Title: Causal inference rules for algorithmic dependences and why they reproduce the arrow of time

Date: Apr 12, 2018 02:00 PM

URL: http://pirsa.org/18040124

Abstract: The causal Markov condition relates statistical dependences to causality. Its relevance is meanwhile widely appreciated in machine learning, statistics, and physics. I describe the *algorithmic* causal Markov condition relating algorithmic dependences to causality, which can be used for inferring causal relations among single objects without referring to statistics. The underlying postulate "no algorithmic dependence without causal relation" extends Reichenbach's Principle to a probability-free setting. I argue that a related postulate called "algorithmic independence of initial state and dynamics" reproduces the non-decrease of entropy according to the thermodynamic arrow of time.

Causal inference rules for algorithmic dependences and why they reproduce the arrow of time

Dominik Janzing

Max Planck Institute for Intelligent Systems
Tübingen, Germany

On leave to
Amazon Research Tübingen

April 2018

Can we infer causal relations from passive observations?

Recent study reports negative correlation between coffee consumption and life expectancy

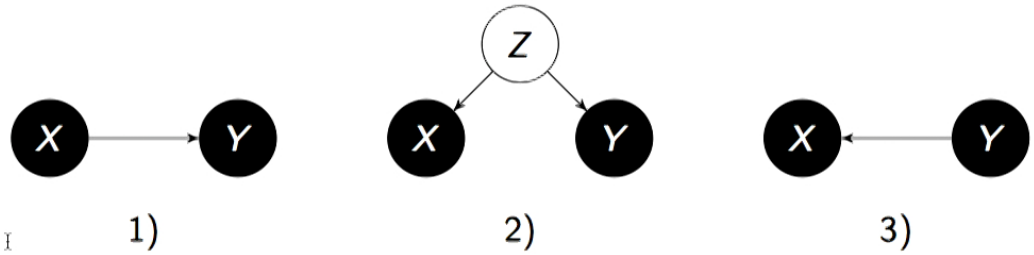
Paradox conclusion:

- drinking coffee is healthy
- nevertheless, strong coffee drinkers tend to die earlier because they tend to have unhealthy habits

$\Rightarrow$ Relation between statistical and causal dependences is tricky

Reichenbach's principle of common cause (1956)

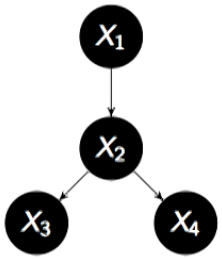
If two variables X and Y are statistically dependent then either



- in case 2) Reichenbach postulated $X \perp\!\!\!\perp Y \,|Z$.
- every statistical dependence is due to a causal relation, we also call 2) "causal".
- distinction between 3 cases is a key problem in scientific reasoning.

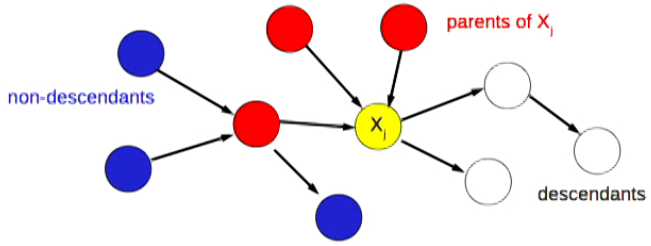
Causal inference problem, general form Spirtes, Glymour, Scheines, Pearl

- Given variables $X_1, \ldots, X_n$
- infer causal structure among them from n-tuples iid drawn from $P(X_1, \ldots, X_n)$
- causal structure = directed acyclic graph (DAG)



Causal Markov condition (3 equivalent versions) Lauritzen et al

- **local Markov condition:** every node is conditionally independent of its non-descendants, given its parents



- **global Markov condition:** If the sets S, T of nodes are d-separated by the set R, then

$$S \perp\!\!\!\perp T \,|\, R\,.$$

- **factorization of joint density:** $p(x_1, \ldots, x_n) = \prod_j p(x_j|pa_j)$ (subject to a technical condition)

Relevance of Markov conditions

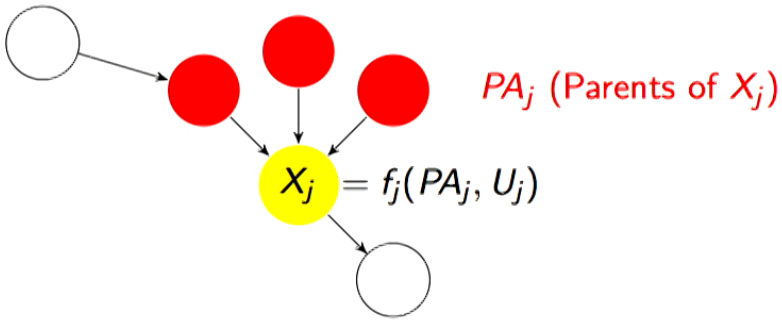
- **local Markov condition:** Most intuitive form, formalizes that every information exchange with non-descendants involves the parents

- **global Markov condition:** graphical criterion describing all independences that follow from the ones postulated by the local Markov condition

- **factorization:** every conditional $p(x_j|pa_j)$ describes a causal mechanism

Justification: Functional model of causality Pearl,...

- every node X_j is a function of its parents and an unobserved noise term U_j



(although this is not justified in quantum theory)

- all noise terms U_j are statistically independent (causal sufficiency)

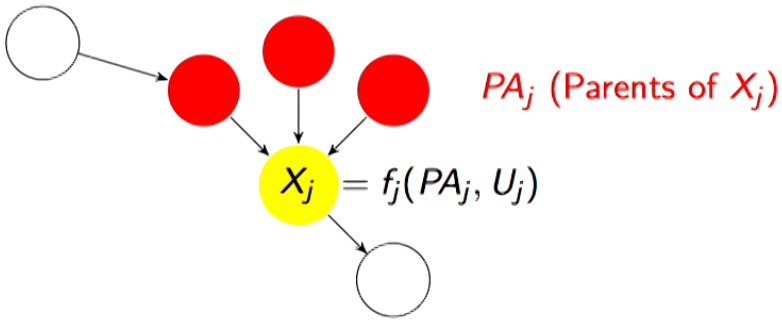
Functional model implies Markov condition

Theorem (Pearl 2000)

If $P(X_1, \ldots, X_n)$ is generated by a functional model according to a DAG G, then it satisfies the 3 equivalent Markov conditions with respect to G.

Justification: Functional model of causality Pearl,...

- every node X_j is a function of its parents and an unobserved noise term U_j



(although this is not justified in quantum theory)

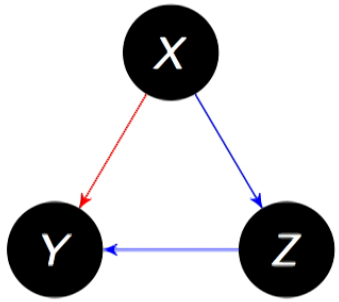
- all noise terms U_j are statistically independent (causal sufficiency)

Causal faithfulness

Spirtes, Glymour, Scheines, 1993

Prefer those DAGs for which all observed conditional independences are implied by the Markov condition

- **Idea:** generic choices of parameters yield faithful distributions
- **Example:** let $X \perp\!\!\!\perp Y$ for the DAG

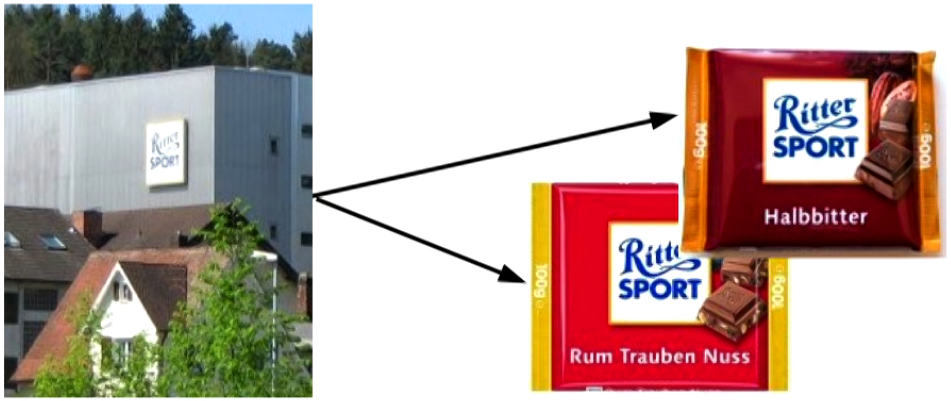


- not faithful, direct and indirect influence compensate
- **Application:** PC and FCI infer causal structure from conditional statistical independences

Causal inference for individual objects

– *why* are these objects so similar?

Conclusion: common history



similarities require an *explanation*

what kind of similarities require an explanation?

- here we would *not* assume that anyone has copied the design because the pattern is too simple
- similarities require an explanation only if the pattern is sufficiently complex

11

consider a binary sequence

Experiment:
2 persons are instructed to write down a string with 1000 digits

Result:
Both write 1100100100001111110110101010001...
(all 1000 digits coincide)

the **naive** statistician concludes

"There must be an agreement between the subjects"

correlation coefficient 1 (between digits) is highly significant for sample size 1000 !

- reject statistical independence
- infer the existence of a causal relation

another mathematician recognizes...

$$11.001001000011111101101010100 1... = \pi$$

- subjects may have come up with this number independently because it follows from a simple law
- superficially strong similarities are not necessarily significant if the pattern is too simple

How do we measure simplicity versus complexity of patterns / objects?

Kolmogorov complexity

(Kolmogorov 1965, Chaitin 1966, Solomonoff 1964)
of a binary string x

- K(x) = length of the shortest program with output x (on a Turing machine)
- interpretation: number of bits required to describe the rule that generates x

  neglect string-independent additive constants; use $\stackrel{+}{=}$ instead of $=$
- strings x, y with low $K(x)$, $K(y)$ cannot have much in common
- $K(x)$ is uncomputable
- probability-free definition of information content

Conditional Kolmogorov complexity

- $K(y|x)$: length of the shortest program that generates y from the input x.
- number of bits required for describing y if x is given
- $K(y|x^*)$ length of the shortest program that generates y from x^*, i.e., the shortest compression x.
- subtle difference: x can be generated from x^* but not vice versa because there is no algorithmic way to find the shortest compression

Algorithmic mutual information

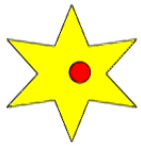
Chaitin, Gacs

Information of x about y (and vice versa)

- $I(x : y) := K(x) + K(y) - K(x, y)$
  $\stackrel{+}{=} K(x) - K(x|y^*) \stackrel{+}{=} K(y) - K(y|x^*)$

- Interpretation: number of bits saved when compressing x, y jointly rather than compressing them independently

Algorithmic mutual information: example

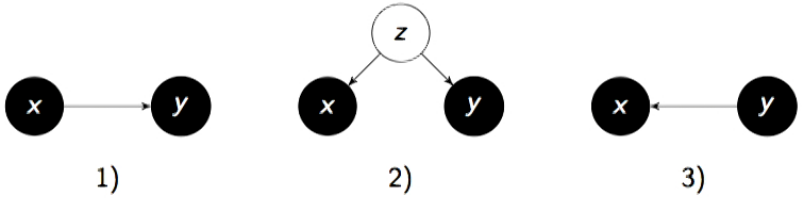
$I($  $:$  $) = K($  $)$

Analogy to statistics:

- replace strings x, y (=objects) with random variables X, Y
- replace Kolmogorov complexity with Shannon entropy
- replace algorithmic mutual information $I(x : y)$ with statistical mutual information $I(X; Y)$

Causal Principle

If two strings x and y are algorithmically dependent then either



- every algorithmic dependence is due to a causal relation
- algorithmic analog to Reichenbach's principle of common cause
- distinction between 3 cases: use conditional independences on more than 2 objects

DJ, Schölkopf IEEE TIT 2010

Relation to Solomonoff's universal prior

- string x occurs with probability $\sim 2^{-K(x)}$
- if generated independently, the pair (x, y) occurs with probability $\sim 2^{-K(x)}2^{-K(y)}$
- if generated jointly, it occurs with probability $\sim 2^{-K(x,y)}$
- hence $K(x, y) \ll K(x) + K(y)$ indicates generation in a joint process
- $I(x : y)$ quantifies the evidence for joint generation

conditional algorithmic mutual information

- $I(x : y|z^*) = K(x|z^*) + K(y|z^*) - K(x, y|z^*)$
- Information that x and y have in common when z is already given
- formal analogy to statistical mutual information:

$$I(X : Y|Z) = S(X|Z) + S(Y|Z) - S(X, Y|Z)$$

- Define conditional independence:

$$I(x : y|z^*) \approx 0 :\Leftrightarrow x \perp\!\!\!\perp y\,|z$$

Equivalence of algorithmic Markov conditions

Theorem (DJ & Schölkopf, IEEE TIT 2010)

For n strings $x_1, ..., x_n$ the following conditions are equivalent

- **Local Markov condition:**

$$I(x_j : nd_j | pa_j^*) \stackrel{+}{=} 0$$

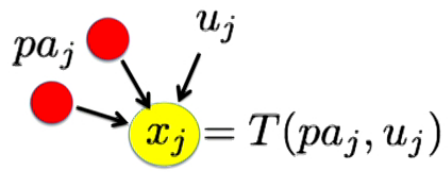
- **Global Markov condition:**
  R d-separates S and T implies $I(S : T|R^*) \stackrel{+}{=} 0$
- **Recursion formula for joint complexity**

$$K(x_1, ..., x_n) \stackrel{+}{=} \sum_{j=1}^{n} K(x_j | pa_j^*)$$

Justification: algorithmic model of causality

Given n causality related strings $x_1, \ldots, x_n$

- each x_j is computed from its parents pa_j and an unobserved string u_j by a Turing machine T



- all u_j are algorithmically independent
- each u_j describes the causal mechanism (the program) generating x_j from its parents
- u_j is the analog of the noise term in the statistical functional model

Relation to Church-Turing-Deutsch Principle

- **Church-Turing-Deutsch Principle:** Every physical process can be simulated on a Turing machine

- **Algorithmic model of causality:** Every distributed process can be simulated by multiple Turing machines whose communication structure resembles the underlying causal structure

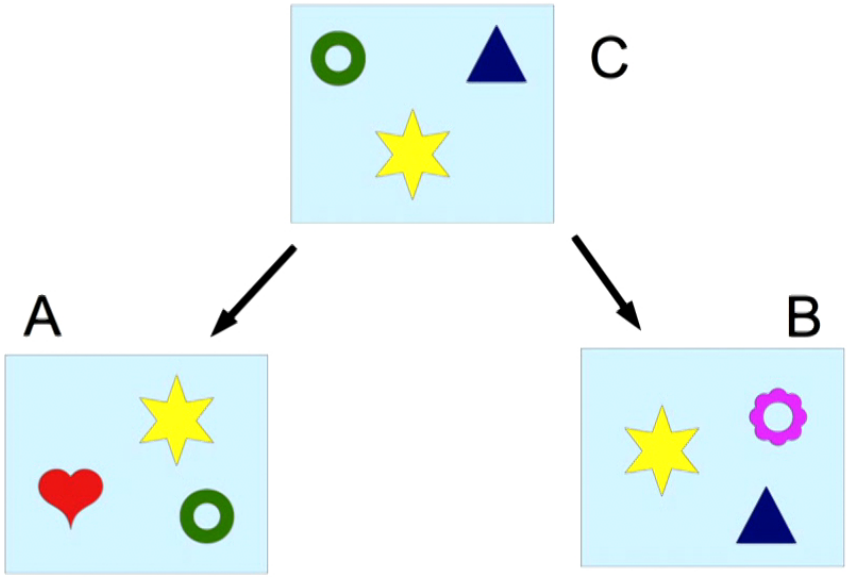
Algorithmic model of causality implies Markov condition

Theorem

If $x_1, \ldots, x_n$ are generated by an algorithmic model of causality according to the DAG G then they satisfy the 3 equivalent algorithmic Markov conditions.

Causal inference for single objects

3 carpets



conditional independence $A \perp\!\!\!\perp B \,||\, C$

How to deal with uncomputability

- **Use algorithmic Markov condition only as foundation for new statistical inference rules (later in this talk)**

- **Approximate K by existing compression schemes** (e.g. infer causal relations between texts by Lempel-Ziv compression. Steudel, DJ, Schölkopf COLT 2010)

  ...or try to *define* conditional independence by compression schemes (don't care whether they are good approximations)

Generalized notion of cond. independence Steudel, DJ, Schölkopf 2010

Given n objects $\mathcal{O} := \{x_1, \ldots, x_n\}$, any monotone function

$$R : 2^{\mathcal{O}} \to \mathbb{R}_0^+$$

that is submodular, i.e.,

$$R(A \cup B \cup C) - R(A \cup B) \leq R(B \cup C) - R(B)$$

(giving 1$ to a rich person increases happiness less than giving it to a poor person)

defines conditional dependence via

$$I(A : C|B) := R(A \cup C) + R(B \cup C) - R(A \cup B \cup C) - R(B) \geq 0$$

Theorem: generalized Markov conditions

for any DAG G, the following conditions are equivalent:

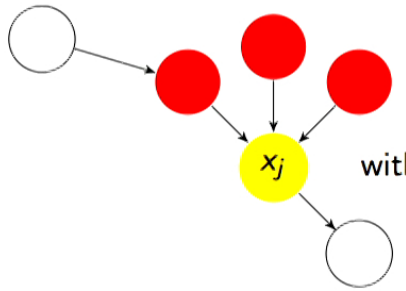
- **local Markov condition:** $x_j \perp\!\!\!\perp nd_j \,|paj$
- **global Markov condition:** d-separation implies independence
- **sum rule:** $R(A) = \sum_{j \in A} R(x_j|pa_j)$ for every ancestral set A of nodes.

but why should the conditions hold w.r.t. to the true causal DAG?

Justification via generalized functional model

the following conditions are sufficient for $x_1, \ldots, x_n$ to be Markovian:

- every objects contains only information from its parents and an unobserved 'noise' object



pa_j (parents of x_j)

with $R(x_j, pa_j, u_j) = R(pa_j, u_j)$

- noise objects are independent

$$R(u_1, \ldots, u_n) = \sum_{j=1}^{n} R(u_j)$$

Submodular information measures

and a corresponding functional model

- **number** of different words of a text
  author takes words from parent texts and from an independent source (e.g. his/her mind)

- **Lempel-Ziv** compression length (approximately submodular)
  concatenate substrings from parents and noise

- logarithm of **least common multiple** of a set of natural numbers (log of period length of a signal)
  periodic signal obtained by linear combination of parent and noise signals

Side remark

submodularity of information plays a role in **information causality** in the context of general probabilistic theories

see: Barnum & Wilce and Pawlowski & Scarani in *'Quantum Theory: Informational Foundations and Foils'*

(monotonicity is violated, however, already in QT)

New statistical inference principle using alg. information

Postulate (Algorithmic Independence of Conditionals)

If $P(X_1, \ldots, X_n)$ is generated by the causal DAG G, then the conditionals $P(X_j|PA_j)$ in the decomposition

$$P(X_1, \ldots, X_n) = \prod_{j=1}^{n} P(X_j|PA_j)$$

are algorithmically independent

DJ & Schölkopf 2010, Lemeire & DJ 2012

Special case: cause effect pair

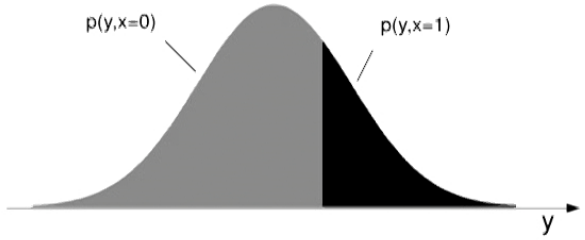
Independence of cause and mechanism (DJ & BS 2010)

P_{cause} and $P_{effect|cause}$ are algorithmically independent, i.e., the shortest description of $P_{cause,effect}$ is given by separate descriptions of P_{cause} and $P_{effect|cause}$.

Toy example

Let X be binary and Y real-valued.

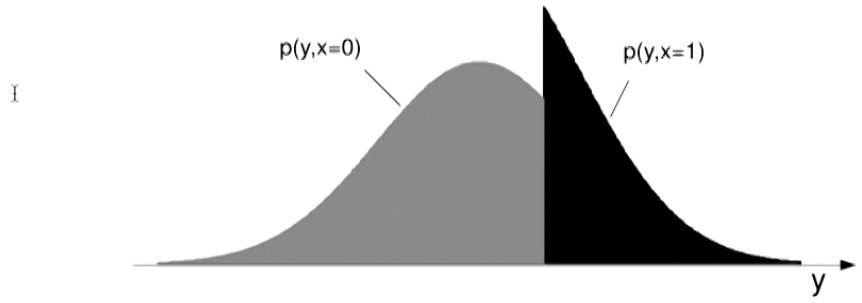
- Let Y be Gaussian and $X = 1$ for all y above some threshold and $X = 0$ otherwise.



- $Y \rightarrow X$ is plausible: simple thresholding mechanism
- $X \rightarrow Y$ requires a strange mechanism:
  look at $P(Y|X = 0)$ and $P(Y|X = 1)$!

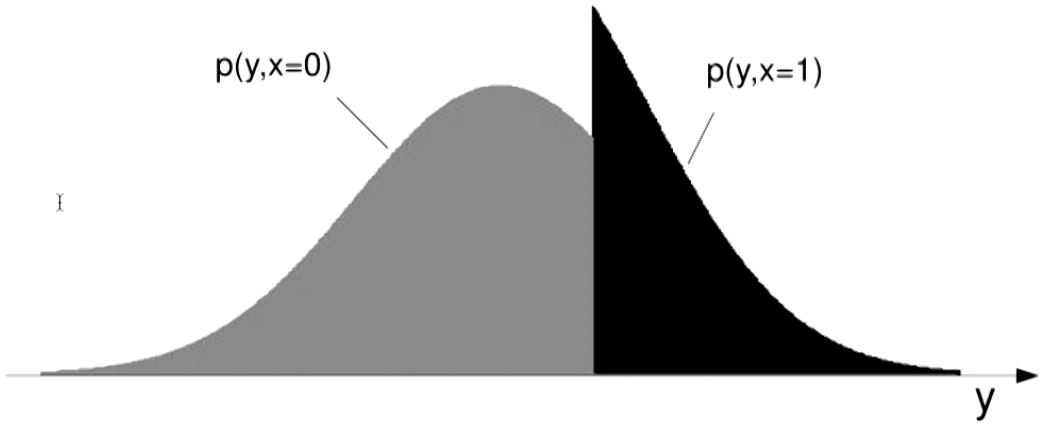
not only $P(Y|X)$ itself is strange...

but also what happens if we change $P(X)$:



Hence, reject $X \rightarrow Y$ because it requires tuning of $P(X)$ relative to $P(Y|X)$.

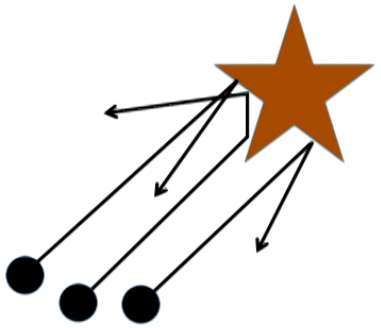
Violation of independence of conditionals



Knowing $P(Y|X)$, there is a short description of $P(X)$, namely 'the unique distribution for which $\sum_x P(Y|x)P(x)$ is Gaussian'.

Asymmetries between past and future

- **photographic images** show the past and not the future
- **particles are scattered at an object:** incoming beam contains no information about the object but maybe the outgoing beam does



- **any energy can be converted into heat** but not vice versa

...although the underlying micro-physical laws are symmetric under time-inversion

Independence of mechanisms as common principle

Postulate (DJ, Chaves, BS (2016))

If s is the initial state of a physical system and M a map describing the effect of applying the system dynamics for some fixed time, then s **and** *M* **are algorithmically independent**

$$I(s : M) \stackrel{+}{=} 0,$$

i.e., knowing s does not enable a shorter description of M and vice versa. Special case: $I(P_{\text{cause}} : P_{\text{effect}|\text{cause}}) \stackrel{+}{=} 0$

Reproduces thermodynamic arrow of time:

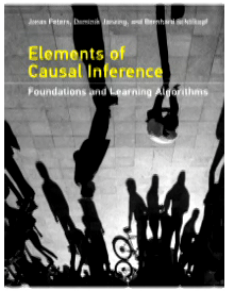
- for bijective M, $s' := M(s)$ cannot have a shorter description than s because $M^{-1}(s')$ is a description of s.
- Kolmogorov complexity (= physical entropy, according to Bennett, Zurek...) cannot decrease

Conclusions

- causal Markov condition holds for any submodular information measure that fits to the underlying causal mechanisms
- not only statistical but also algorithmic (in)dependences reveal causal information
- algorithmic independence of P_{cause} and $P_{\text{effect}|\text{caue}}$ is part of a general principle of independence of mechanisms
- principle explains asymmetry between cause and effect analog to asymmetry between past and future
  $\rightarrow$ is causal inference a kind of thermodynamics?

References

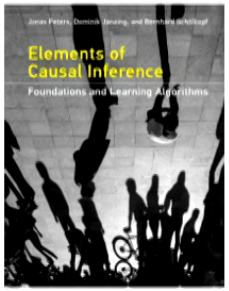
Our recent book can be downloaded at
`https://mitpress.mit.edu/books/elements-causal-inference`



- Janzing, Schölkopf: **Causal inference using the algorithmic Markov condition.** IEEE TIT (2010).
- Lemeire, Janzing: **Replacing causal faithfulness with the algorithmic independence of conditionals**, Minds & Machines (2012).
- Steudel, Janzing, Schölkopf: **Causal Markov condition for submodular information measures.** COLT (2010)

References

Our recent book can be downloaded at
`https://mitpress.mit.edu/books/elements-causal-inference`



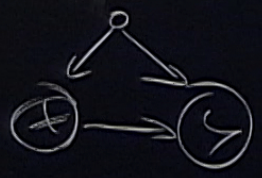
- Janzing, Schölkopf: **Causal inference using the algorithmic Markov condition.** IEEE TIT (2010).
- Lemeire, Janzing: **Replacing causal faithfulness with the algorithmic independence of conditionals**, Minds & Machines (2012).
- Steudel, Janzing, Schölkopf: **Causal Markov condition for submodular information measures.** COLT (2010)

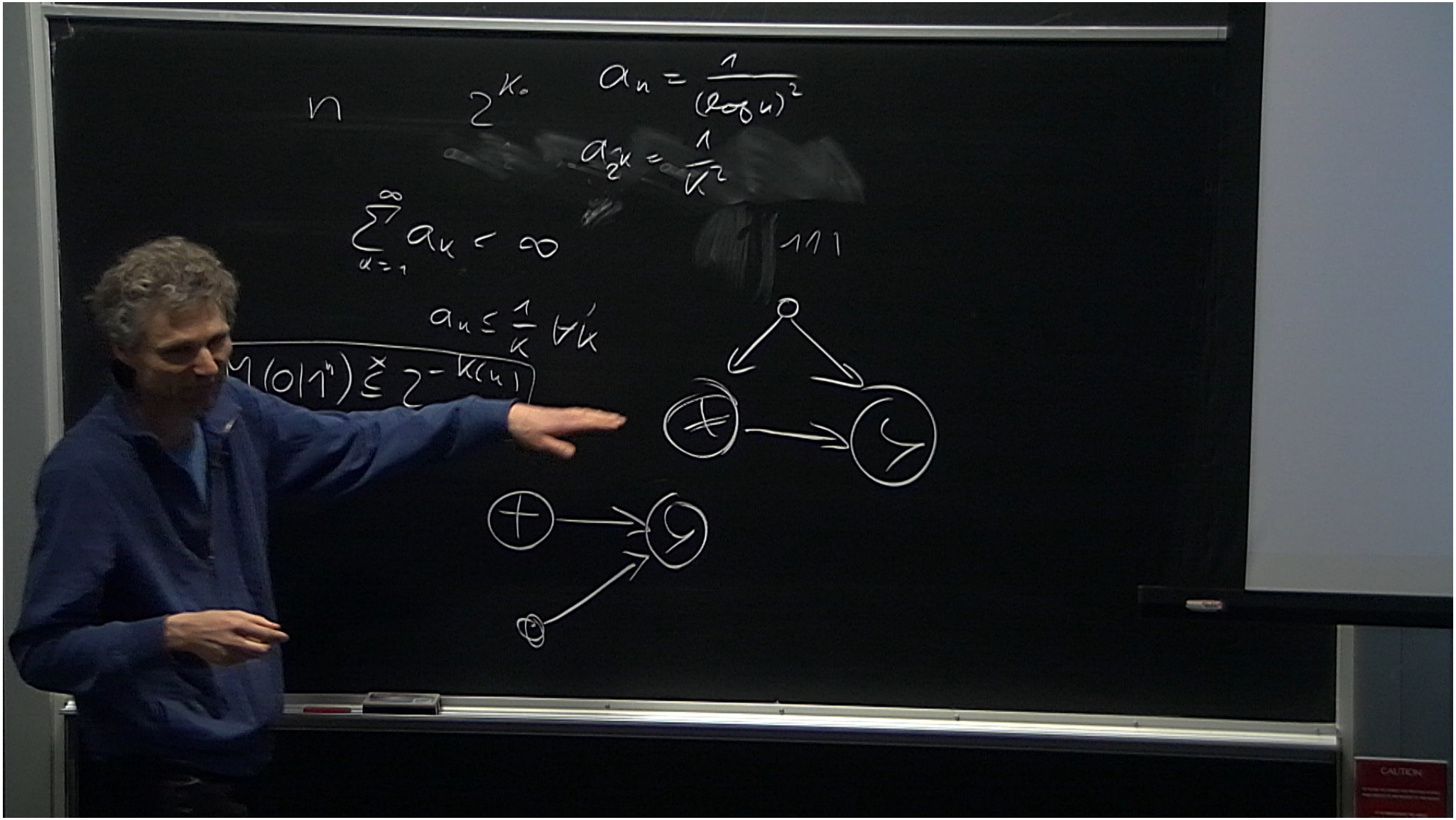
Submodular information measures

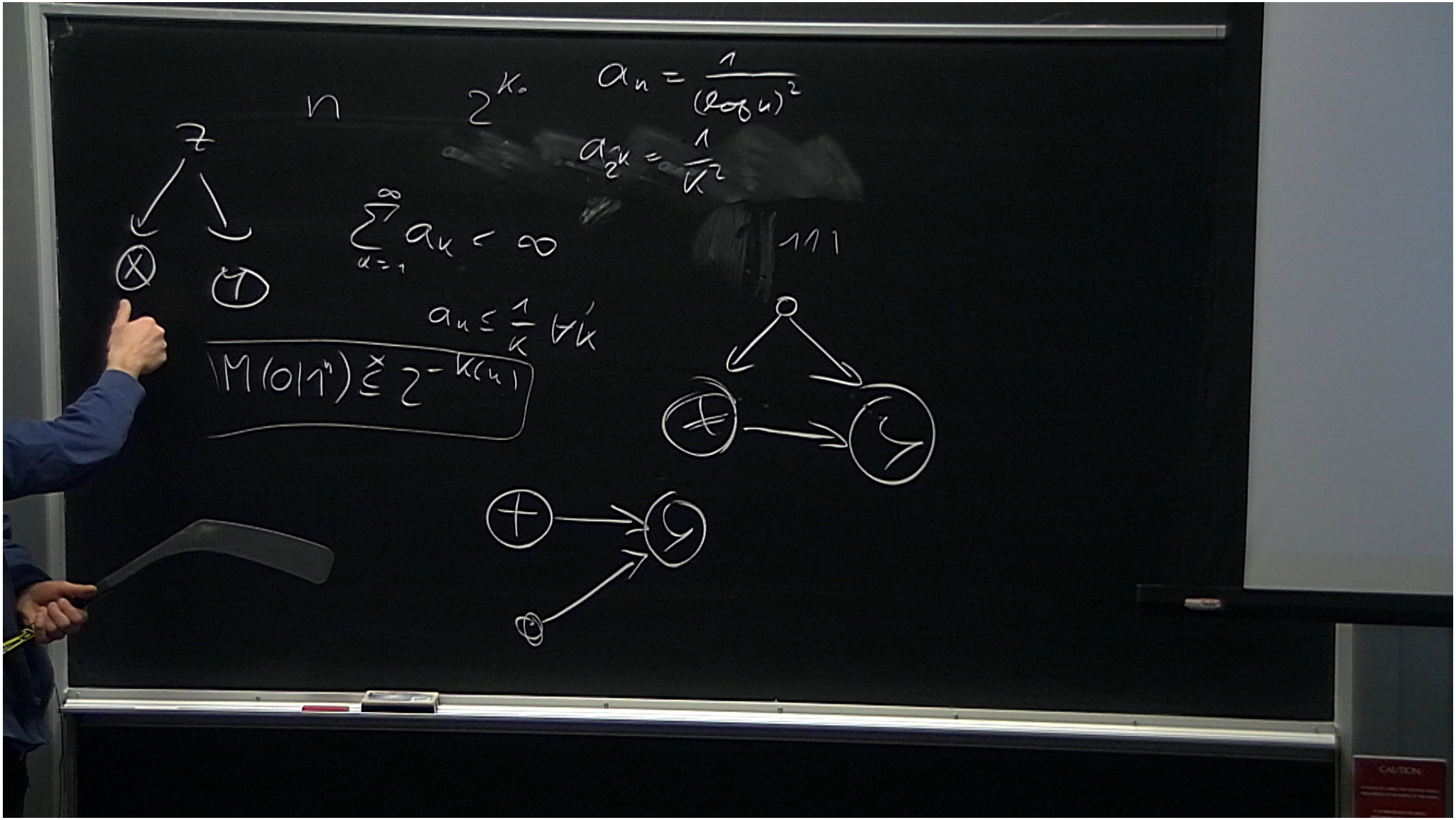
and a corresponding functional model

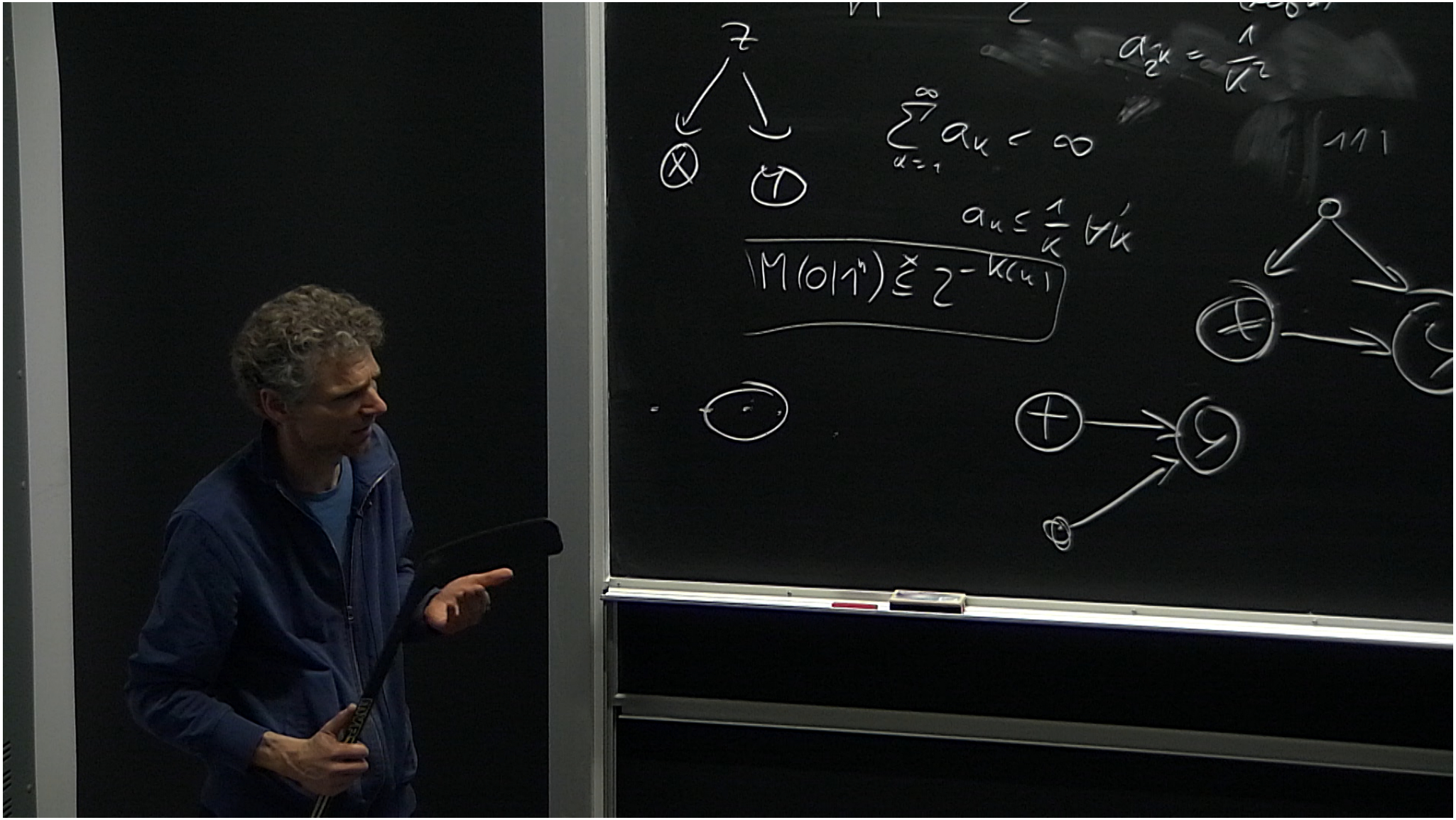
- **number** of different words of a text
  author takes words from parent texts and from an independent source (e.g. his/her mind)
- **Lempel-Ziv** compression length (approximately submodular)
  concatenate substrings from parents and noise
- logarithm of **least common multiple** of a set of natural numbers (log of period length of a signal)
  periodic signal obtained by linear combination of parent and noise signals

32




n
2^{k_n}
$a_n = \frac{1}{(\log n)^2}$
$a_{2^k} = \frac{1}{k^2}$
$\sum_{n=1}^{\infty} a_n < \infty$
$a_n \le \frac{1}{k} \forall k$


n
2^k
$a_n = \frac{1}{(\log n)^2}$
$a_{2^k} = \frac{1}{k^2}$
$\sum_{n=1}^{\infty} a_n < \infty$
$a_n \le \frac{1}{k} \forall k$
$M(0|1^n) \in 2^{-K(n)}$
111


$\sum_{k=1}^{\infty} a_k < \infty$
$a_k \leq \frac{1}{k} \quad \forall k$
$|M(0|1^n)| \leq 2^{-k(n)}$