Introduction to Classical Information Theory

What is Information and how to quantify it?

https://iuk.one/203-0002

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

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

May 15, 2026

Overview

1. Motivation
2. Conceptual Difficulties
3. Algorithmic Information Theory
4. Measuring Sets
5. Shannon Information Theory
6. Information Sources
7. Products and Compounds

1. Motivation

Why would a physicist want to study information theory?

Information and Physics

Norbert Wiener

Information is information **not matter or energy**. [Wie61], p132

Rolf Landauer

Information is Physical. [Lan91]

John Archibald Wheeler

It from a bit.

Every physical quantity, every it,
derives its ultimate significance from bits, binary yes-or-no indications. [Whe89, Whe90]

David Deutsch

It from qubit. [Deu04]

Attempts to Define Information

Information is a

1. Means for **constructing** objects — (will talk a bit on this)
   - **Algorithmic information theory** — Chaitin, Solomonov, Kolmogorov, Martin-Löf
   - **Complexity theory** — Blum
2. Choice of the **actual among the potential** — (will talk a lot on this)
   - **Probabilistic information theory** — Wiener, Shannon, Nyquist, Hartley
   - **Frequency analysis** of empirical outcomes. — [Haj19]
3. Human cognitive **behavioral construct** — (will not talk about this)
   - **Belief:** Calculus of human belief. — Bayes, Pearl. [Tal08], [Pea09]
   - **Propensity:** Tendency of favoring an outcome. — Peirce, Popper. [Whi72], [Pop59]
   - **Economy:** Readyness to engage in a bet. — Ramsey [Ram16], [BR11]

Concepts: Determinism and Non-Determinism

Hypothesis of Determinism

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

Famous Quote by Einstein

God does not play dice.

Hypothesis of Non-Determinism (“Regellosigkeit”)

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

Famous Quote by Bohr

Einstein, stop telling god what to do.

Important Observation

Lack of Rules

Auch völlige Regellosigkeit führt zu Regeln.
Even a complete lack of rules – has rules as a consequence.

Example: Central Limit Theorem

The random variable which is constructed as arithmetic mean of a large number of independent and identically distributed real random variables over the same probability space is approximately a Gaussian distribution.

Questions

In a Non-deterministic Theory

What are the rules emerging from a **lack of rules**.

In a Deterministic Theory

What are the rules emerging from

- not knowing the state of the observer
- not knowing the full state of the universe

Is Nature Deterministic or Non-Deterministic?

Theory:

- We know *both* deterministic *and* non-deterministic quantum theories.
- Think of: De Broglie-Bohm and Schrödinger/Heisenberg/von Neumann.
- The non-deterministic theory is socially more accepted and easier to extend.

Experiment:

- We cannot settle the question empirically.
- Reason: We, in principle, cannot do the *exact* same experiment twice.

The question is speculative and not scientific (it cannot be settled).

The answer seems to be an artifact of our modeling.

2. Conceptual Difficulties

Important differences between mathematics and physics.

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: Is an empirical science of **modeling observations**.

- If the experiment says different, the theory is dismissed as wrong.
- Theory may remain as useful approximation. (eg: Thermodynamics, classical mechanics)

Mathematics: Is a sciences of **cognitive constructions**.

- There is no experiment.
- Isn't mathematics restricted by some "natural" laws of logic?
- **No!**
- Mathematics is only restricted by the decisions of its cognitive design.
- But: Resulting math might not be practically useful.

Question 1: Was god restricted by some natural laws of logic?

Question 2: Is logic empirical? [Put68], [Dum76].

Our Methodology for Information Theory

Application of our understanding this far:

1. **Wrong:** "Information has certain properties."
2. **Correct:** "Our concept of information has certain properties."
   "Are these properties useful structures for modeling observations."

1. Which intuition do we want to achieve?
   Formulate axioms!
2. What are the consequences of the axioms?
   Prove theorems on existence, uniqueness and properties.
3. Does it describe physical observations?
   For **classical** physics yes, for **quantum** physics no.
4. Must redesign information theory to a quantum information theory!

How is the Classical World? (1)

There are **objects** and objects **have properties**.

Properties:

1. can be measured and **measurements tell us** about properties.
2. good measurements **minimize disturbances** of properties.
3. ideally measurements are **arbitrarily precise**
4. **exist before** the measurement — realism
5. exist **independently of** the measurement — conterfactual definiteness
6. exist **independently of** simultaneously measured properties — non-contextual
7. can **coexist** with all atomic properties — fully compatible
8. correlations propagate **inside the light cone** — causal
9. **propositions** about properties form a **Boolean** algebra / lattice / logic — Boolean

How is the Classical World? (2)

These **statements** are

1. **ontologic**: The world "is" like that,
2. **assumptions**: We have no deeper proof
3. **empirical**: Driven by every-day human experience.
4. **macroscopic**: Occur especially at (very) large particle counts.

They drive the cognitive framework which forms our concepts, logics, formal symbolic tools for modeling our world.

Thus our classical world looks like the classical assumptions which we plug in to the modeling tools employed.

Where is the Quantum World different?

Most of these classical assumptions are "wrong" in a quantum world.

"Wrong": Experiments falsify the predictions of classical theories,

1. Especially in the **microscopic** domain (Hydrogen, radiation laws)
2. but **not only** (Schrödingers cat, Wigners friend, measurement paradox)
3. Formal consequences of empirically better theories
   contradict formal consequences of classical theories. (CHSH, Bell, Kochen-Specker)
4. Cognitive consequences of empirically better theories **at odds** with
   cognitive consequences of classical theories. (Wave-particle dualism)
5. Mathematical tools seem peculiar (complex Hilbert space)
   but produce theories which are, thus far, not-falsified.

The more we look into the quantum world and its experiments
the more *our every-day cognition lacks concepts* for a good description.

Not completely clear yet, which ontologic assumptions will replace the classical ones.

Attempts to Describe the Quantum World

Random: In some experiments *seemingly* identical things *seem* to randomly produce different device states / measurement results (eg. u and d).

Idea 1: Describe states as **convex** random mix.

Example: $p_u \cdot u + p_d \cdot d$ with $p_u + p_d = 1$

But: Fails to describe experiments properly.

Idea 2: Some phenomena remind of the interference of waves. Describe them using amplitude / phase metaphoras.

Example: $\alpha \cdot u + \beta \cdot d$ with $\bar{\alpha} \cdot \alpha + \bar{\beta} \cdot \beta = 1$ plus Born rule

But: Fails to grasp degenerate situations.

Idea 3: Measurements are idempotent / projective: $P^2 = P$. Describe them using maximal rank projectors.

Example: $\lambda \cdot P_\lambda + \mu \cdot P_\mu$

Quantum Idealization: Hermitean Operators

Let Ψ be an *n*-**level quantum system** described by a vector ψ in *n*-**dimensional complex Hilbert space** H.

Let $\mathcal{A}$ be an **observable** described by a **Hermitean operator** A on H.

A decomposes into weighted sum of pairwise orthogonal projections.

$$A = \sum_{\lambda \text{ eigenvalue of } A} \lambda \cdot P_\lambda$$

P_λ maximal-rank orthogonal projection on the eigenspace with the eigenvalue λ **Measurement results** of $\mathcal{A}$ are eigenspaces of A.

$$\langle \psi | P_\lambda(\psi) \rangle$$

Probability of eigenspace with eigenvalue λ by **Born rule**.

$$\langle \psi | P_\lambda(\psi) \rangle = \langle\ \psi |\, \langle \vec{v}_\lambda | \psi \rangle \cdot \vec{v}_\lambda\ \rangle = \langle \psi | \vec{v}_\lambda \rangle \cdot \langle \vec{v}_\lambda | \psi \rangle$$

Generic, non-deg case.
$P_\lambda = |\vec{v}_\lambda\rangle\langle\vec{v}_\lambda|$

Quantum World Superposition

State spaces: Often (but not well) respresented as

- Cartesian products (as in $\mathbb{C} \times \mathbb{C}$) — Specific choice of a basis.
- Hilbert spaces — Ignores relative role of phase.

Better: Projective spaces.

(Pure) state space of a 2-level system:

- Is a 2-sphere.
- Fixing any two antipodes I can distinguish longitude (probability $\angle$) latitude (phase $\angle$)

Superposition: Often perceived in strange ways by classical minds.

- A particle is "*at the same time*" at position x_1 "*and*" position x_2.
- A thing is "*at the same time*" in states α "*and*" β
  and when asked randomly chooses what it is "*really*".
- We never see objects in a "superposition" of states (Schrödinger cat)

Superposition: Geometry of complex projective spaces.

Classical World: String Idealization

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

Let $A^* := \{a_1 a_2 \dots a_n \mid a_j \in A, n \in \mathbb{N}\} \cup \{\varepsilon\}$ 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.

Roadmapto the 4 lectures

Algorithmic Information Theory shows how one discipline of quantifying the *choice of the actual among the potential* is theoretically great but practically a dead-end road.

Measuring Sets shows how we can quantify
the choice of the actual among the potential.

Shannon Information Theory provides more precise notions
for quantifying specific situations.

Information Sources describes how generating information can be quantified.

Compounds and Products describes systems of subsystems –
similar to but different from entanglement.

Transinformation and KLD provide notions for describing the flow of information.

Information Channels use flow of information in classical computation.

3. Algorithmic Information Theory

Information as
means for constructing objects.

Example 1: Naïve Repetition

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

```
yyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy
```

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

```
print("yyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy");

for (var num=0; num < 70; num++) {print("y");}    // shorter program

for(var i=0;i<70;i++)print("y")                   // still shorter

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

Src. 1: Four programs for printing 70 copies of "y".

Example 2: More Advanced

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

```
56789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz
```

```
1
2 print("56789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\\]^_\`
3   abcdefghijklmnopqrstuvwxyz");
4
5 for (var num=53; num <= 122; num++) {printChar(num);}
6
7 for (var n=53;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.
Rather: This is **nearly never** the case. Can be made more precise!

Definition: Algorithmic Information Content

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.

3 Problems in Algorithmic Information Theory

Problem 1: We need a **more precise notion of a programming language**.

- Example 1: A Java program is fine.
- Example 2: A definite integral is fine, provided we can approximate the value.
- Counterexample: An arbitrary possibly "inconstructive" specification is **not** fine.

Problem 2: Different programming languages 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 may lead to different lengths.

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

Solution 1: Informal Notion of a Programming Language

Ingredients:

1. A **finite alphabet** A
2. A **recursively enumerable** subset $\mathcal{L} \subseteq A^*$.
3. A **partial recursive function** $\beta \colon \mathcal{L} \times \mathbb{N}^* \hookrightarrow \mathbb{N}$

Used to encode a language $\mathcal{L}$

Conditions:

1. For every partial recursive function $f \colon \mathbb{N}^* \to \mathbb{N}$ there exists a program $P \in \mathcal{L}$ such that $f = \beta(P)$.
2. The **UTM** property: We can build a **U**niversal **T**uring **M**achine (universal interpreter)
3. The **SMN** property: We can do partial evaluation.

Usually built up via notions of a **T**uring **M**achine or a **R**andom **A**ccess **M**achine.

More precisely: Lecture series in theoretical computer science or [Odi92], [Cha87].

Problem 1: ✓

Chaitin-Kolmogorov-Solomonoff Complexity

Let $\beta\colon \mathcal{L} \times \mathbb{N}^* \hookrightarrow \mathbb{N}$ represent a programming language.

Definition

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

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

[a] Natural numbers in some encoding.

Problem 2: γ_β depends on the computational concept β.

Problem 3: γ_β depends on the encoding of $\mathcal{L}$.

Solution 2 and 3: Transformation Theorem

Solution: The dependency is not strong: [Cha66, Cha87], [Kol68], [Sol64a, Sol64b].
Nearly as in the conversion of physical units.

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:

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

Limitation 1: 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:

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

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 $b_1 b_2 b_3 \ldots$ which is not in this list
and thus has no program printing it
and thus escapes every analysis by algorithmic information theory.

Limitation 2: 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. [STZDG14]

Very 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.

4. Measuring Sets

Quantifying the size of sets.

What do we want to achieve?

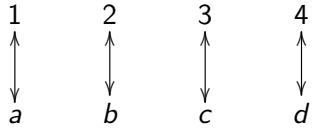
Goal: Information as choice of the actual in the set of the potential.
We want to quantify the size of a set.

Ansatz 1: Intuition of **counting**, leads to the concept of **cardinality**.
Problem: Has inconvenient artifacts which **cannot** be resolved.

Ansatz 2: Intuition of **contents**, leads to the concept of a **measure**.
Problem: Has inconvenient artifacts which **can** be resolved.
Issue: Requires a bit of technical overhead.

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.



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

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: The Intuition

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:

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

(3) Rotation and Reflection Invariance:

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

(4) σ-Additivity: For every pairwise disjoint family $(A_j)_{j\in\mathbb{N}}$ of subsets
i.e. $i \neq j \Rightarrow A_i \cap A_j = \emptyset$ we have
$\mu(\uplus_{j\in J} A_j) = \sum_{j\in J} \mu(A_j)$

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

Problems with this Intuition of a Measure

Theorem by Vitali: There are no measures! [Vit05].
Functions satisfying our conditions (1), (2), (3), (4) do not exist.

Paradox of Banach-Tarski: [BT24], [Tao10], [Str79].
The unit ball in $\mathbb{R}^3$, i.e. $\mathbb{B}_3 = \{\vec{x} \in \mathbb{R}^3 \mid \|\vec{x}\| = 1\}$

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 $\{\vec{x} \mid \|\vec{x}\| = 2\}$

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

"Repairing" Measure Theory

Dropping requirements for a measure by restricting the existence of a measure to certain *measurable* sets.

Consequence 1: Can no longer prove the Theorem by Vitali.
Thus: Measures do exist.

Consequence 2: Banach-Tarski paradoxon gets an explanation.
From the 5 subsets in the paradoxon, not all are measurable.
Blow-up of the volume happens with non-measurable sets.

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*: $\forall (A_j \in \mathcal{A})_{j \in \mathbb{N}} \colon \cup_{j \in \mathbb{N}} \in \mathcal{A}$

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 Countably Infinite

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

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": [Hal13], [Coh13], [Tao11].

- $\mathcal{A}$ is well-defined ("smallest") and μ can be uniquely extended from intervals to $\mathcal{A}$.
- Can be extended to $\mathbb{R}^n$ using "cubes" and to topological spaces.
- Concepts of integrals and density functions may be introduced.

5. Shannon Information Theory

5.1. Probability

5.2. Conditional Probability

5.3. 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 [Haj19].

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**.

Proposition: 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: $\Omega = [0,5] \qquad \mu([a,b]) = b-a \qquad \mu(\Omega) = 5$ 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: $P([a,b]) = \int_a^b \varphi(x)dx = \Phi(b) - \Phi(a) \qquad \Phi(u) = \int_0^u \varphi(x)dx$

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

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:	Only consider events where the validity of a set B of properties is ensured.
Consequence:	Values no longer sum up to 1 but to smaller value.
Fix this:	Renormalize probability by scaling factor to still sum up to 1.

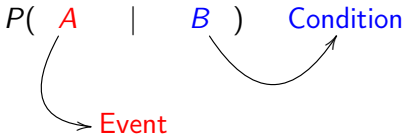
Let $(\Omega, \mathcal{A}, P)$ be a probability space and let $B \in \mathcal{A}$ with $P(B) \neq 0$.

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

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

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 $(\Omega, \mathcal{A})$.
Essential 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 $(B, \mathcal{A} \cap B)$.
Essential proof obligation: Show proper definition on base sets.

$p:$	$\mathcal{A}$	$\to$	$[0,1]$	original probability measure	
$p_{\|B}:$	$\mathcal{A}$	$\to$	$[0,1]$	modified measure on $\mathcal{A}$	(1)
$p_{\|B}:$	$\mathcal{A} \cap B$	$\to$	$[0,1]$	modified measure on modified algebra $\mathcal{A} \cap B \overset{id}{\hookrightarrow} \mathcal{A} \overset{p_{\|B}}{\longrightarrow} [0,1]$	(2)

These two interpretations of a conditional probability are theoretically different but practically nearly the same.

Thus we use identical notations.

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 Rules

Classical Bayes Rule: Swapping event and condition

$$P(A \mid B) = \frac{P(A)}{P(B)}\, P(B \mid A) \qquad \text{holds for } A, B \text{ with } P(A), P(B) \neq 0$$

Other forms and consequences of this:

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

Splitting Events on a Condition (1)

An event may be split on a single condition B

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

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

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

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

Now: This observation carries over to probabilities.

Now: Apply Bayes Chain Rule twice.

Splitting Events on a Condition (2)

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

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

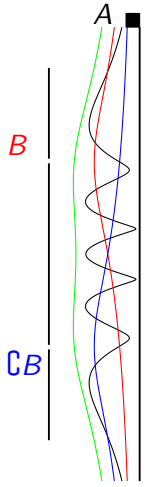
General case: Assume: $X_1, X_2, \ldots, X_n$ is a partition of Ω with $\forall i : P(X_i) > 0$.

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

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

Splitting Rule and Double Slit Experiment (1)

Shoot photons (Young) or electrons (Davisson-Germer) on a double-slit.



Classical theory expects green curve due to splitting rule.

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

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

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?

Speculations which of our assumptions leading to $P(A) = P(A)$ are not met by nature:

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 which we did not really measure have a definite value.
   Cannot theoretize on the value red might have while actually measuring blue.
4. **Distributivity:** Quantum logic is not distributive but needs an *orthomodular* law. [BVN36]
5. **Interference** as property of the wavelike nature of particles is neglected.

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.
Gives *better intuitive understanding* of independence,
but would be *worse definition* (works only if 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*:

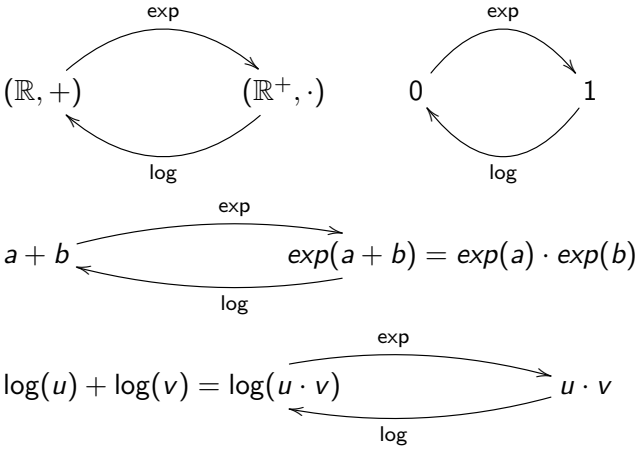
$$P(X \cap Y) = P(X) \cdot P(Y) \Rightarrow I(X \cap Y) = I(X) + I(Y)$$

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 probabilities *multiply*. Independent information *adds*.



6. Information Sources

6.1. Basic Definitions

6.2. Entropy and Redundancy

6.3. Examples

6.4. 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 [Haj19].

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 and Expectation Values

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

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

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

Random Variables and Conditional Expectation Values

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).

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

Note the two different but ultimately equal ways of defining this.

Toy Example: Dice as Information Source (1)

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

$A = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\} \qquad (\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot) \overset{p}{\mapsto} (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$

$(\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot) \overset{f}{\mapsto} (1,2,3,4,5,6) \qquad \mathcal{E}_{\mathcal{Q}}(f) = \mathcal{E}_{(A,p)}(f) = \vec{f} \cdot \vec{p} = \sum_{j=1}^{6} \frac{j}{6} = \frac{7}{2}$

$\textbf{Even} := \{\boxdot, \boxdot, \boxdot\} \qquad p(\textbf{Even}) = 1/2$

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

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

Toy Example: Dice as Information Source (2)

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

$$p_{|\mathbf{Even}}(\{\boxdot\}) \cdot f(\boxdot) + p_{|\mathbf{Even}}(\{\boxdot\}) \cdot f(\boxdot) + p_{|\mathbf{Even}}(\{\boxdot\}) \cdot f(\boxdot) +$$

$$p_{|\mathbf{Even}}(\{\boxdot\}) \cdot f(\boxdot) + p_{|\mathbf{Even}}(\{\boxdot\}) \cdot f(\boxdot) + p_{|\mathbf{Even}}(\{\boxdot\}) \cdot f(\boxdot)$$

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

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.

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

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: $A = \{0, 1\}$.

Two probabilities: $P(0)$ and $P(1)$.

Condition: $P(0) + P(1) = 1$.

One parameter: $P(0) =: q$
$P(1) = 1 - P(0) = (1 - q)$

Entropy:

$$H(q) = -q \log_2(q) - (1 - q) \log_2(1 - q)$$

Maximal Entropy:
At $q = P(0) = P(1) = 1/2$
Value: $H_{max}(2) = \log_2(2) = 1$.

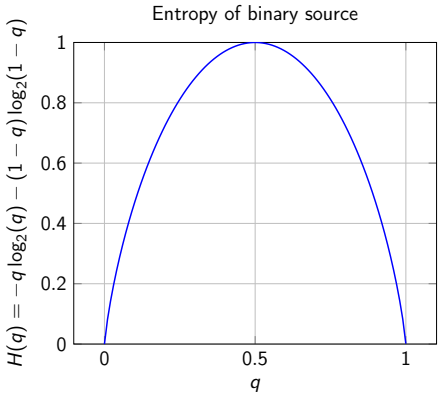


Fig. 1: Entropy of binary source as 1-parameter object.

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

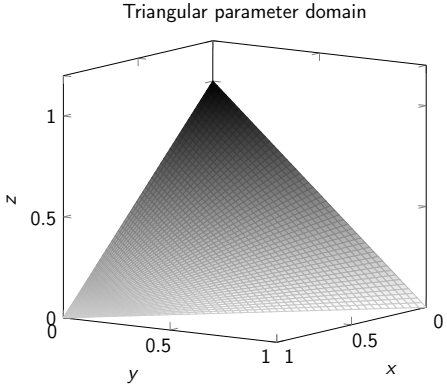


Fig. 2: Twodimensional triangular parameter domain of ternary sources as a plane in three-dimensional space.

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 $P(0) = P(1) = P(2) = 1/3$ Value;
$H_{max}(3) = log_2(3) = 1.585\ldots$

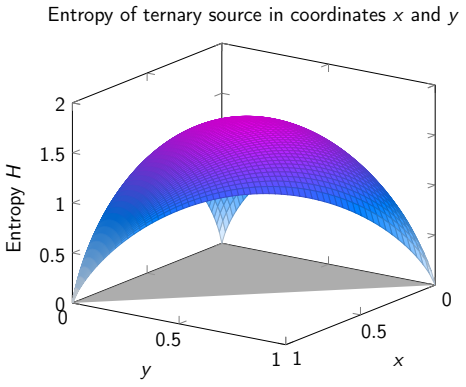


Fig. 3: Entropy of ternary source, x-y coordinates.

Example: Recoding Ternary Sources (1)

Let $A = \{a, b, c\}$ be the alphabet a ternary information source.

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

Idea: Use coding $a \mapsto 00 \quad b \mapsto 10 \quad c \mapsto 11$

Assume: Probabilities of the source:

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

Observe: Average length of recoded word: $2x + 2y + 2(1 - x - y-) = 2$.

Question: Can we do better?

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 unique decoding.

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

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

Example: 0001110 decodes as aaacb

Counterex: The coding $a \mapsto 1 \quad b \mapsto 10 \quad c \mapsto 11$ is not prefix-free
11 would decode as c or as aa.

Example: Recoding Ternary Sources (2)

Idea: Consider the coding $a \mapsto 0 \quad b \mapsto 10 \quad c \mapsto 11$

Check: The coding is prefix-free: ✓

Assume: Probabilities of the source:

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 whose scalars comprise $\mathbb{R}$ 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:

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

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

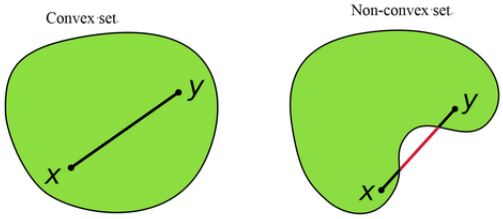


Fig. 4: Convex and non-convex set.

Convex Hull and Extreme Points

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 whose scalars comprise $\mathbb{R}$ is the following 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 for the convex hull:

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

Convex Functions

Let V be a vector space whose scalars comprise $\mathbb{R}$

A function $f : V \to \mathbb{R}$ is called

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

Classification of functions:

1. Convex
2. Concave
3. Others

Note 1: Convex and concave are **dual** to each other.

Note 2: Concave = not-convex is **simply wrong**.

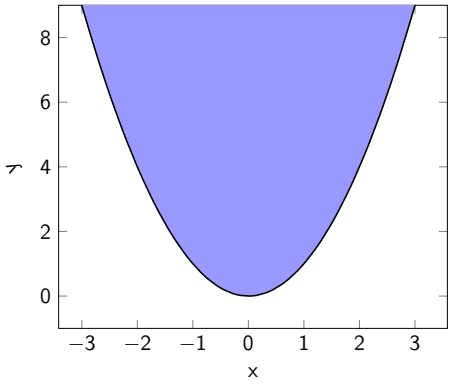


Fig. 5: The **epigraph** of a function consists of the graph and all points "above": $\text{epi}(f) := \{(x, y) \mid x \in \text{dom}(f) \wedge y \geq f(x)\}$. Obviously, this function **is convex**.

Criterion for Convex Function

Definition: $f: V \to \mathbb{R}$ convex, iff epigraph convex.

Criterion: $f: V \to \mathbb{R}$ is convex iff

$$\forall \vec{x_1}, \vec{x_2} \in V: \forall p_1, p_2 \in [0,1], p_1 + p_2 = 1$$

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

Question: Can this be generalized? Maybe to:

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

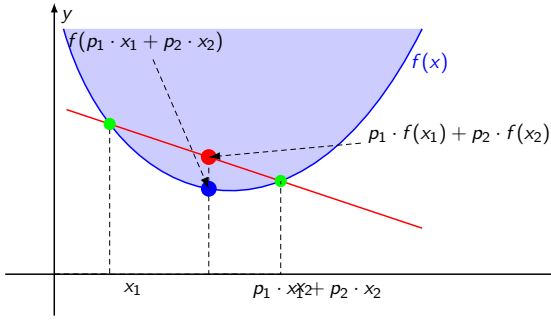


Fig. 6: Convex function and inequalities: The red dot is above the blue dot. As f is convex the epigraph (above the blue line) is convex. Thus the points on the red line between the two green dots are in the epigraph. Thus the red dot in the epigraph is above the blue dot on its boundary.

Theorem: Jensen Inequality

When $f: V \to \mathbb{R}$ is convex, then: For $p_i \geq 0$ with $\sum p_i = 1$

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

Jensen Inequality

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 in Mathematics

Convex functions defined **over convex** sets:

- **Maxima:** If existant lie on the boundaries of the convex set.
- **Minima:** A Local minimum is also a global minimum.
- Many more nice properties.

Krein-Milman Theorem: Convex sets are (often) the *convex hull of their extreme points*.

- **Math:** Only need to know the extremal points of convex sets.
- **Physics:** Only need to study pure states.

Many more technically important results.

Convexity and Information Theory

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 $\sum_i P(a_i) = 1$.

Classical probability is (pretty much exactly) convex geometry.

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

- $\mathfrak{I}$ is **convex** and an $(n-1)$-dimensional **simplex** in $\mathbb{R}^n$.
- $\mathfrak{I}$ is the **convex hull** of its corners: Knowing the corners means knowing the set.
- **Entropy** function on $\mathfrak{I}$ is **concave**.
- **Negentropy** (*neg*ative *entropy*) on $\mathfrak{I}$ is **convex**
- Negentropy **maximal at extremals** of $\mathfrak{I}$, **local minimum** in interior is **global**.
- Entropy **minimal at extremals** of $\mathfrak{I}$, **local maximum** in interior is **global**.

Example: Ternary Source as Convex Object

Observations:

1. Three corners are the **extremals**.
2. **State space** is their convex hull.
3. **Entropy** is maximal in an inner point.
4. **Negentropy** is maximal in the extremals,

Interpretations:

High negentropy = more order
High entropy = more disorder
= more information

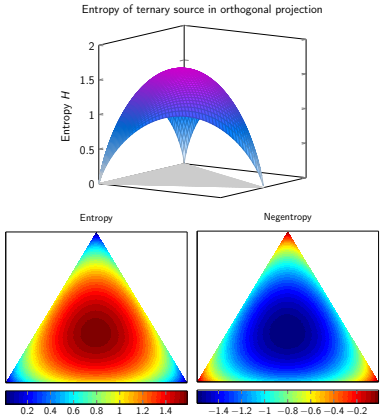


Fig. 7: Entropy and negentropy of ternary source.

Convextity and Quantum Theory

Density operators:	Form a **convex set**.
State is **pure:**	Iff it is an extremal point in the set of density operators.
State is **mixed:**	Iff it is a convex combination of pure states.
State is **separable:**	Iff it can be written as convex combination of product states.

Fundamental Differences in Theories

State:

- **Classical:** $0.3 \cdot a + 0.7 \cdot b$ not state, string or character.
  Merely an abstract, stochastically mixed information source.
- **Quantum:** $(1/\sqrt{2}) \cdot a + (i/\sqrt{2}) \cdot b$ pure physical state
  Not a stochastic mixture

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 Two significantly different concepts of state combination.

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

7. 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.

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 $\mathcal{S}_A \times \mathcal{S}_B := (A \times B, p_{\alpha\otimes\beta})$

where the measure $p_{\alpha\otimes\beta} = \alpha \otimes \beta$ 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 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

First source: $\mathcal{S}_A = (A, \alpha) \quad A := \{a_1, \ldots, a_n\} \quad \alpha_i := \alpha(\{a_i\})$

Second source: $\mathcal{S}_B = (B, \beta) \quad B := \{b_1, \ldots, b_m\} \quad \beta_j := \beta(\{b_j\})$

Product: $A \times B = \{(a_1, b_1), (a_1, b_2), \ldots, (a_1, b_m), (a_2, b_1), \ldots, \ldots, (a_n, b_m)\}$

$p_{ij} = p\ (\ \{\ (a_i, b_j)\ \}\ ) = \alpha_i \cdot \beta_j$ probabilities defined by **product** (independence)

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

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.

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

and the probabilities are **arbitrary**.

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

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

$$\begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \end{array} \begin{array}{c} \begin{array}{ccc} b_1 & b_2 & b_3 \end{array} \\ \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} \end{array} \quad \begin{array}{l} p_{1\bullet} = p_{11} + p_{12} + p_{13} = p_A(a_1) = p(\{\ (a_1, b_1),\ (a_1, b_2),\ (a_1, b_3)\ \}) \\ p_{2\bullet} = p_{21} + p_{22} + p_{23} = p_A(a_2) = p(\{\ (a_2, b_1),\ (a_2, b_2),\ (a_2, b_3)\ \}) \\ p_{3\bullet} = p_{31} + p_{32} + p_{33} = p_A(a_3) = p(\{\ (a_3, b_1),\ (a_3, b_2),\ (a_3, b_3)\ \}) \end{array}$$

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

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

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

Defined by *summing up to the matrix margin*

Definition: Marginals as Sums

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 Notation in Probability

Warning: Probability theory often heavily abuses and breaks notational clarity.

Remember: Introduce $p(\texttt{Set} : X)$ as $p\colon \mathcal{A} \subseteq 2^{\Omega} \to [0,1]$
Then use: $p(A \mid B)$ where $A \mid B$ is not a set.

Here again: Define a 2-variable function $p\colon A \times B \to [0,1]$ with $p(a,b)$
Write $p(a)$ instead of $p_A(a) = p(\ \{a\} \times B\ )$
Write $p(b)$ instead of $p_B(b) = p(\ A \times \{b\}\ )$

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

Set notation: Buys clarity at the expense of more brackets. 👍
Is always unambiguous. 👍
As in $p(\ \{a_1\} \times B\ )$ or $p(\ \{\sigma\} \times B \mid A \times \{\lambda\}\ )$

Notation: Conventions for Compounds

Shorthand notation:

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

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

Conditionals in **set notation**:

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

Conditionals in **shorthand notation**:

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

Same syntax as for single source
completely *different semantics*
division by marginal

Connection between Conditionals, Joints & 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:

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

Proof: Conditionals, Joints & Marginals

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:

$$
\begin{aligned}
&p(a,b) = \text{ go to set notation} \\
&= p(\ \{(a,b)\}\ ) \\
&= p\Big(\ (\ \{a\} \times B\ )\ \cap\ (\ A \times \{b\}\ )\ \Big) = \\
&\text{use definition of conditional } p\Big(\mathrm{X} \cap \mathrm{Y}\Big) = p\Big(\mathrm{X} \mid \mathrm{Y}\Big) \cdot p\Big(\mathrm{Y}\Big) \\
&= p\Big(\{a\} \times B \mid A \times \{b\}\Big) \cdot p\Big(A \times \{b\}\Big) = \text{ go back to shorthand notation} \\
&= p(a|b) \cdot p_B(b)
\end{aligned}
$$

Solvable Technical Problem with Marginals

Just as a short reminder...that we can cheat conceptually only in the finite cases.

Problem: Strictly, a compound is **not** $p\colon A \times B \to [0,1]$
but rather $p\colon \mathcal{C} \to [0,1]$ with $\mathcal{C} \subseteq A \times B$ a suitable σ-algebra

Solution: With finite A and B,we can use $\mathcal{C} := 2^{A \times B}$
For $U \subseteq A \times B$ we get a finite sum $p(U) = \sum_{u \in U} p(\{u\})$.

So how would we define marginals in the more general case?

Alternative Definition 1: Marginals as Compositions

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

Marginals are **compositions**

$\pi_A \colon A \times B \to A$

$$A \times B \xrightarrow{\pi_A} A \qquad (a,b) \longmapsto a$$

$$A \xrightarrow{\pi_A^{-1}} 2^{A\times B} \qquad a \longmapsto (\{a\} \times B)$$

$$2^A \xrightarrow{\pi_A^{-1}} 2^{A\times B} \qquad U \longmapsto (U \times B)$$

$$2^A \xrightarrow{\pi_A^{-1}} 2^{A\times B} \xrightarrow{p} [0,1] \qquad U \longmapsto U \times B \longmapsto p(U \times B)$$

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

Note 1: We still did not provide proper σ-algebra conditions.

Note 2: Better definition. More general. Works in σ-algebra situations.

Expectation Values: Extension to Vector Values

If $q\colon B \to [0,1]$ is a probability and $f\colon B \to \mathbb{R}$ a real function we can define an expectation value

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

This can be generalized from $\mathbb{R}$ to arbitrary real vector spaces V.
So let $f\colon B \to V$ be a vector valued function.

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

Note: This also can be done for σ-algebra situations.

Reinterpreting: Partial Conditionals as Vectors

Step 1: Consider **conditional** probabilities:

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

Step 2: Consider them as **vector-valued** functions of one (here: first, A) variable:

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

For fixed $a \in A$ the function $p(a \mid \cdot)\colon B \to [0,1]$
may be considered the vector $p(a \mid \cdot)$ of conditional probabilities

$$p(a \mid b_1), p(a \mid b_2), \ldots, p(a \mid b_n)$$

Marginals as Expectation Value of Conditionals

Given the function (vector) $p(a \mid \cdot)$ of conditional probabilities

$$p(a \mid b_1), p(a \mid b_2), \ldots, p(a \mid b_n)$$

and given for every condition, which here is b_j, its (marginal) probability $p_B(b_j)$.

We get as **expectation value** of the **vector of conditional probabilities**:

$$\mathcal{E}_{p_B}(p(a \mid \cdot)) = \sum_{b \in B} p_B(b) \cdot p(a|b)$$

But this equals $p_A(a)$.

The A-**marginals** are the **expectation values** of the **conditional probabilities** weighted by the B-marginals.

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.

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

Product: Probability depends only on color and size alone
Compound: There might be 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 1: $\{a, b, c, d\}$ (bad example, as it indicates a specific factorization)
Example 2: $\{a, e, i, u\}$ (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".

Some concepts of quantum information can be abstractly reformulated in classical language – despite the **big** conceptual differences in some aspects.

Factorizables versus Compounds in Physics Speculation!)

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

Now:	We analyze a quantum experiment as classical compound.
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 $A := \|\vec{a}_1\rangle\langle\vec{a}_1\| - \|\vec{a}_2\rangle\langle\vec{a}_2\|$ $\quad$ $B := \|\vec{b}_1\rangle\langle\vec{b}_1\| - \|\vec{b}_2\rangle\langle\vec{b}_2\|$
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

Probabilities suggested by quantum theory and confirmed by experiment:

	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: Joint and marginal probabilities of the "Bell" compound source.

Perspective 1: θ is a parameter.

Perspective 2: θ is angle between the real, 3-dim Bloch vectors belonging to A and B.

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.
Note 2: We thus **only study** (a_1, b_1) and (a_2, b_1).
Note 3: **Marginals are independent** of θ and symmetric (always $1/2$)
Note 4: 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.

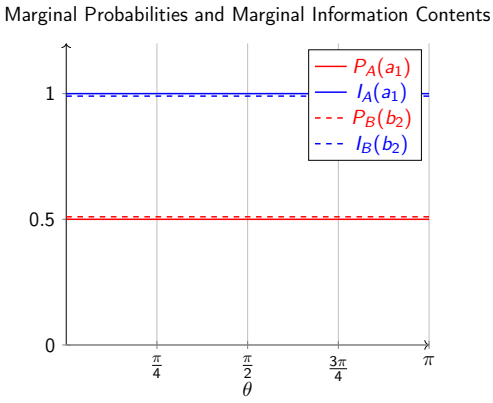


Fig. 8: Marginal probabilities (red) and marginal information contents (blue) of the "Bell" compound source are independent of the parameter θ.

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?

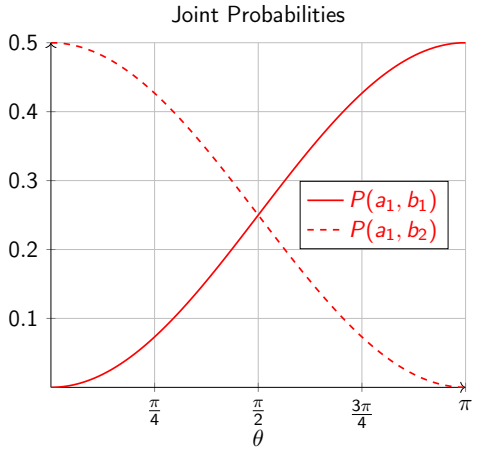


Fig. 9: Joint probabilities (red). Dashed versions shows a different pair.

Joints (Using Graphs)

Observations:

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

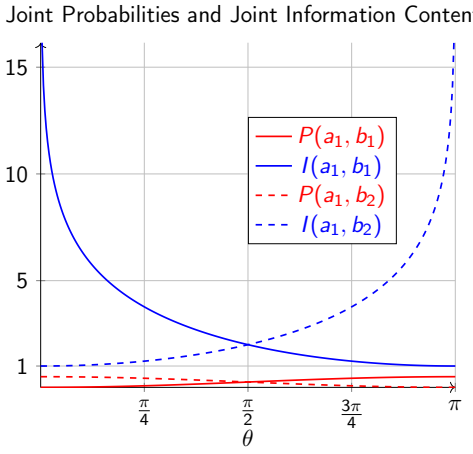


Fig. 10: Joint probabilities (red) and joint information contents (blue) of the "Bell" compound source. Dashed versions show a different pair.

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.

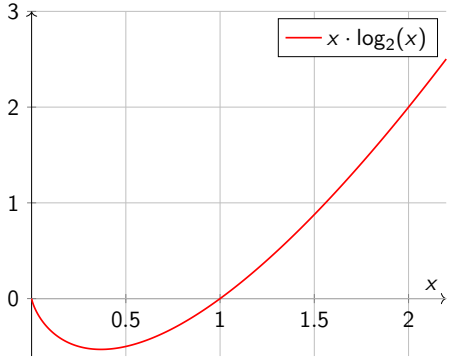
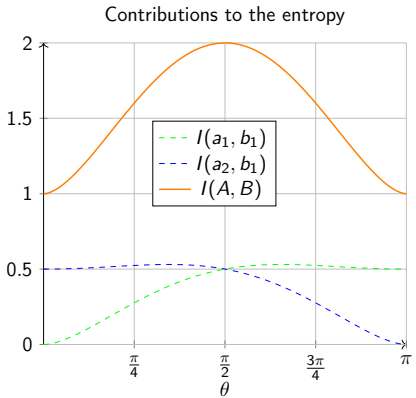


Fig. 11: Additive contribution of a symbol to the entropy.

Total Contributions of Pairs to Entropy



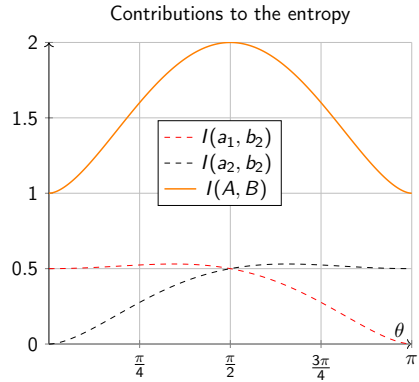


Fig. 12: Contributions of the four pairs (a_1, b_1), (a_1, b_2), (a_2, b_1) and (a_2, b_2) to the to the total entropy of the source.

Relative Contributions of Pairs to Entropy

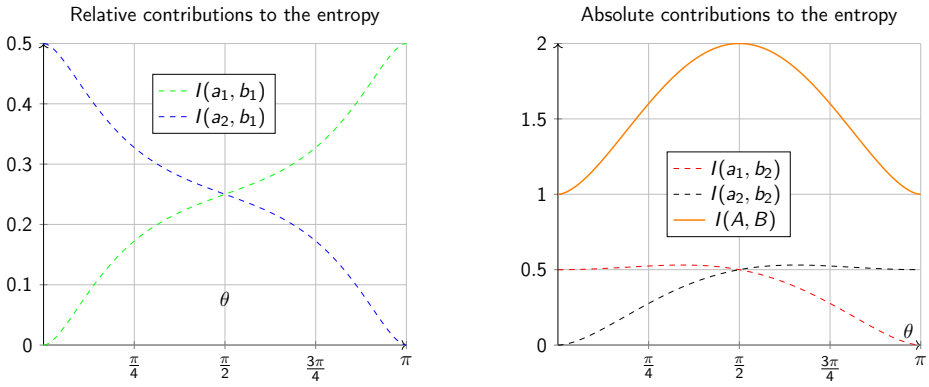


Fig. 13: Absolute and relative contributions of the pairs to the total entropy of the source.

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

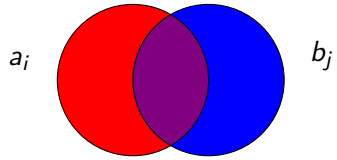


Fig. 14: Venn diagram for two sets motivates the definition of an overlap.

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

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

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 $;$ versus $,$ (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.

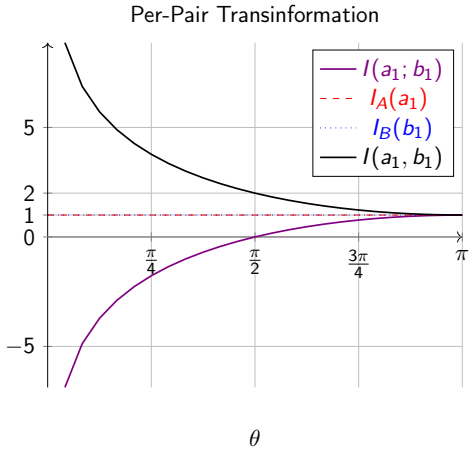


Fig. 15: Per-pair transinformation for the Bell example.
$I(a_i\ ;\ b_j) := I_A(a_i) + I_B(b_j) - I(a_i\ ,\ b_j)$

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 122.

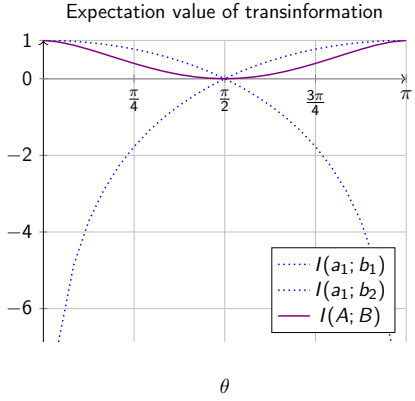


Fig. 16: The expectation value of the transinformation is non-negative, although the contribution of some individual pairs may be negative.

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

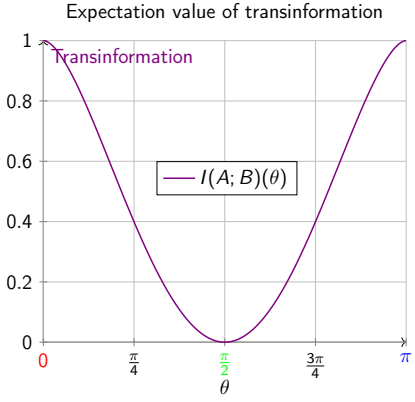


Fig. 17: The expectation value of the transinformation in a better magnified plot.

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: (Expected) transinformation is non-negative.

Proof:

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

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

Bibliography

[BR11] Luc Bovens and Wlodek Rabinowicz.
Bets on hats: on Dutch books against groups, degrees of belief as betting rates, and group-reflection.
Episteme, 8(3):281–300, 2011.
URL: http://eprints.lse.ac.uk/49667/1/Bovens_Bets_on_hats_2011.pdf.
DoI: 10.3366/epi.2011.0022.
5

[BT24] Stefan Banach and Alfred Tarski.
Sur la décomposition des ensembles de points en parties respectivement congruentes.
Fund. math, 6(1):244–277, 1924.
http://kielich.amu.edu.pl/Stefan_Banach/pdf/oeuvres1/12.pdf.
36

[BVN36] Garrett Birkhoff and John Von Neumann.
The Logic of Quantum Mechanics.
Annals of mathematics, pages 823–843, 1936.
URL: https://www.cs.huji.ac.il/~lehmann/nonmon/Birkhoff_vonNeumann.pdf.
52

[Cha66] Gregory J Chaitin.
On the length of programs for computing finite binary sequences.
Journal of the ACM (JACM), 13(4):547–569, 1966.
29

[Cha87] Gregory. J. Chaitin.
Algorithmic Information Theory.
Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, 1987.
DoI: 10.1017/CBO9780511608858.
27, 29

[Coh13] Donald L Cohn.
Measure theory.
Springer, 2013.
40

[Deu04] David Deutsch.
It from qubit.
Science and Ultimate Reality. Quantum Theory, Cosmology, and Complexity, pages 90–102, 2004.
https://www.daviddeutsch.org.uk/wp-content/ItFromQubit.pdf.
4

[Dum76] Michael Dummett.
Is Logic Empirical?
Contemporary British Philosophy, (4):45–68, 1976.
11

[Haj19] Alan Hajek.
Interpretations of Probability.
Stanford Encyclopedia of Philosophy, 2002 and 2019.
URL: https://plato.stanford.edu/entries/probability-interpret/.
Version of August 28, 2019.
5, 42, 57

[Hal13] Paul R Halmos.
Measure theory, volume 18.
Springer, 2013.
40

[Kol68] Andrei Nikolaevich Kolmogorov.
Three approaches to the quantitative definition of information.
International journal of computer mathematics, 2(1-4):157–168, 1968.
29

[Lan91] Rolf Landauer.
Information is Physical.
Physics Today, (44 (5)):23–29, 1991.
URL: https://doi.org/10.1063/1.881299.
https://doi.org/10.1063/1.881299.
4

[Odi92] Piergiorgio Odifreddi.
Classical recursion theory: The theory of functions and sets of natural numbers.
Elsevier, 1992.
27

[Pea09] Judea Pearl.
Causality.
Cambridge University Press, 2009.
5

[Pop59] Karl R Popper.
The propensity interpretation of probability.
The British journal for the philosophy of science, 10(37):25–42, 1959.
5

[Put68] Hilary Putnam.
Is Logic Empirical?
Boston Studies in the Philosophy of Science, (5):216–241, 1968.
11

[Ram16] Frank P Ramsey.
Truth and probability.
In *Readings in Formal Epistemology*, pages 21–45. Springer, 2016.
URL: https://www.sapili.org/subir-depois/en/mc000224.pdf.
5

[Sol64a] Ray J Solomonoff.
A formal theory of inductive inference. Part I.
Information and control, 7(1):1–22, 1964.
https://doi.org/10.1016/S0019-9958(64)90223-2.
29

[Sol64b] Ray J Solomonoff.
A formal theory of inductive inference. Part II.
Information and control, 7(2):224–254, 1964.
https://core.ac.uk/reader/81988200.
29

[Str79] Karl Stromberg.
The Banach-Tarski paradox.
The American Mathematical Monthly, 86(3):151–161, 1979.
https://pdfs.semanticscholar.org/3bf8/25626beb94f940db7e355973cb3a4587042e.pdf.
36

[STZDG14] Fernando Soler-Toscano, Hector Zenil, Jean-Paul Delahaye, and Nicolas Gauvrit.
Calculating Kolmogorov complexity from the output frequency distributions of small Turing machines.
PloS one, 9(5):e96223, 2014.
31

[Tal08] William Talbott.
Bayesian Epistemology.
Standford Encyclopedia of Philosophy, 2001, 2008.
URL: https://plato.stanford.edu/entries/epistemology-bayesian/index.html.
5

[Tao10] Terence Tao.
An epsilon of room, I: real analysis, volume 1.
American Mathematical Soc., 2010.
36

[Tao11] Terence Tao.
An introduction to measure theory.
2011.
40

[Vit05] Giuseppe Vitali.
Sul problema della misura dei Gruppi di punti di una retta: Nota.
Tip. Gamberini e Parmeggiani, 1905.
36

[Whe89] John Archibald Wheeler.
Information, Physics, Quantum: The Search for Links.
In *Proceedings III International Symposium on Foundations of Quantum Mechanics*, pages 354–358. 1989.
4

[Whe90] John A Wheeler.
A journey into gravity and spacetime.
jigs, 1990.
4

[Whi72] Alan R White.
The propensity theory of probability.
The British Journal for the Philosophy of Science, 23(1):35–43, 1972.
5

[Wie61] Norbert Wiener.
Cybernetics or Control and Communication in the Animal and the Machine.
MIT Press, second edition, paper back. edition, 1948, 1961.
ISBN ISBN 0-262-23007-0.
4