\sum_{a} p_A(a) \cdot \log_2 p_A(a) - \sum_{a,b} p(a,b) \cdot \log_2 p(b|a) = H_A(\mathcal{Q}) + H_{|B}(\mathcal{Q}) = H(A) + H(A|B)$$

For **entropies** we obtained: $H(A,B) = H(A) + H(B|A) = H(B) + (A|B)$

Channel Equations: Derivation from Transinformation

$$I(A\,;B) = \sum_{a\in A\ b\in B} p(a,b)\cdot\log_2\frac{p(a,b)}{p_A(a)\cdot p_B(b)} =$$

$$+\sum_{a,b} p(a,b)\cdot\log_2 p(a,b) - \sum_{a,b} p(a,b)\cdot\log_2 p_A(a) - \sum_{a,b} p(a,b)\cdot\log_2 p_B(b) =$$

$$\underbrace{-H(A,B)+H(A)}_{-H(B|A)} + H(B) = H(B) - H(B|A) = \text{(similarly)} = H(A) - H(A|B)$$

We obtain the **channel equations:**

$$H(A) = H(A|B) + I(A\,;B)$$

$$H(B) = H(B|A) + I(A\,;B)$$

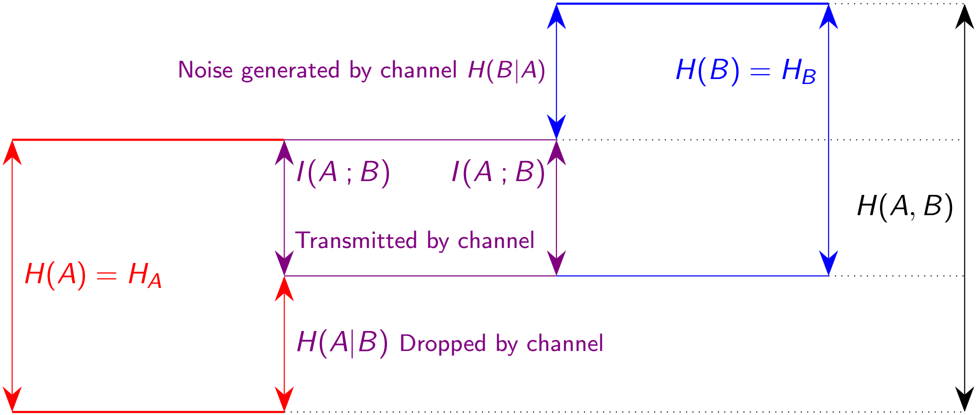
$$H(A,B) = H(A|B) + I(A\,;B) + H(B|A)$$

Channel Equations: Graphical Ilustration

$H(A) = H(A|B) + I(A\,;B)$

$H(B) = H(B|A) + I(A\,;B)$

$H(A,B) = H(A|B) + I(A\,;B) + H(B|A)$



Example: Deterministic Channel

An information channel (A, c, B) is called **deterministic**, iff $\forall a \in A : \exists b \in B : c(a, b) = 1$.

Properties of a deterministic channel:

$H(Y|X) = 0.$

$I(X \| Y) = H(Y).$

10. Kullback-Leibler Divergence

Definition: Kullback-Leibler Divergence

Let A be a set of symbols.
Let $\mathcal{P} = (A, p)$ and $\mathcal{Q} = (A, q)$ two information sources over this set A.
Assume: q vanishes for no symbol. This allows to condition on every $a \in A$ for $\mathcal{Q}$.

The **Kullback-Leibler divergence** is defined as

$$\mathcal{D}(p, q) := \sum_{a \in A} p(a) \cdot \log_2 \frac{p(a)}{q(a)} = -\sum_{a \in A} p(a) \cdot \log_2 \frac{q(a)}{p(a)}$$

Motivation: Kullback-Leibler Divergence

Sufficiently general formula structure

1. **Expectation value** of a
2. **logarithm** of a
3. **conditioned**
4. **probability**

Motivation can be found in the possible usages.

Fig. 27: Kullback-Leibler divergence is a swiss army knife of information theory. Rights see appendix.

Example: Binary Sources (1)

Consider two binary sources $\mathcal{P}$ and $\mathcal{Q}$ over $\{0,1\}$.

The sources are given by a parameter p and a parameter q as follows:

$$p_{\mathcal{P}}(0) = p \qquad p_{\mathcal{P}}(1) = 1 - p$$

$$p_{\mathcal{Q}}(0) = q \qquad p_{\mathcal{Q}}(1) = 1 - q$$

For the **Kullback-Leibler divergence** we get:

$$\mathcal{D}(p, q) = p \cdot \log_2 \frac{p}{q} + (1-p) \cdot \log_2 \frac{1-p}{1-q}$$

Example: Binary Sources (2)

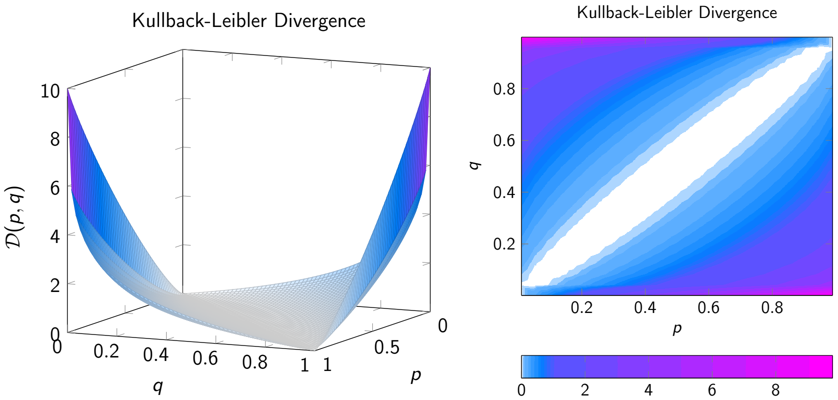


Fig. 28: Kullback-Leibler divergence $\mathcal{D}(p,q)$ of two binary sources, characterized by parameter p and q, respectively. Note that $\mathcal{D}(p,q) = 0 \Leftrightarrow p = q$. This prompts the **question** whether it is a metric!

Example: Binary Sources (3)

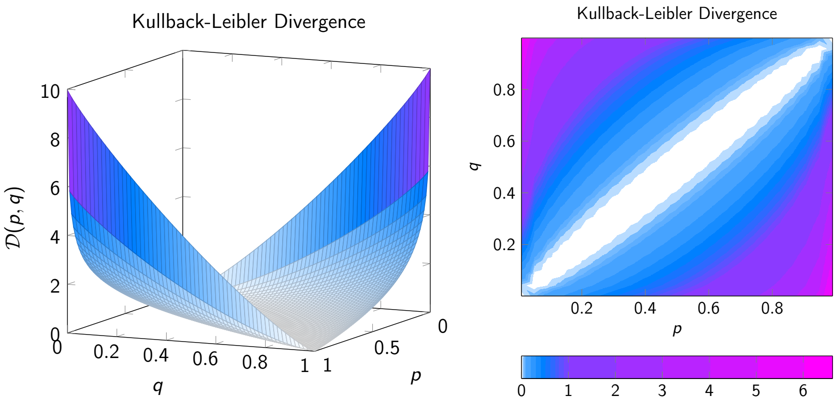


Fig. 29: Kullback-Leibler divergence $(p, q) \mapsto \mathcal{D}(q, p)$ as opposed to $(p, q) \mapsto \mathcal{D}(p, q)$ in the earlier plot. All plot parameters are the same. Comparison – short of rounding effects – suggests that it is not symmetric.

Is the Kullback-Leibler Divergence a Metric?

No it is not a metric.

Positive definite: $\forall p, q : \mathcal{D}(p,q) \geq 0$ and $\mathcal{D}(p,q) = 0 \Leftrightarrow p = q$

Not symmetric: ~~$\mathcal{D}(p,q) = \mathcal{D}(q,p)$~~

$\mathcal{D}$ could be made symmetric: $\mathcal{D}_s(p,q) = \frac{\mathcal{D}(p,q)+\mathcal{D}(q,p)}{2}$

No triangle inequality: ~~$\mathcal{D}(p,q) + \mathcal{D}(q,r) \geq \mathcal{D}(p,r)$~~

Asymmetry

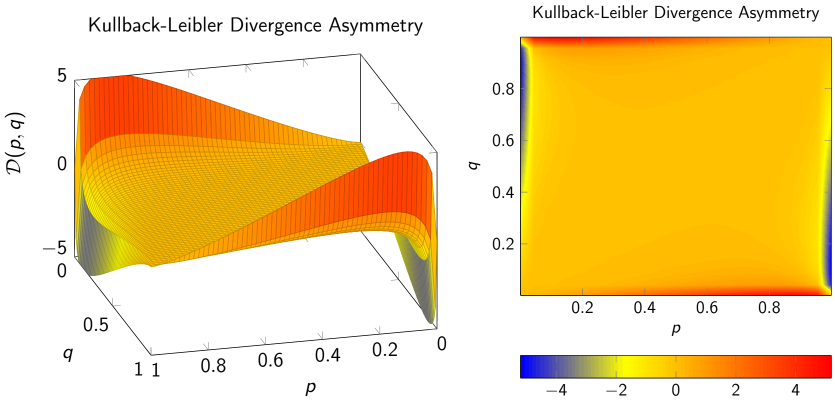


Fig. 30: Plotting $\mathcal{D}(p, q) - \mathcal{D}(q, p)$ to illustrate the asymmetry of the Kullback-Leibler divergence. Note, how this difference is zero for $p = q$ (positive definite). Note, how this difference is zero for $p = 1 - q$ (symmetry of the binary source).

Symmetrized Variant

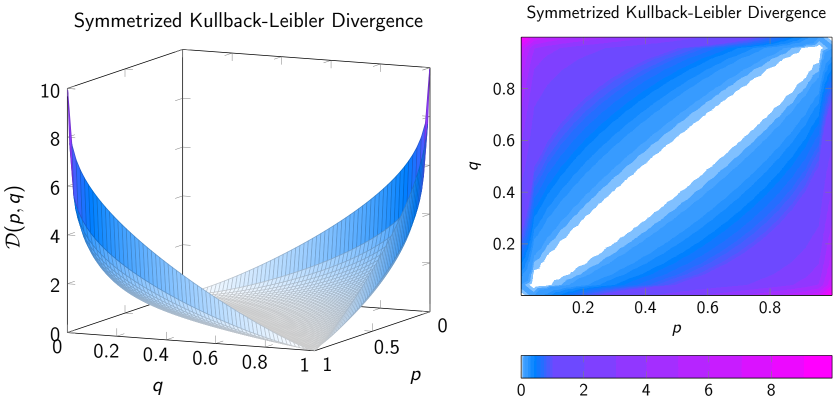


Fig. 31: Symmetrized Kullback-Leibler divergence $\mathcal{D}_s(p,q) = (\mathcal{D}(p,q) + \mathcal{D}(q,p))/2$ of two binary sources, characterized by parameter p and q, respectively. Note that $\mathcal{D}(p,q) = 0 \Leftrightarrow p = q$

Kullback-Leibler Connections

Transinformation as KLD from the joints to the product of the marginals:

$$I(A\,;B) = \mathcal{D}(p, p_A \otimes p_B) = \mathcal{D}_{a,b}(p(a,b), p_A(a) \cdot p_B(b))$$

Where $p\colon A \times B \to [0,1]$, $p_A\colon A \to [0,1]$ and $p_B\colon B \to [0,1]$,
with $p_A \otimes p_B\colon A \times B \to [0,1]$ as $(p_A \otimes p_B)(a,b) = p_A(a) \cdot p_B(b)$.

Redundancy as KLD to the equi-distribution:

$$R(p) = \mathcal{D}(p, u_n)$$

Where $|A| = n$ and $u_n\colon A \to [0,1]$ be the equi-distribution on A.

Entropy as co-KLD to the equi-distribution

$$H(p) = H_{\text{max}} - \mathcal{D}(p, u_n)$$

Conditional Entropy as KLD:

$H(A|B) = H(A) - I(A\,;B)$ and both right-side terms are KLDs

11. Overview on Coding Theorems

The Problems of Coding Theory

Problem 1: Source Inefficiency
Non-equidistribution in source symbols make a source less efficient.
It produces less information per symbol on the average than theoretically possible.

Problem 2: Source Adaption
An information source may not be optimally adjusted to the channel.
The transinformation of a source/channel compound is smaller than channel capacity.

Problem 3: Channel Information Loss
The randomness in a channel may lead to loss of source information.
Concept of channel capacity points to channel-source adaptation.

Problem 4: Channel Noise
The randomness in a channel may introduce additional, unwanted noise.

The Solutions by Shannon

All these problems may be dealt with.

The solutions can be made **asymptotically optimal**.

Shannon Coding Theorems

Source coding theorem: Every source can be *recoded* so as to *asymptotically* achieve *nearly* the maximally possible entropy.

Channel coding theorem: By suitable coding at the input and output port, a channel can be used such that *asymptotically* at the same time two goals can be achieved:

1. the error rate through the system is *nearly* zero.
2. the channel capacity is utilized as *nearly* as good as possible *given* a specific error rate.

Nearly: As close as we want but not completely. (Uses some $\varepsilon - \delta$ definition.)

Definitions: Monoids and Codings

Let A and B be two finite alphabets and let ε denote the empty word.
Let $B^0 := \{\varepsilon\}$
Let $B^* := \cup_{n \in \mathbb{N}_0} B^n$

A **coding** is a function $f\colon A \to B^*$.

The **extension** of a coding $f\colon A \to B^*$ is the function $f^*\colon A^* \to B^*$, uniquely defined by

$$f(\varepsilon) = \varepsilon \qquad \forall x, y \in A^* : f(xy) = f(x)f(y)$$

The ***n*-th extension** $f^n\colon A^n \to B^*$ of a coding $f\colon A \to B^*$ is the restriction of its extension $f^*\colon A^* \to B^*$ to A^n.

Definition: Source Extensions

The *n*-th **extension** of an information source $\mathcal{S} = (A, s)$ is the information source $\mathcal{S}^n = (A^n, s^n)$ where

$$s^n(a_1 \dots a_n) := s(a_1) \cdot \dots \cdot s(a_n)$$

Interpretation:

1. n pairwise independent copies of the information source.
2. We may think *sequential* or *parallel*.

Motivation: When using sources and channels we are not interested in single but in multiple usage.

Definition: Decoding

A coding $f\colon A \to B^*$ is called **uniquely decodeable**,
iff its extension $f^*\colon A^* \to B^*$ is injective.

Prefix Free

Definition: A subset $X \subseteq B^*$ is called **prefix free** iff $\forall w \in X : \forall u \in X : \forall v \in B^* : u \neq wv$.

Interpretation: No prefix of a word from X is in X.

Definition: A coding $f: A \to B^*$ is called **prefix free** iff its image $f(A)$ is prefix free.

Propositon: (1) A prefix free coding is uniquely decodable.
(2) The converse is not true.

Motivation of the Prefix-Free Condition

We consider 3 situations:

$f(a) = 11$ and $f(b) = 111$. What does 111111 encode? $f^*(aaa) = 111111 = f^*(bb)$

$f(a) = 01$ and $f(b) = 1$. What does 011 encode? We get $f^*(ab) = 011$, which is the only pre-image, since no code-word in $\{01, 1\}$ occurs as prefix of another code word.

$f(a) = 10$ and $f(b) = 1$. What does 110 encode? We get $f^*(ba) = 110$, which is the only pre-image, since no code-word in $\{10, 1\}$ occurs as suffix of another code word. However, we do not understand this while reading 110 from left to right – we only realize it at the end.

Prefix-freedom allows *unique decoding* while reading the code *from start to end*.

Note: Nomenclature in literature is "prefix code" instead of "prefix-*free*" code (which I consider misleading).

Example: Non Prefix-Free Decoding

We consider

$f(a) = 01$ and $f(b) = 011$.

The code is not prefix-free, since 01 is a prefix of 011.

However, the code is uniquely decoding.

Efficiency of a Source-Coding

The **average code word length** of a coding $f\colon A^n \to B^*$ for a source $\mathcal{S} = (A, s)$ is

$$L_{\mathcal{S},f} = \sum_{a \in A^n} s(a) \cdot \mathrm{len}(f(a))$$

The **relative efficiency** of a coding is

$$E_{\mathcal{S},f} = \frac{H(\mathcal{S})}{L_{\mathcal{S},f} \cdot \log_2 |B|}$$

Intuitively clear: $E_{\mathcal{S},f} \leq 1$.

Shannon Source Coding Theorem

Theorem: Every finite memoryless information source $\mathcal{S} = (A, \alpha)$ may be coded to asymptotic optimal efficiency.

More precisely: For every $\varepsilon > 0$ there exists an $n \in \mathbb{N}$ and a uniquely decodeable coding $f : A^n \to B^*$ such that

$$1 - \varepsilon < \frac{H(\mathcal{S})}{L_{\mathcal{S},f} \cdot \log_2 |B|} \leq 1$$

Interpretation: The left inequality tells us how good we should be able to code.
The right inequality tells us how good it can get at most.

Shannon-Weaver Communication System (1)

A **Shannon-Weaver Communication System** consists of the following components:

1. An information source $\mathcal{S} = (A, s)$ over A
2. A source encoding function $e\colon A^n \to B^*$
3. A channel (B, c, D)
4. A decoding function $d\colon D^* \to A^n$

$$\xrightarrow{\mathcal{S}^n} A^n \xrightarrow{e} B^* \xrightarrow{\mathcal{C}} D^* \xrightarrow{d} A^n$$

Shannon-Weaver Communication System (2)

1. The information source delivers a message $\vec{a} \in A^n$. This is random.
2. The source encoding turns this into a word over B. This is deterministic.
3. The channel transmits the individual characters according to c. This is random again.
4. The decoding function transforms this back into a word in A^n. This is deterministic.
5. The decoding might correct some errors.

In toto, a communication system can be regarded as a compound of the form $A^n \times A^n$ with a probability $p\colon A^n \to A^n$, generated as described.

In $p(\vec{i}, \vec{o})$ the word $\vec{i}$ is called the **input** and the word $\vec{o}$ is called the **output**.

The **error probability** of a communication system is given by

$$p(\vec{i} \neq \vec{o}) = p(\ \{(\vec{a}, \vec{b})\} \in A^n \times A^n \mid \vec{a} \neq \vec{b}\})$$

Shannon Channel Coding Theorem

Situation: Let $\mathcal{S} = (A, s)$ be a finite, memoryless information source and $\mathcal{C} = (B, c, D)$ a channel. Let $\varepsilon > 0$ and $\delta > 0$ be small positive real values.

Theorem: We can find an n, a source encoding function $e\colon A^n \to B^*$ and a decoding function $d\colon B^* \to A^n$ such that the resulting communication system satisfies:

Error: The probability of an error is smaller than ε
Transfer: The achievable transinformation is at least $R_{\mathcal{C}}(\varepsilon) - \delta$

For a given maximal error probability ε the achievable transinformation is always less-or-equal to the **rate function:**

$$R_{\mathcal{C}}(\varepsilon) = \frac{\mathfrak{C}(\mathcal{C})}{1 - H_2(\varepsilon)}$$

$H_2(x) = -x \log_2(x) + (1-x)\log_2(1-x)$ is the entropy function of a binary source.
$\mathfrak{C}(\mathcal{C})$ is the capacity of the channel.

Appendix

Bibliography

List of Slides

6. Shannon Information Theory

6.1. Probability

6.2. Conditional Probability

6.3. Information

7. Information Sources

7.1. Basic Definitions

7.2. Entropy and Redundancy

7.3. Examples

7.4. Convexity

8. Products and Compounds

8.1. Basic Definitions

8.2. Remarks on Marginals

8.3. Factorization

8.4. Example of a Compound

8.5. Transinformation

9. Information Channels

9.1. Transforming Information and Processing Data

9.2. Concept of a Channel

9.3. Symmetric Binary Channel

9.4. Channel Capacity and Conditional Entropy

Legend:
- continuation slide
- slide without title header
- image slide