Comparison of statistical experiments beyond the discrete case

Tobias Fritz

joint work with Tomáš Gonda, Paolo Perrone and Eigil Fjeldgren Rischel

November 2020

Introduction to synthetic probability via Markov categories

Tobias Fritz

November 2020

References

- Peter V. Golubtsov,
  Axiomatic description of categories of information converters. *Problemy Peredachi Informatsii* 35(3), 80–98 (1999).
  (And other similar papers by Golubtsov.)

- Kenta Cho and Bart Jacobs,
  Disintegration and Bayesian inversion via string diagrams.
  Math. Struct. Comp. Sci. 29, 938–971 (2019). arXiv:1709.00322.

- Tobias Fritz,
  A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. *Adv. Math.* 370, 107239 (2020).
  arXiv:1908.07021.

- Tobias Fritz and Eigil Fjeldgren Rischel,
  The zero-one laws of Kolmogorov and Hewitt–Savage in categorical probability.
  Compositionality 2, 3 (2020). arXiv:1912.02769.

- Evan Patterson,
  The algebra and machine representation of statistical models, PhD thesis.
  arXiv:2006.08945.

- Tobias Fritz, Tomáš Gonda, Paolo Perrone, Eigil Fjeldgren Rischel,
  Representable Markov Categories and Comparison of Statistical Experiments in Categorical Probability, arXiv:2010.07416.

Teaser

Theorem (Generalized Blackwell-Sherman-Stein theorem)

Let

- X, Y and Θ be standard Borel spaces,
- $(P_\theta)_{\theta\in\Theta}$ and $(Q_\theta)_{\theta\in\Theta}$ measurably indexed statistical models,
- m a probability measure on Θ (prior).

Then the following are equivalent:

(a) There is a Markov kernel $c : X \to Y$ such that

$$Q_\theta = c(P_\theta)$$

for m-almost all θ.

(b) The standard measures $\hat{f}_m$ and $\hat{g}_m$ on $P\Theta$ satisfy the second-order dominance relation

$$\hat{f}_m \sqsubseteq \hat{g}_m.$$

Teaser

- It generalizes the classical result in the discrete case.
- This is the first result in probability and statistics which is:
  - proven "synthetically",
  - apparently new even within traditional measure-theoretic probability!
- Unrelated to existing measure-theoretic generalizations (as far as we know).

Theorem (Prior-independent Blackwell-Sherman-Stein theorem)

Let

- ▷ X, Y and Θ be standard Borel spaces,
- ▷ $(P_\theta)_{\theta\in\Theta}$ and $(Q_\theta)_{\theta\in\Theta}$ measurably indexed statistical models.

Then the following are equivalent:

(a) There is a Markov kernel $c : X \times P\Theta \to Y$ such that

$$Q_\theta = c(P_\theta, m)$$

for m-almost every θ and every m.

(b) The standard measures $\hat{f}_m$ and $\hat{g}_m$ on $P\Theta$ satisfy the second-order dominance relation

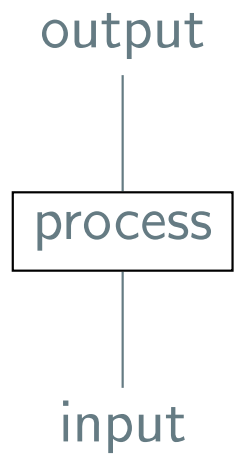
$$\hat{f}_m \sqsubseteq \hat{g}_m.$$

for every prior $m \in P\Theta$, as witnessed by a measurably m-dependent dilation $P\Theta \to P\Theta$.

Both theorems are instances of the same abstract result!

Ideas

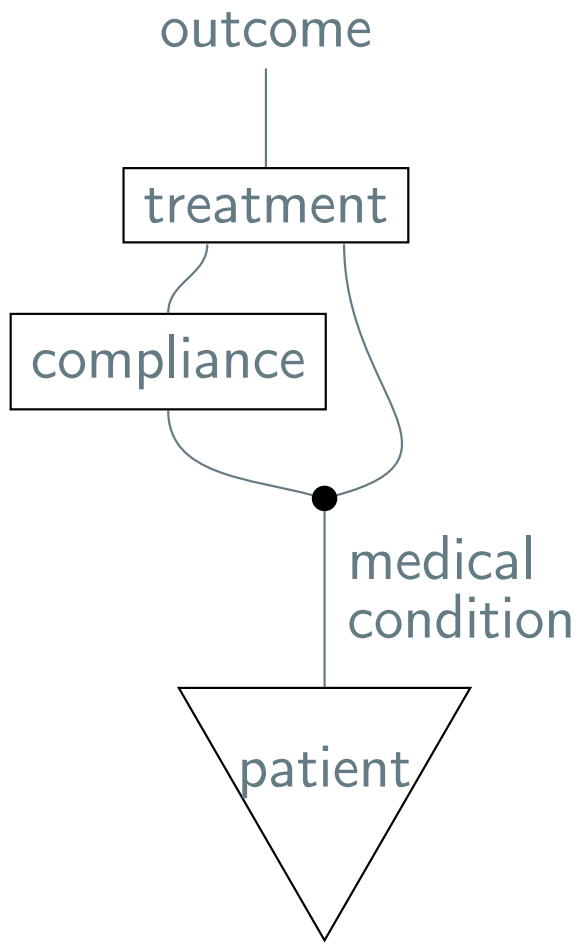
- The central objects of probability theory are not probability distribution, but **Markov kernels**



  which can be interpreted as
  - communication channels,
  - statistical models,
  - or **statistical experiments**.

- Do not say what a Markov kernel is — rather, say how it behaves!

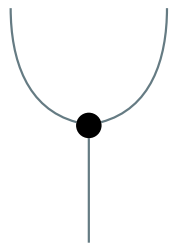
Suppose that we want to reason about **flow of information** in a medical trial. Then we seem to need diagrams like this:



$\rightarrow$ Medical condition has an influence on **both** trial compliance and on treatment outcome!

Ideas

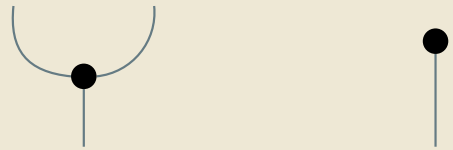
- Processes can have any number of inputs and outputs.
- **Distributions** are special processes with no inputs.
- To describe information flow, have additional pieces of structure:
  - copying information:



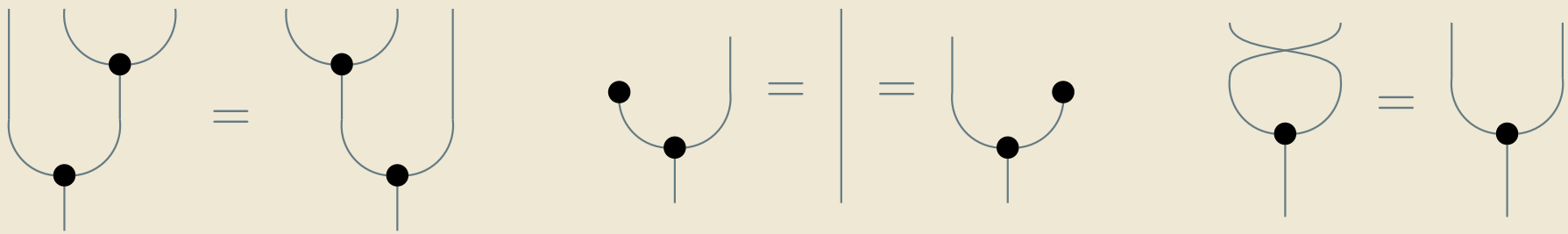
  - deleting information:

Definition

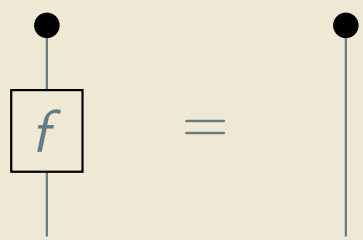
A **Markov category C** is a symmetric monoidal category supplied with **copying** and **deleting** operations on every object,



giving commutative comonoid structures



which interact well with the monoidal structure, and such that



(do we really want this?)

A basic example

One of the paradigmatic Markov categories is **FinStoch**, the category of finite sets and **stochastic matrices**:

- A morphism $f : X \to Y$ is

$$(f(y|x))_{x\in X, y\in Y} \in \mathbb{R}^{X\times Y}$$

  with

$$f(y|x) \geq 0, \qquad \sum_y f(y|x) = 1.$$

- Composition is the **Chapman-Kolmogorov formula**,

$$(gf)(z|x) := \sum_y g(z|y)\, f(y|x).$$

- A morphism $p : 1 \to X$ is a **probability distribution**.

- A general morphism $X \to Y$ has many names: **Markov kernel**, probabilistic mapping, information transformer, ...
- The monoidal structure implements **stochastic independence**,

$$(g \otimes f)(xy|ab) := g(x|a)\, f(y|b).$$

- The copy maps are

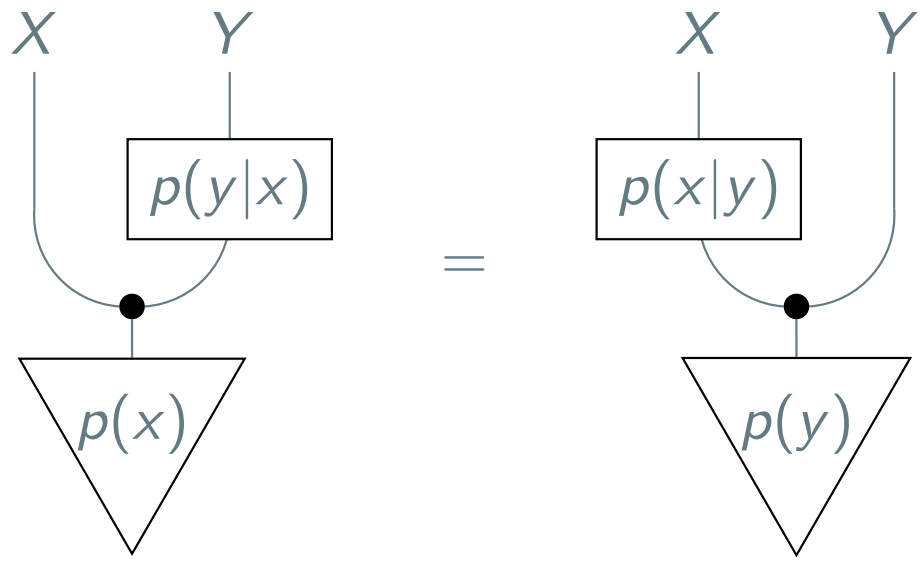
$$\mathrm{copy}_X : X \longrightarrow X \times X, \qquad \mathrm{copy}_X(x_1, x_2|x) = \begin{cases} 1 & \text{if } x_1 = x_2 = x, \\ 0 & \text{otherwise.} \end{cases}$$

- The deletion maps are the unique morphisms $X \to 1$.

The Markov category **BorelStoch** is defined similarly, with standard Borel spaces instead of finite sets.

A first theoretical development: Bayesian inversion

Bayes' rule takes the form:



$$p(y|x)\,p(x) \quad = \quad p(x|y)\,p(y)$$

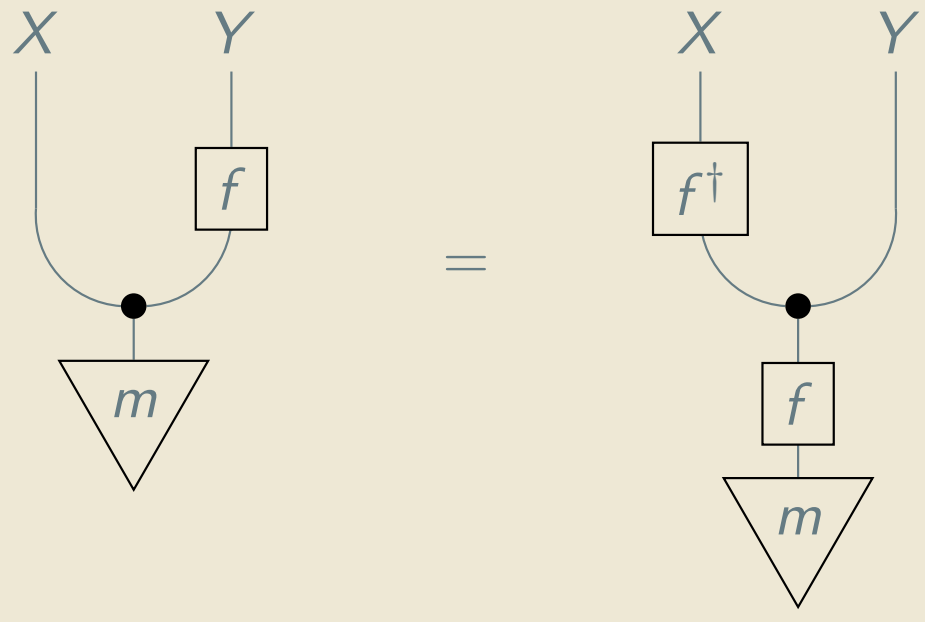
These exist (non-uniquely) in any Markov category **with conditionals**.

Bayesian inversion

Generally:

Definition

The **Bayesian inverse** of a process $f : X \to Y$ with respect to a distribution $\mu : I \to X$ is any $f^\dagger : Y \to X$ such that

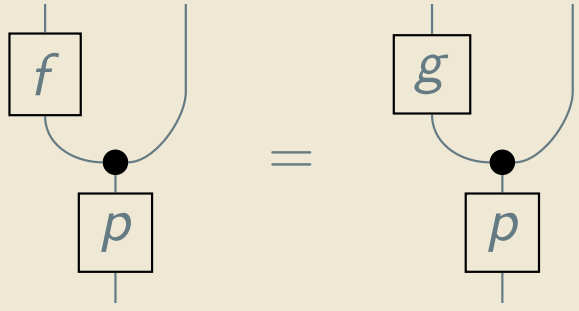


Almost sure equality

Definition

Let $p : A \to X$ and $f, g : X \to Y$.

f and g are **equal p-almost surely**, $f =_{p\text{-a.s.}} g$, if

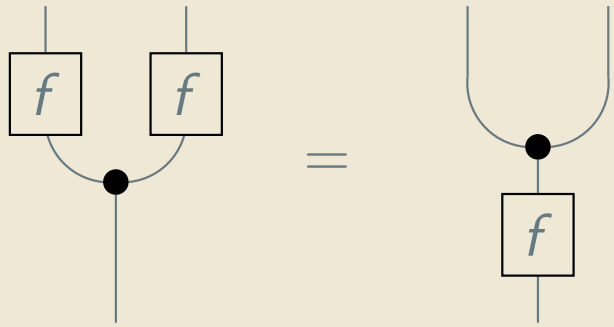


- ▷ **Intuition:** f and g behave the same on all inputs produced by p.
- ▷ Other concepts (besides equality) also relativize with respect to p-almost surely.

Determinism

Definition

In a Markov category, a morphism $f : X \to Y$ is **deterministic** if it commutes with copying,



- ▷ **Intuition:** Applying f to copies of input $=$ copying the output of f.
- ▷ The deterministic morphisms form a cartesian monoidal subcategory.

Representability

- In a **representable Markov category**, there is a bijection morphisms

$$f : A \to X$$

and deterministic morphisms

$$f^\sharp : A \to PX$$

where PX plays the role of the object of distributions on X.

- Under this correspondence, the deterministic identity $PX \to PX$ corresponds to the **sampling map**

$$\mathrm{samp} : PX \to X.$$

so that $\mathrm{samp}^\sharp = \mathrm{id}_{PX}$.

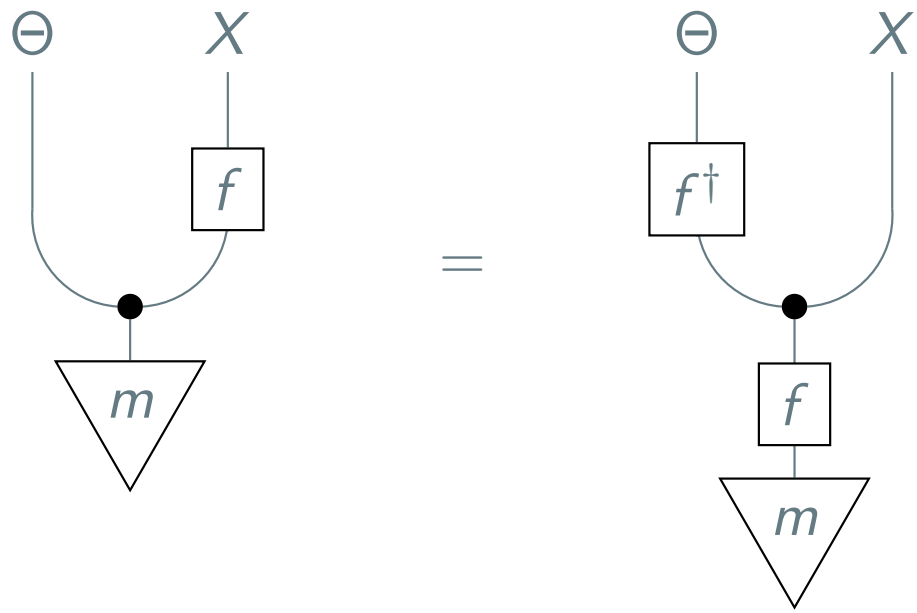
- Suppose that

$$f : \Theta \to X$$

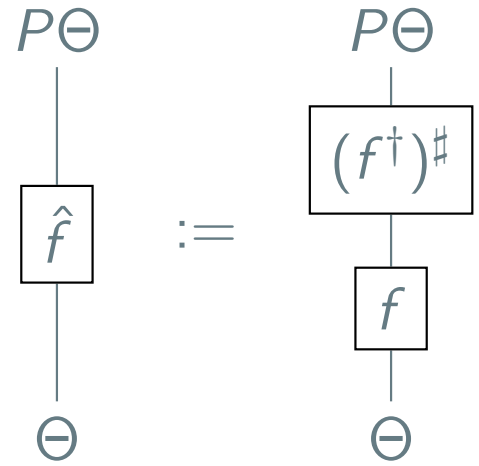
  is a statistical experiment, and $m : I \to \Theta$ a **prior** over hypotheses.

- The Bayesian inverse $f^\dagger : X \to \Theta$ computes the posterior from the experiment outcome.

- By definition

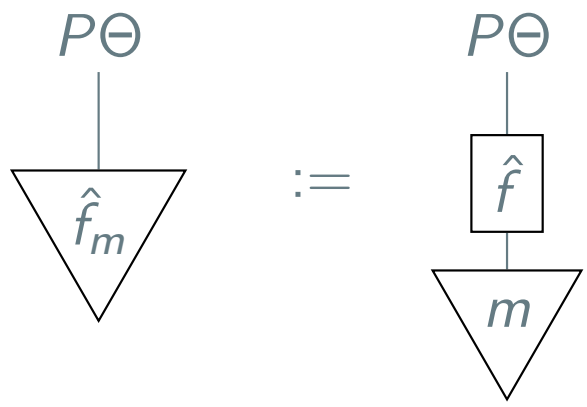


▷ The **standard experiment** is



It assigns to every hypothesis the resulting distribution over posteriors.

▷ The **standard measure** is



It is a distribution on $P\Theta$, namely the expected distribution over posteriors (with respect to the prior m).

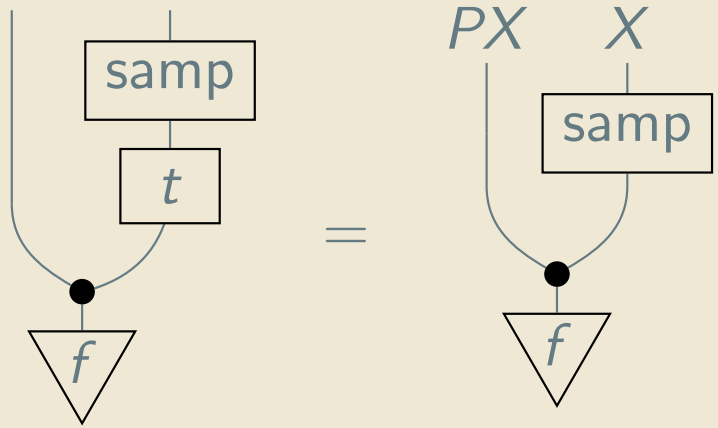
Second-order stochastic dominance

Definition

Given a distribution $f : I \to P\Theta$, an f-**dilation** is a morphism

$$t : P\Theta \to P\Theta$$

such that



Idea:

t preserves the expected distribution of a distribution over distributions, at least f-almost surely.

Second-order stochastic dominance

Definition

Given distributions $f, g : I \to P\Theta$, we say that g second-order dominates f,

$$f \sqsubseteq g$$

if there is an f-dilation $t : P\Theta \to P\Theta$ such that

$$f = tg.$$

This makes f "more spread out" than g.

Comparison of statstical experiments

Definition

Let $f : \Theta \to X$ and $g : \Theta \to Y$ be statistical experiments.
Then f is **more informative** than g if there is $c : X \to Y$ such that

$$g = cf\,.$$

Also consider the informativeness preorder up to almost sure equality with respect to prior m, where only

$$g =_{m\text{-a.s.}} cf$$

is needed.

The categorical Blackwell-Sherman-Stein theorem

Theorem

Let $\mathbf{C}$ be an a.s.-compatibly representable Markov category with conditionals.

Consider two morphisms $f : \Theta \to X$ and $g : \Theta \to Y$ in $\mathbf{C}$.

Then the following are equivalent:

(a) There exists a morphism $c : X \to Y$ such that

$$g =_{m\text{-a.s.}} cf\,.$$

(b) $\hat{f}_m \sqsubseteq \hat{g}_m$.

Theorems from beginning: follow upon instantiation on suitable $\mathbf{C}$.

$\top\hbar e\ \mathbb{E}\eta\partial$

A Markov category for information theory?

There are well-known analogies between probability and information theory:

- Conditional probability: $P(A|B) = \frac{P(A \cap B)}{P(B)}$.
- Conditional entropy: $H(A|B) = H(AB) - H(B)$.

Question

Is there a Markov category for information theory explaining these analogies?

Maybe like this:

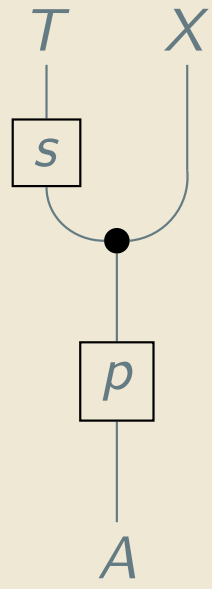
- Objects are finite sets,
- Morphisms $f : X \to Y$ are compatible families of stochastic maps

$$(f_n : X^{\times n} \to Y^{\times n})_{n \in \mathbb{N}}$$

modulo some suitable asymptotic equivalence as $n \to \infty$.

Definition

- A **statistical model** on X is a morphism $p : A \to X$.
- A **statistic** for p is a deterministic morphism $s : X \to T$.
- The statistic is **sufficient** if



displays $A \perp X \mid T$.

There is a version of the **Fisher–Neyman factorization theorem**.

Theorem

Suppose that $\mathbf{C}$ is strictly positive.

A statistic $s : X \to T$ is sufficient for $p : A \to X$ if and only if there is $\alpha : T \to X$ with $\alpha sp = p$.

There are versions of other classical theorems of statistics.

Basu's theorem

A complete sufficient statistic for p is independent of any ancillary statistic.

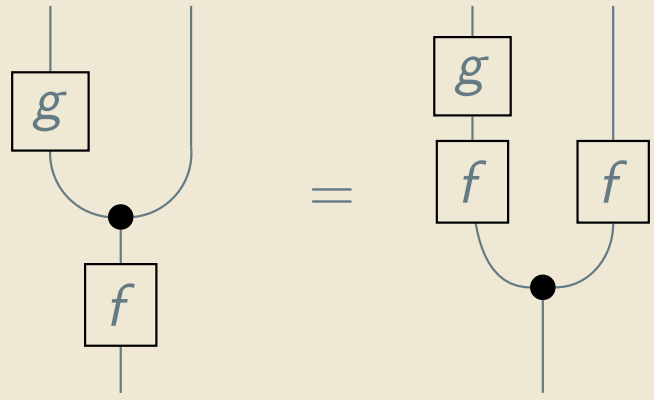
Bahadur's theorem

If a minimal sufficient statistic exists, then a complete sufficient statistic is minimal sufficient.

Explaining these would first require stating the relevant additional definitions, for which I don't have time.

Definition

C is **positive** if whenever gf is deterministic for composable f and g, then also

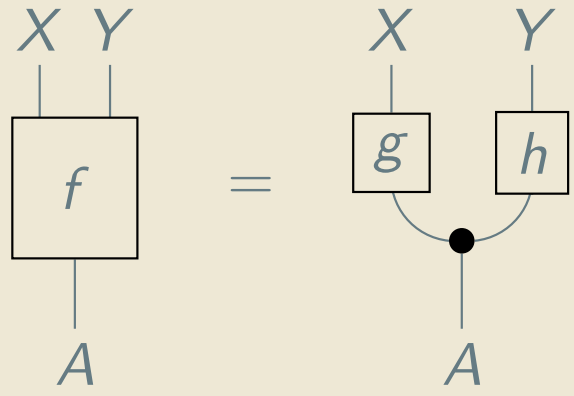


- **Intuition:** If a deterministic process has a random intermediate result, then that result can be computed independently from the process.

- Not every Markov category is positive.

Definition

$f : A \to X \otimes Y$ **displays the conditional independence** $X \perp Y \,||\, A$ if there are g and h such that



- **Intuition:** The outputs X and Y can be produced independently.
- Note the difference from the earlier definition of conditional independence!

Definition

Let $(X_i)_{i \in I}$ be a family of objects. The **infinite tensor product**

$$X_I := \bigotimes_{i \in I} X_i$$

is the cofiltered limit of the finite tensor products $X_F := \bigotimes_{i \in F} X_i$, if this limit exists and is preserved by every $- \otimes Y$.

Definition

An infinite tensor product X_I is a **Kolmogorov product** if the limit projections $\pi_F : X_I \to X_F$ are deterministic.

- ▷ This additional condition fixes the comonoid structure on X_I.

A piece of probability theory

One of the fundamental theorems of probability is the **law of large numbers**:

$$\mathbf{P}\left[\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mathbb{E}[X]\right] = 1. \qquad (*)$$

I don't yet know how to state and prove this in terms of Markov categories. But we have proven a closely related classical result synthetically.

Hewitt–Savage zero-one law

Let $(X_i)_{i\in\mathbb{N}}$ be independent and identically distributed random variables, and A any event depending only on the X_i and invariant under finite permutations.

Then $\mathbf{P}(A) \in \{0, 1\}$.

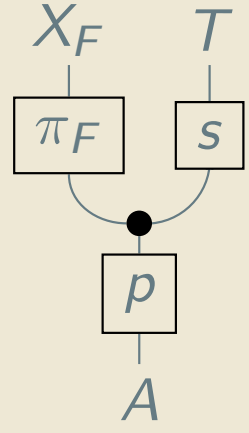
This implies that $(*)$ is 0 or 1, but we don't know which!

Theorem (Kolmogorov zero–one law)

Let X_I be a Kolmogorov product of a family $(X_i)_{i \in I}$.

If

- $p : A \to X_I$ makes the X_i independent and identically distributed, and
- $s : X_I \to T$ is such that

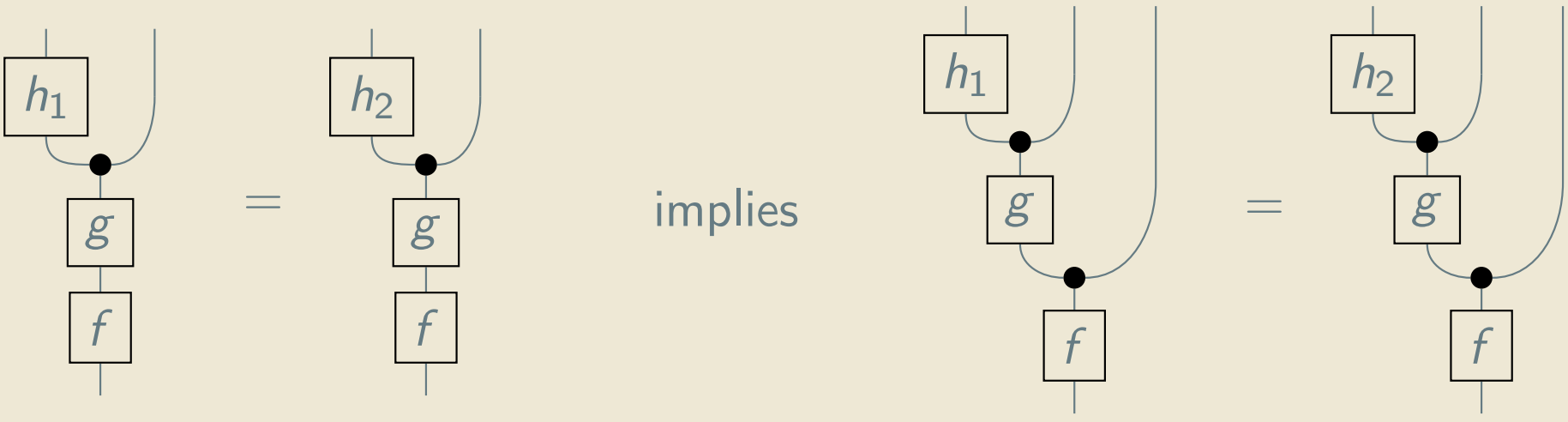


  displays $X_F \perp T \,||\, A$ for every finite $F \subseteq I$,

then sp is deterministic.

Definition

C is **causal** if



- **Intuition:** The choice between h_1 and h_2 in the "future" of g does not influence the "past" of g.

- Not every Markov category is causal.

Theorem (Hewitt–Savage zero–one law)

Suppose that $\mathbf{C}$ is causal, I infinite and $X_I := \bigotimes_{i \in I} X$ a Kolmogorov product of the same X with itself.

If

- $p : A \to X_I$ makes the X_i independent and identically distributed, and
- $s : X_I \to T$ is deterministic and invariant under finite permutations,

then sp is deterministic.

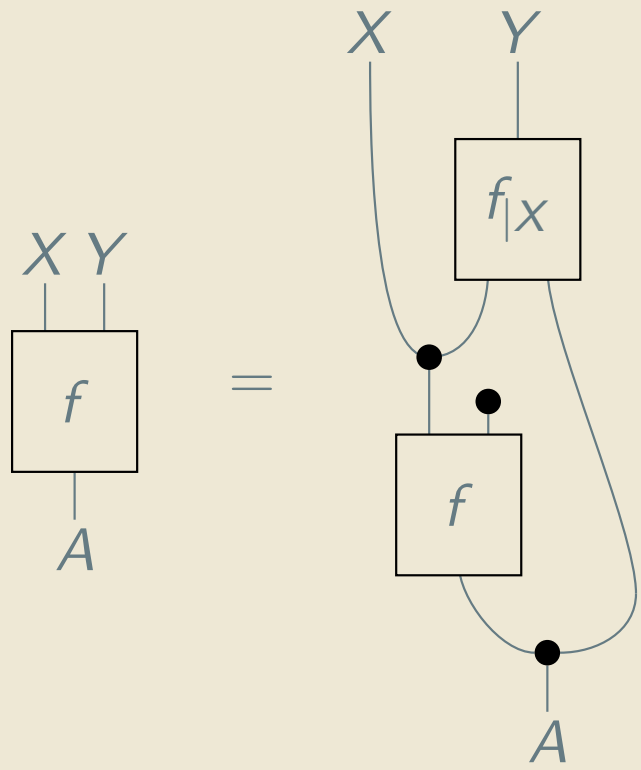
Proof is by string diagrams, but far from obvious!

Example

If $\prod_{i \in I} X$ is an infinite product of the same topological space, Y a Hausdorff space and $f : \prod_i X \to Y$ continuous and invariant under finite permutations, then f is constant.

Definition

C has conditionals if for $f : A \to X \otimes Y$ there is $f_{|X} : X \otimes A \to Y$ with



- If **C** has conditionals, then it is both strictly positive and causal.
- The positivity and causality axioms (partly?) eliminate the relevance of conditionals!