Information Geometry (2018) 1:237–285
https://doi.org/10.1007/s41884-018-0013-5

RESEARCH PAPER

Asymptotic dependency structure of multiple signals

Asymptotic equipartition property for diagrams of probability spaces

Rostislav Matveev[1] · **Jacobus W. Portegies**[2]

Received: 7 February 2018 / Revised: 16 August 2018 / Published online: 12 October 2018

Abstract
We formalize the notion of the dependency structure of a collection of *multiple* signals, relevant from the perspective of information theory, artificial intelligence, neuroscience, complex systems and other related fields. We model multiple signals by commutative diagrams of probability spaces with measure-preserving maps between some of them. We introduce the asymptotic entropy (pseudo-)distance between diagrams, expressing how much two diagrams differ from an information-processing perspective. If the distance vanishes, we say that two diagrams are asymptotically equivalent. In this context, we prove an asymptotic equipartition property: any sequence of tensor powers of a diagram is asymptotically equivalent to a sequence of homogeneous diagrams. This sequence of homogeneous diagrams expresses the relevant dependency structure.

Keywords Asymptotic equipartition property · Entropy distance · Diagrams of probability spaces · Multiple signals

1 Introduction

According to usual modeling assumptions in information theory, a discrete signal is cut into a collection of long words of length n, whose particular representation is irrelevant (each word is considered as an atomic object without inner structure), and small errors are allowed. If the signal is modeled as a sequence of independently, identically distributed random variables, there is only one relevant variable determining

✉ Rostislav Matveev
matveev@mis.mpg.de

Jacobus W. Portegies
j.w.portegies@tue.nl

[1] Max-Plank-Institut für Mathematik in der Naturwissenschaften, Inselstr. 22, 04103 Leipzig, Germany

[2] Eindhoven University of Technology, De Zaale, Eindhoven, The Netherlands

the signal, namely the entropy: the exponential growth rate of the number of typical words of length n. We elaborate on this point of view below in Sect. 1.1.

Similarly, if one probes a measure-preserving dynamical system at a discrete sequence of times with a finite-output measurement device and counts measurement trajectories of length n, while discarding rarely appearing, untypical ones, one arrives at the notion of entropy of a system-measurement pair. Entropy, in this case, is the exponential growth rate of the number of typical trajectories with respect to the length n. The supremum of such entropies over varying measurement devices is the Kolmogorov–Sinai entropy of a measure-preserving dynamical system. According to a theorem of Ornstein [17], the entropy is the only invariant of the isomorphism classes of certain types of dynamical systems (Bernoulli shifts).

In information theory, but also in artificial intelligence, neuroscience and the theory of complex systems, one usually studies *multiple* signals at once. Likewise, a dynamical system is often observed with multiple measurement devices simultaneously. In these cases, one assumes in addition that the relations between the signals are essential. In this article we characterize, under these modeling assumptions, the relevant invariants in multiple signals, that are obtained as i.i.d. samples from random variables. We explain this in more detail in Sects. 1.3 and 1.4.

We will now explain our point of view on entropy for a single signal, that is, for a single probability space.

1.1 Probability spaces and their entropy

First we consider a finite probability space $X = (S, p)$, where S is a finite set, and p is a probability measure on S. For simplicity, assume for now that the measure p has full support. Next, we consider the, so-called, Bernoulli sequence of probability spaces

$$X^n = (S^n, p^{\otimes n})$$

where S^n denotes the n-fold Cartesian product of S, and $p^{\otimes n}$ is the n-fold product measure.

The *entropy* of X is the exponential growth rate of the *observable cardinality* of tensor powers of X. The observable cardinality, loosely speaking, is the cardinality of the set X^n after the biggest possible subset of small measure has been removed. It turns out that the observable cardinality of X^n might be much smaller than $|S|^n$, the cardinality of the whole of X^n, in the following sense.

The *Asymptotic Equipartition Property* states that for every $\varepsilon > 0$ and $n \gg 0$ one can find a, so-called, *typical subset* $A_\varepsilon^{(n)} \subset S^n$, defined as a subset that takes up almost all of the mass of X^n and the probability distribution on $A_\varepsilon^{(n)}$ is almost uniform on the normalized logarithmic scale, as stated in the following theorem, see [8].

Theorem 1.1 (Asymptotic equipartition property) *Suppose $X = (S, p)$ is a finite probability space. Then, for every $\varepsilon > 0$ and every $n \gg 0$ there exists a subset $A_\varepsilon^{(n)} \subset X^n$ such that*

1. $p^{\otimes n}(A_\varepsilon^{(n)}) > 1 - \varepsilon$
2. *For any* $a, a' \in A_\varepsilon^{(n)}$ *holds*

$$\left| \frac{\ln p^{\otimes n}(a)}{n} - \frac{\ln p^{\otimes n}(a')}{n} \right| < \varepsilon.$$

Moreover, if $A_\varepsilon^{(n)}$ *and* $B_\varepsilon^{(n)}$ *are two subsets of* X^n *satisfying the two conditions above, then their cardinalities satisfy*

$$\left| \frac{\ln |A_\varepsilon^{(n)}|}{n} - \frac{\ln |B_\varepsilon^{(n)}|}{n} \right| < 2\varepsilon. \tag{1}$$

The cardinality $|A_\varepsilon^{(n)}|$ may be much smaller than $|S|^n$, but it will still grow exponentially with n. Even though there are generally many choices for such a set $A_\varepsilon^{(n)}$, in view of the property (1) in Theorem 1.1, the exponential growth rate with respect to n is well-defined up to 2ε.

The limit of the growth rate as $\varepsilon \to 0+$ is called the entropy of X

$$\mathsf{Ent}(X) := \lim_{\varepsilon \downarrow 0} \lim_{n \to \infty} \frac{1}{n} \ln |A_\varepsilon^{(n)}|. \tag{2}$$

This point of view on entropy goes back to the original idea of Boltzmann [3,4], according to which entropy is the logarithm of the number of equiprobable states, that a system, comprised of many identical weakly interacting subsystems, may take on. It was further developed and applied to Information Theory by Shannon [19], and in the context of dynamical systems by Kolmogorov and Sinai [12,13,21].

Entropy is especially easy to evaluate if the space is uniform, since for any finite probability space with the uniform distribution the observable cardinality is equal to the cardinality of the whole space and therefore

$$\mathsf{Ent}(X) = \ln |X|. \tag{3}$$

For non-uniform spaces, the entropy can be evaluated by the well-known formula

$$\mathsf{Ent}(X) = -\sum_{x \in S_X} p_X(x) \ln p_X(x).$$

1.2 Asymptotic equivalence

If X_1 and X_2 are probability spaces with the same entropy, there is a bijection between their typical sets of sequences of length n, for the plain reason that they can be chosen to have the same cardinality. It means that up to a change of code (of representation) and an error that becomes small as n gets large, the spaces X_1^n and X_2^n are equivalent. In the same sense, both X_1^n and X_2^n are equivalent to a uniform measure space with cardinality $\mathbf{e}^{n\mathsf{Ent}(X_i)}$.

In [10], Gromov formalized this concept of asymptotic equivalence. With his definition, two Bernoulli sequences of measure spaces X_1^n and X_2^n are asymptotically equivalent if there exists an "almost-measure-preserving" "almost-bijection"

$$X_1^n \overset{f}{\dashleftarrow\!\!\dashrightarrow} X_2^n$$

Even though we were greatly influenced by ideas in [10], we found that Gromov's definition does not extend easily to situations in which *multiple signals* are processed at the same time, or when a dynamical system is probed with several measurement devices at once.

1.3 Diagrams of probability spaces

We model multiple signals by *diagrams* of probability spaces. By a diagram of probability spaces we mean a commutative diagram of probability spaces and measure-preserving maps between some of them. We will give a precise definition in Sect. 2.4, but will now consider particular examples of diagrams called *two-fans*.

A two-fan

$$X \xleftarrow{\pi_X} U \xrightarrow{\pi_Y} Y$$

is a triple of probability spaces $X = (\underline{X}, p_X)$, $Y = (\underline{Y}, p_Y)$ and $U = (\underline{U}, p_U)$, and two measure-preserving maps π_X and π_Y. We will restrict ourselves for now to the case in which the underlying set of U is the Cartesian product of the underlying sets of X and Y, $\underline{U} = \underline{X} \times \underline{Y}$, and π_X and π_Y are the ordinary projections. Such a situation arises, for example, when a complex dynamical system, such as a living cell or a brain, is observed via two measuring devices.

Generalizing from the case of single signals, we might want to say that two-fans

$$\begin{aligned} X_1 &\longleftarrow U_1 \longrightarrow Y_1 \\ X_2 &\longleftarrow U_2 \longrightarrow Y_2 \end{aligned}$$

are asymptotically equivalent if for large n there exist almost measure-preserving, almost-bijections

$$\begin{array}{ccc} X_1^n & U_1^n & Y_1^n \\ \uparrow\downarrow f & \uparrow\downarrow h & \uparrow\downarrow g \\ X_2^n & U_2^n & Y_2^n \end{array}$$

Without additional assumptions, asymptotic equivalence classes for two-fans would be completely determined by the entropies of the constituent spaces.

However, such an asymptotic equivalence relation would be too coarse. Consider the three examples of two-fans are shown in Fig. 1, which is to be interpreted in the following way. Each of the spaces X_i and Y_i, $i = 1, 2, 3$, have cardinality six and a uniform distribution, where the weight of each atom is $\frac{1}{6}$. The spaces U_i have cardinality 12 and the distribution is also uniform with all weights being $\frac{1}{12}$. The

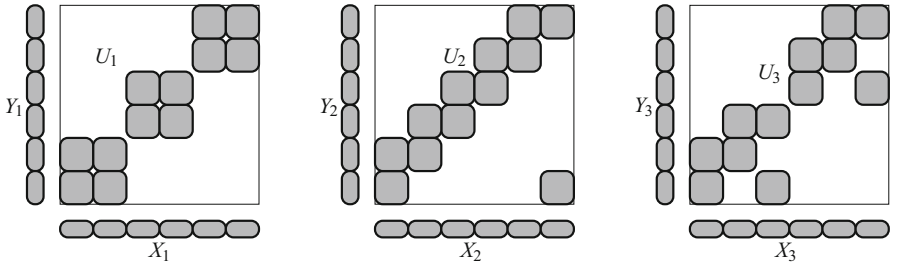


Fig. 1 Examples of pairs of probability spaces together with joint distributions

support of the measure on U_i's is colored grey on the pictures. The maps from U_i to X_i and Y_i are coordinate projections.

In view of Eq. (3) we have for each $i = 1, 2, 3$,

$$\begin{aligned} \mathsf{Ent}(X_i) &= \ln 6 \\ \mathsf{Ent}(Y_i) &= \ln 6 \\ \mathsf{Ent}(U_i) &= \ln 12. \end{aligned}$$

However, common information-processing techniques can still differentiate between the two-fans, by calculating solutions to information-optimization problems. This observation is sometimes expressed by saying that mutual information does not capture the full dependency structure that is relevant from an information-processing perspective. Information-optimization problems play an important role in information theory [25], causal inference [18], artificial intelligence [23], information decomposition [5], robotics [1], and neuroscience [9].

The additional assumption that the relations between the signals is relevant and should be preserved by the asymptotic equivalence results in the requirement that the following diagram commutes

$$\begin{array}{ccccc} X_1^n & \longleftarrow & U_1^n & \longrightarrow & Y_1^n \\ \uparrow f & & \uparrow h & & \uparrow g \\ X_2^n & \longleftarrow & U_2^n & \longrightarrow & Y_2^n \end{array}$$

However, with this generalization of an asymptotic equivalence to diagrams, we were not able to prove a corresponding Asymptotic Equipartition Property or even to prove the transitivity of the relation.

1.4 The entropy distances and asymptotic equivalence for diagrams

Instead of finding an almost measure-preserving bijection between large parts of the two spaces, we consider a stochastic coupling (transportation plan, joint distribution) between a pair of spaces and measure its deviation from being an isomorphism of probability spaces (a measure-preserving bijection). Such a measure of deviation from being an isomorphism then leads to the notion of intrinsic entropy distance, and its

stable version—the asymptotic entropy distance, as explained in Sect. 3. We say two sequences of diagrams are asymptotically equivalent if the asymptotic entropy distance between them vanishes.

The intrinsic entropy distance is an intrinsic version of a distance between random variables going by many different names, such as entropy distance, shared information distance and variation of information. It was reinvented many times by different people, among them Shannon [20], Kolmogorov, Sinai and Rokhlin. It appears in the proof of the theorem about generating partitions for ergodic systems by Kolmogorov and Sinai, see for example [22].

The intrinsic version of the entropy distance between probability spaces was introduced by Kovacevic et al. [14] and by Vidyasagar [24]. They showed that the involved minimization problem is NP-hard. Methods to find approximate solutions are discussed in [6,11].

1.5 Asymptotic equipartition property

With the notion of asymptotic equivalence induced by the asymptotic entropy distance, we *can* prove an asymptotic equipartition property for diagrams. Whereas the asymptotic equipartition property for single probability spaces states that high tensor powers of probability spaces can be approximated by *uniform* measure spaces, the Asymptotic Equipartition Property Theorem for diagrams, Theorem 6.1, states that sequences of successive tensor powers of a diagram can be approximated in the asymptotic entropy distance by a sequence of *homogeneous* diagrams.

Homogeneous diagrams have the property that the symmetry group acts transitively on the support of the measures of the constituent spaces. Two-fans shown on Fig. 1 are particular examples of homogeneous diagrams.

Homogeneous probability spaces are just uniform probability spaces, while homogeneous *diagrams* are, unlike homogeneous probability spaces, rather complex objects. Nonetheless, they seem to be simpler than arbitrary diagrams of probability spaces for the types of problems that we would like to address.

In a subsequent article we show that the optimal values in Information-Optimization problems only depend on the asymptotic class of a diagram and that they are continuous with respect to the asymptotic entropy distance; in many cases, the optimizers are continuous as well. The Asymptotic Equipartition Property implies that for the purposes of calculating optimal values and approximate optimizers, one only needs to consider homogeneous diagrams and this can greatly simplify computations.

Summarizing, the Asymptotic Equipartition Property and the continuity of Information-Optimization problems are important justifications for the choice of asymptotic equivalence relation and the introduction of the intrinsic and asymptotic Kolmogorov–Sinai distances.

1.6 Definitions and results in random variable context

In this article, we use the language of probability spaces and their commutative diagrams rather than the language of random variables, because we often encounter

situations in which their joint distributions are not defined, are variable, or even do not exist.

Some relations between the probability spaces can be easily represented by commutative diagrams of probability spaces, such as by a diamond diagram, Sect. 2.5.5, while the description with random variables is complex and not easily interpretable. The diagrams also provide a geometric overview of various entropy identities and inequalities.

Since the language of random variables will be more familiar to many readers, we now present our main result in these terms.

For random variables $\mathsf{X}, \mathsf{Y}, \mathsf{Z}$ etc., we denote by $\underline{X}, \underline{Y}, \underline{Z}$ the target sets, and by X, Y, Z the probability spaces with the induced distributions.

In general, there is a relation between k-tuples of random variables and diagrams of a certain type, involving a space for every subset $I \subset \{1, \dots, k\}$.

For example, a pair of random variables X, Y (defined on the same probability space) gives rise to a two-fan,

$$X \longleftarrow X \times Y \longrightarrow Y$$

where X, Y and $X \times Y$ are the target spaces of the random variables X, Y and (X, Y) endowed with their respective laws (i.e. the pushforward of the probability measure).

However, not every type of diagram corresponds to a tuple of random variables.

The entropy distance between two k-tuples $\mathsf{X} = (\mathsf{X}_1, \dots, \mathsf{X}_k)$ and $\mathsf{Y} = (\mathsf{Y}_1, \dots, \mathsf{Y}_k)$ is defined by

$$\mathrm{kd}(\mathsf{X}, \mathsf{Y}) := \sum_{I \subset \{1,\dots,k\}} \big(2\mathsf{Ent}(\mathsf{X}_I \mathsf{Y}_I) - \mathsf{Ent}(\mathsf{X}_I) - \mathsf{Ent}(\mathsf{Y}_I)\big).$$

A random k-tuple $(\mathsf{H}_1, \dots, \mathsf{H}_k)$ is called homogeneous if for every two elements

$$(h_1, \dots, h_k) \in \underline{H}_1 \times \cdots \times \underline{H}_k$$
$$(h'_1, \dots, h'_k) \in \underline{H}_1 \times \cdots \times \underline{H}_k$$

there exists a k-tuple of invertible maps $f_1 : \underline{H}_1 \to \underline{H}_1, \dots, f_k : \underline{H}_k \to \underline{H}_k$ such that $(f_1 \circ \mathsf{H}_1, \dots, f_k \circ \mathsf{H}_k)$ is equal to $(\mathsf{H}_1, \dots, \mathsf{H}_k)$ in distribution and $f_i(h_i) = h'_i$. This condition is strictly stronger than the requirement that all the distributions are uniform.

Given n random k-tuples

$$\mathsf{X}(i) = \big(\mathsf{X}_1(i), \dots, \mathsf{X}_k(i)\big), \quad i = 1, \dots, n$$

we can naturally construct the k-tuple

$$\big(\mathsf{X}(1), \dots, \mathsf{X}(n)\big)$$

defined by

$$\big(\mathsf{X}(1),\dots,\mathsf{X}(n)\big)_j = \big(\mathsf{X}_j(1),\dots,\mathsf{X}_j(n)\big),\quad j=1,\dots,k.$$

It is difficult to formulate our main result, Theorem 6.1, in full generality using the language of random variables. However, the following theorem is an immediate corollary.

Theorem 1 *Let $(X(i) : i \in \mathbb{N})$ be a sequence of i.i.d. random tuples defined on a standard probability space.*

Define random k-tuples $Y(n)$ by

$$Y(n) := (X(1),\dots,X(n)).$$

Then, there exists a sequence of homogeneous random k-tuples $H(n) = (H_1(n),\dots,H_k(n))$, where $H_i(n)$ takes values in $\underline{X}_i^n$, such that

$$\frac{1}{n} kd(Y(n), H(n)) = O\left(n^{-1/2}\ln^{3/2}(n)\right).$$

2 Category of probability spaces and diagrams

In this section we present the basic setup used throughout the article. We will start by explaining how probability spaces and (equivalence classes) of measure-preserving maps between them form a *category*. This point of view on probability theory was already advocated in [2,10].

Category theory yields simple definitions of diagrams of probability spaces and morphisms between them and allows for precise and relatively short proofs. The setup is also convenient when couplings (joint distributions) between probability spaces are absent or variable.

2.1 Categories

Below we briefly review elementary category theory. We refer the reader to the first chapter of [15] for a more extensive introduction.

A *category* $\mathbf{C}$ is an abstract mathematical structure that captures the idea of a collection of spaces and structure-preserving maps between them, such as groups and homomorphisms, vector spaces and linear maps, and topological spaces and continuous maps. Categories consist of a collection of objects (which need not to be sets), a collection of morphisms (which need not to be maps), and a rule for composing morphisms.

More formally a category consists of

- A class of objects $\mathrm{Obj}_{\mathbf{C}}$;
- A class of morphisms $\mathrm{Hom}_{\mathbf{C}}(A, B)$ for every pair of objects $A, B \in \mathrm{Obj}_{\mathbf{C}}$. For a morphism $f \in \mathrm{Hom}_{\mathbf{C}}(A, B)$ one usually writes $f : A \to B$. Object A will be

called the domain and B the target of f, and we say that f is a morphism from A to B;
– For each triple of objects A, B and C, a binary, associative operation, called composition,

$$\begin{aligned} \circ : \mathrm{Hom}_{\mathbf{C}}(B, C) \times \mathrm{Hom}_{\mathbf{C}}(A, B) &\to \mathrm{Hom}_{\mathbf{C}}(A, C) \\ (g, f) &\mapsto g \circ f \end{aligned}$$

– For every object $A \in \mathrm{Obj}_{\mathbf{C}}$ an identity morphism $\mathbf{1}_A : A \to A$, with the property that for every $f : A \to B$ and every $g : B \to A$,

$$f \circ \mathbf{1}_A = f, \quad \mathbf{1}_A \circ g = g.$$

A morphism $f : A \to B$ is an *isomorphism* if there exists a morphism $g : B \to A$ such that $f \circ g = \mathbf{1}_B$ and $g \circ f = \mathbf{1}_A$.

Category theory becomes a very powerful tool when *functors* and their *natural transformations* are considered. Functors can be seen as homomorphisms between categories. In turn, *natural transformations* are homomorphisms between functors.

A (covariant) *functor* $\mathscr{X} : \mathbf{C} \to \mathbf{D}$ between two categories $\mathbf{C}$ and $\mathbf{D}$, maps objects and morphisms in $\mathbf{C}$ to objects and morphisms in $\mathbf{D}$, respectively. It satisfies the following additional properties: For every morphism $f : A \to B$ in $\mathbf{C}$ the image $\mathscr{X}(f)$ is a morphism from $\mathscr{X}(A)$ to $\mathscr{X}(B)$ in $\mathbf{D}$ and composition is preserved,

$$\mathscr{X}(g \circ f) = \mathscr{X}(g) \circ \mathscr{X}(f)$$

for any pair of morphisms $f : A \to B$ and $g : B \to C$.

A *natural transformation* between functors $\mathscr{X}, \mathscr{Y} : \mathbf{C} \to \mathbf{D}$ is a family η of morphisms in category $\mathbf{D}$, indexed by objects in $\mathbf{C}$: For every $A \in \mathrm{Obj}_{\mathbf{C}}$, there is a morphism $\eta_A : \mathscr{X}(A) \to \mathscr{Y}(A)$, such that for every morphism $f : A \to B$ the diagram

$$\begin{array}{ccc} \mathscr{X}(A) & \xrightarrow{\mathscr{X}(f)} & \mathscr{X}(B) \\ \downarrow{\scriptstyle \eta_A} & & \downarrow{\scriptstyle \eta_B} \\ \mathscr{Y}(A) & \xrightarrow{\mathscr{Y}(f)} & \mathscr{Y}(B) \end{array}$$

commutes, that is

$$\eta_B \circ \mathscr{X}(f) = \mathscr{Y}(f) \circ \eta_A.$$

2.2 Probability spaces and reductions

We will now describe the category **Prob**. The objects in **Prob** are *finite probability spaces*. A finite probability space X is a pair (S, p), where S is a (not necessarily finite) set and $p : 2^S \to [0, 1]$ is a probability measure, such that there is a *finite* subset of S with full measure. We denote by $\underline{X} = \mathrm{supp}\, p$ the support of the measure and by $|X| := |\mathrm{supp}\, p_X|$ its cardinality. Slightly abusing the language, we call this quantity

the *cardinality* of X. We will no longer explicitly mention that the probability spaces we consider are finite. We will also write p_X where we truly mean its density with respect to the counting measure.

We say that a map $f\colon X \to Y$ between two probability spaces X and Y is *measure-preserving* if the push-forward $f_* p_X$ equals p_Y. This means that for every $A \subset Y$,

$$(f_* p_X)(A) := p_X(f^{-1}A) = p_Y(A).$$

We say that two measure-preserving maps $f\colon X \to Y$ are equivalent if they agree on a set of full measure. We call an equivalence class of measure-preserving maps from X to Y a *reduction*.

The morphisms in the category **Prob** are exactly the *reductions* between finite probability spaces. At this stage one might want to check that **Prob** is indeed a category, and this is guaranteed as the composition of two reductions is again a reduction.

2.3 Isomorphisms, automorphisms and homogeneity

Now that we have organized probability spaces and reductions into a category, we get concepts such as isomorphism for free: Two probability spaces X and Y are isomorphic in the category **Prob** if and only if there exists a measure-preserving bijection between the supports of the measures on X and Y. If X and Y are isomorphic, they have the same cardinality. The automorphism group $\mathrm{Aut}(X)$ is the group of all self-isomorphisms of X.

A probability space X is called *homogeneous* if the automorphism group $\mathrm{Aut}(X)$ acts transitively on the support $\underline{X}$ of the measure. For the category **Prob**, this turns out to be a complicated way of saying that the measure on X is uniform on its support, but when we consider *diagrams* later, there will be no such simple implication. Homogeneity is an isomorphism invariant and we will denote the subcategory of homogeneous spaces by $\mathbf{Prob_h}$.

There is a product in **Prob** (which is not a product in the sense of category theory!) given by the Cartesian product of probability spaces, that we will denote by $X \otimes Y := (\underline{X} \times \underline{Y}, p_X \otimes p_Y)$, where $p_X \otimes p_Y$ is the (independent) product measure. There are canonical reductions $X \otimes Y \to X$ and $X \otimes Y \to Y$ given by projections to factors. For a pair of reductions $f_i\colon X_i \to Y_i$, $i = 1, 2$ their tensor product is the reduction $f_1 \otimes f_2\colon X_1 \otimes X_2 \to Y_1 \otimes Y_2$, which is equal to the class of the Cartesian product of maps representing f_i's. The product leaves the subcategory of homogeneous spaces invariant. If one of the factors in the product is replaced by an isomorphic space, then the product stays in the same isomorphism class.

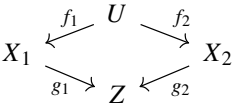
We close this section with a technical remark. The category **Prob** is not a *small category*. However it has a small full subcategory, that contains an object for every isomorphism class in **Prob** and for every pair of objects in it, it contains all the available morphisms between them and is closed under the product. From now on we imagine that such a subcategory was chosen and fixed and replaces **Prob** in all considerations below.

2.4 Diagrams of probability spaces

Essentially, a diagram $\mathscr{X} = \{X_i; f_{ij}\}$ is a commutative diagram in **Prob** consisting of a finite number of probability spaces and reductions between some of them. We have seen an example of the two-fan diagram in the introduction

$$X \xleftarrow{f} U \xrightarrow{g} Y$$

Another example is the diamond diagram



We require the diagram to be commutative, that is

$$g_1 \circ f_1 = g_2 \circ f_2.$$

A morphism $\rho\colon \mathscr{X} \to \mathscr{Y}$ between two diagrams $\mathscr{X} = \{X_i; f_{ij}\}$ and $\mathscr{Y} = \{Y_i; g_{ij}\}$ of the same shape is a collection of reductions between corresponding individual objects $\rho_i\colon X_i \to Y_i$, that commute with the reductions within each diagram, $\rho_j \circ f_{ij} = g_{ij} \circ \rho_i$.

We need to keep track of the combinatorial structure of the collection of reductions within a diagram. There are several possibilities for doing so:

- the reductions form a directed, acyclic graph which is transitively closed;
- the spaces in the diagram form a poset;
- the underlying combinatorial structure could be recorded as a finite category.

The last option seems to be most convenient since it has many operations, that are necessary for our analysis, already built-in. Besides, we need at times to iterate the construction of commutative diagrams, to create diagrams of diagrams, which is readily available in the category-theory framework but is cumbersome in the other contexts.

A (finite) poset category $\mathbf{G}$ is a finite category such that for every two objects O_1 and O_2 there is at most one morphism between them in either direction:

$$\left|\mathrm{Hom}_{\mathbf{G}}(O_1, O_2) \cup \mathrm{Hom}_{\mathbf{G}}(O_2, O_1)\right| \leq 1.$$

For instance, the poset category $\boldsymbol{\Lambda}_2$,

$$O_1 \xleftarrow{\pi_1} O_{12} \xrightarrow{\pi_2} O_2$$

is a category with three objects $\{O_1, O_{12}, O_2\}$ and two non-identity morphisms $\pi_1\colon O_{12} \to O_1$ and $\pi_2\colon O_{12} \to O_2$.

A two-fan is then a diagram indexed by $\boldsymbol{\Lambda}_2$: we assign to each object in $\boldsymbol{\Lambda}_2$ a probability space and to each morphism in $\boldsymbol{\Lambda}_2$ a reduction.

In general, then, a *diagram of probability spaces* indexed by a poset category $\mathbf{G}$ is a functor $\mathscr{X}\colon \mathbf{G} \to \mathbf{Prob}$. The requirement that $\mathscr{X}$ is a functor and not just a map between objects and morphisms (combined with the assumption that there is only one morphism between objects), is exactly the requirement that the diagrams should be commutative.

The collection of all diagrams of probability spaces indexed by a fixed poset category $\mathbf{G}$ forms the so-called category of functors

$$\mathbf{Prob}\langle\mathbf{G}\rangle := [\mathbf{G}, \mathbf{Prob}].$$

The objects of $\mathbf{Prob}\langle\mathbf{G}\rangle$ are diagrams, that is functors from $\mathbf{G}$ to $\mathbf{Prob}$, while morphisms in $\mathbf{Prob}\langle\mathbf{G}\rangle$ are natural transformations between them. We will refer to the morphisms in $\mathbf{Prob}\langle\mathbf{G}\rangle$ as reductions as well.

Let us go through the simple example of two-fans: we look at a reduction $\eta\colon \mathscr{X} \to \mathscr{Y}$ between two-fan diagrams $\mathscr{X}, \mathscr{Y}\colon \boldsymbol{\Lambda}_2 \to \mathbf{Prob}$. The reduction η, being a natural transformation between $\mathscr{X}$ and $\mathscr{Y}$, is illustrated by the commutative diagram

$$\begin{array}{ccccc}
\mathscr{X}(O_1) & \xleftarrow{\mathscr{X}(\pi_1)} & \mathscr{X}(O_{12}) & \xrightarrow{\mathscr{X}(\pi_2)} & \mathscr{X}(O_2) \\
\downarrow{\scriptstyle \eta(O_1)} & & \downarrow{\scriptstyle \eta(O_{12})} & & \downarrow{\scriptstyle \eta(O_2)} \\
\mathscr{Y}(O_1) & \xleftarrow{\mathscr{Y}(\pi_1)} & \mathscr{Y}(O_{12}) & \xrightarrow{\mathscr{Y}(\pi_2)} & \mathscr{Y}(O_2)
\end{array}$$

Thus, a reduction of a two-fan is a family of reductions of probability spaces indexed by the objects in the poset category $\boldsymbol{\Lambda}_2$ such that the diagram commutes.

For a diagram $\mathscr{X} \in \mathbf{Prob}\langle\mathbf{G}\rangle$, the poset category $\mathbf{G}$ will be called the *combinatorial type* of $\mathscr{X}$. For a poset category $\mathbf{G}$ or a diagram $\mathscr{X} \in \mathbf{Prob}\langle\mathbf{G}\rangle$ we denote by $[\![\mathbf{G}]\!] = [\![\mathscr{X}]\!]$ the number of objects in the category $\mathbf{G}$.

An object O in a poset category $\mathbf{G}$ will be called a *source*, if it is not a target of any morphism except for the identity. Likewise a *sink* object is not a domain of any morphism, except for the identity morphism. If a category contains a *unique* source object, the object is called the initial object and such a category will be called complete.

The above terminology transfers to diagrams indexed by $\mathbf{G}$: A source space in $\mathscr{X} \in \mathbf{Prob}\langle\mathbf{G}\rangle$ is one that is not a target space of any reduction within the diagram, a sink space is not the domain of any non-trivial reduction and $\mathscr{X}$ is called complete if $\mathbf{G}$ is, i.e. if it has a unique source space.

The tensor product of probability spaces extends to a tensor product of diagrams. For $\mathscr{X}, \mathscr{Y} \in \mathbf{Prob}\langle\mathbf{G}\rangle$, such that $\mathscr{X} = \{X_i;\, f_{ij}\}$ and $\mathscr{Y} = \{Y_i;\, g_{ij}\}$ define

$$\mathscr{X} \otimes \mathscr{Y} := \{X_i \otimes Y_i;\, f_{ij} \otimes g_{ij}\}.$$

The construction of the category of commutative diagrams could be applied to any category, not just **Prob**. Two additional cases will be of interest to us.

Denote by **Set** the category of finite sets and surjective maps. Then all of the above constructions could be repeated for sets instead of probability spaces. Thus we could talk about the category of diagrams of sets $\mathbf{Set}\langle\mathbf{G}\rangle$.

Given a reduction $f\colon X \to Y$ between two probability spaces, the restriction $\underline{f}\colon \underline{X} \to \underline{Y}$ is a well-defined surjective map. Given a diagram $\mathscr{X} = \{X_i;\, f_{ij}\}$ of

probability spaces, there is an underlying diagram of sets, obtained by taking the supports of measures on each level and restricting reductions on these supports. We will denote it by $\underline{\mathscr{X}} = \left\{\underline{X}_i; \underline{f}_{ij}\right\}$, where $\underline{X}_i := \operatorname{supp} p_{X_i}$. Thus we have a forgetful functor

$$\underline{\cdot} : \mathbf{Prob}\langle\mathbf{G}\rangle \to \mathbf{Set}\langle\mathbf{G}\rangle.$$

We could also repeat the construction of commutative diagrams to form a category of diagrams of diagrams. Thus given two poset categories $\mathbf{G}$ and $\mathbf{H}$ we can form a category $\mathbf{Prob}\langle\mathbf{G}, \mathbf{H}\rangle := \mathbf{Prob}\langle\mathbf{G}\rangle\langle\mathbf{H}\rangle$. We will rarely need anything beyond a two-fan of diagrams. There is a natural isomorphism

$$\mathbf{Prob}\langle\mathbf{G}\rangle\langle\mathbf{H}\rangle \equiv \mathbf{Prob}\langle\mathbf{H}\rangle\langle\mathbf{G}\rangle.$$

Thus, for example, a two-fan of $\mathbf{G}$-diagrams could be equivalently considered as a $\mathbf{G}$-diagram of two-fans, see also Sect. 2.5.3.

2.5 Examples of diagrams

We now consider some examples of poset categories and corresponding diagrams, that will be important in what follows.

2.5.1 Singleton

We denote by $\bullet$ the poset category with a single object. Clearly diagrams indexed by $\bullet$ are just probability spaces and we have $\mathbf{Prob} \equiv \mathbf{Prob}\langle\bullet\rangle$.

2.5.2 Chains

The chain $\mathbf{C}_n$ of length $n \in \mathbb{N}$ is a category with n objects $\{O_i\}_{i=1}^n$ and morphisms from O_i to O_j whenever $i \geq j$. A diagram $\mathscr{X} \in \mathbf{Prob}\langle\mathbf{C}_n\rangle$ is a chain of reductions

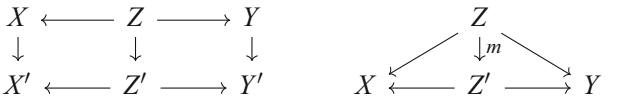
$$\mathscr{X} = (X_n \to X_{n-1} \to \cdots \to X_1).$$

2.5.3 Two-fan

The two-fan $\boldsymbol{\Lambda}_2$ is a category with three objects $\{O_1, O_{12}, O_2\}$ and two non-identity morphisms $O_{12} \to O_1$ and $O_{12} \to O_2$. A diagram indexed by a two-fan will also be called a two-fan.

Essentially, a two-fan $(X \leftarrow Z \to Y)$ is a triple of probability spaces and a pair of reductions between them.

A reduction of a two-fan $\mathscr{F} = (X \leftarrow Z \to Y)$ to another two-fan $\mathscr{F}' = (X' \leftarrow Z' \to Y')$ is a triple of reductions $Z \to Z'$, $Y \to Y'$ and $X \to X'$ that commute with the reductions within each fan, so that the diagram on Fig. 2a is commutative.



a Reduction of two-fans **b** Minimization of a two-fan

Fig. 2 Two-fans, their reductions and minimizations

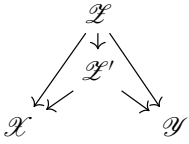
A two-fan $(X \leftarrow Z \rightarrow Y)$ is called *minimal* if for any super-diagram $\{X, Y, Z, Z'\}$ shown on Fig. 2b the reduction $m : Z \rightarrow Z'$ must be an isomorphism. Minimal two-fans are also called *couplings* in probability theory.

For any two-fan $(X \leftarrow Z \rightarrow Y)$ of probability spaces there always exist a unique (up to isomorphism), minimal two-fan $(X \leftarrow Z' \rightarrow Y)$, that can be included in the diagram shown on Fig. 2b. The minimization can be constructed by taking $Z' := \underline{X} \times \underline{Y}$ as a set and considering a probability distribution on Z' induced by a map $Z \rightarrow Z'$, that is the Cartesian product of the reductions $\underline{Z} \rightarrow \underline{X}$ and $\underline{Z} \rightarrow \underline{Y}$ in the original two-fan. Thus, the inclusion of a pair of probability spaces X and Y as sink vertices in a minimal two-fan is equivalent to specifying a joint distribution on $\underline{X} \times \underline{Y}$.

Note that minimality of a two-fan is defined in purely categorical terms. Even though the definition applies to two-fans of morphisms in any category, the minimization need not to exist. However as the next proposition asserts, if minimization of any two-fan exists in a category $\mathbf{C}$, then it also exists in a category of diagrams over $\mathbf{C}$.

Proposition 2.1 *Let* $\mathbf{G}$ *be a poset category, and let* $\mathscr{X} = \{X_i; a_{ij}\}$, $\mathscr{Y} = \{Y_i; b_{ij}\}$ *and* $\mathscr{Z} = \{Z_i; c_{ij}\}$ *be three* $\mathbf{G}$*-diagrams. Then*

1. *A two-fan* $\mathscr{F} = (\mathscr{X} \leftarrow \mathscr{Z} \rightarrow \mathscr{Y}) \in \mathbf{Prob}\langle \mathbf{G}, \boldsymbol{\Lambda}_2 \rangle$ *of* $\mathbf{G}$*-diagrams is minimal if and only if the constituent two-fans of probability spaces* $\mathscr{F}_i = (X_i \leftarrow Z_i \rightarrow Y_i)$ *are all minimal.*
2. *For any two-fan* $\mathscr{F} = (\mathscr{X} \leftarrow \mathscr{Z} \rightarrow \mathscr{Y})$ *of* $\mathbf{G}$*-diagrams its minimal reduction exists, that is, there exists a minimal two-fan* $\mathscr{F}' = (\mathscr{X} \leftarrow \mathscr{Z}' \rightarrow \mathscr{Y})$ *included in the following diagram*

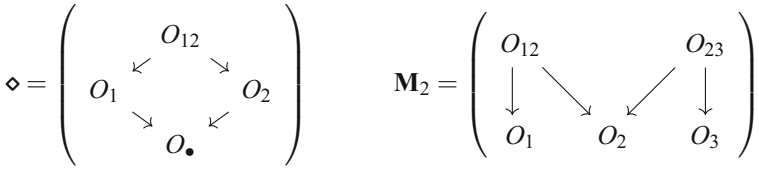


The proof of Proposition 2.1 can be found on page 274.

2.5.4 Co-fan

A co-fan $\mathbf{V}$ is a category with three objects and morphisms

$$\mathbf{V} = (O_1 \rightarrow O_\bullet \leftarrow O_2)\,.$$



a Diamond category **b** Two-tents category

Fig. 3 Diamond and two-tents categories

2.5.5 A diamond diagram

A "diamond" diagram is indexed by a diamond category $\diamond$, that consists of a two-fan and a co-fan, as shown on Fig. 3.

Of course, there is also a morphism $O_{12} \to O_\bullet$, which lies in the transitive closure of the given four morphisms. We will often skip writing morphisms that are implied by the transitive closure.

A diamond diagram will be called *minimal* if the top two-fan in it is minimal.

2.5.6 "Two-tents" diagram

The "two-tents" category $\mathbf{M}_2$ consists of five objects, of which two are sources and three are sinks, and morphisms are as in Fig. 3b.

Thus, a typical two-tents diagram consists of five probability spaces and reductions as in

$$\mathscr{X} = (X \leftarrow U \to Y \leftarrow V \to Z).$$

The probability spaces U and V are sources and X, Y and Z are sinks.

2.5.7 Full diagram

The full category $\boldsymbol{\Lambda}_n$ on n objects is a category with objects $\{O_I\}_{I \in 2^{\{1,\dots,n\}} \setminus \{\emptyset\}}$ indexed by all non-empty subsets $I \in 2^{\{1,\dots,n\}}$ and a morphism from O_I to O_J, whenever $J \subseteq I$.

A diagram $\mathscr{X}$ indexed by a full category will be called *minimal*, if for every two-fan in it, it also contains a minimal two-fan with the same sink vertices. If $\mathscr{X} \in \mathbf{Prob}\langle \boldsymbol{\Lambda}_n \rangle$ is minimal full diagram of probability spaces, then the set $\underline{\mathscr{X}(O_I)}$ can be considered as a subset of the product $\prod_{i \in I} \underline{\mathscr{X}(O_i)}$, while reductions are just coordinate projections.

For an n-tuple of random variables $\mathsf{X}_1, \dots, \mathsf{X}_n$ one may construct a minimal full diagram $\mathscr{X} \in \mathbf{Prob}\langle \boldsymbol{\Lambda}_n \rangle$ by considering all joint distributions and "marginalization" reductions. We denote such a diagram by $\langle \mathsf{X}_1, \dots, \mathsf{X}_n \rangle$. On the other hand, the reductions from the initial space to the sink vertices of a full diagram can be viewed as random variables on the domain of definition given by the (unique) initial space.

Suppose $\mathscr{X} \in \mathbf{Prob}\langle \mathbf{\Lambda}_n \langle$ is a minimal full diagram with sink vertices $X_1, \ldots, X_n$. It is convenient to view $\mathscr{X}$ as a distribution on the Cartesian product of the underlying sets of the sink vertices:

$$p_{\mathscr{X}} \in \Delta(\underline{X}_1 \times \cdots \times \underline{X}_n)$$

where ΔS stands for the space of all probability distribution on a finite set S.

Once the underlying sets of the sink spaces are fixed, there is a one-to-one correspondence between the full minimal diagrams and distributions as above.

As a corollary of Proposition 2.1 we also obtain the following characterization of minimal full diagrams of any **G**-diagrams of probability spaces.

Corollary 2.2 *Let* **G** *be an arbitrary poset category. Then*

1. *A full diagram* $\mathscr{F}$ *of* **G***-diagrams is minimal, if and only if the constituent full diagrams of probability spaces* $\mathscr{F}_i$ *are all minimal.*
2. *For any full diagram* $\mathscr{F} \in \mathbf{Prob}\langle \mathbf{G}, \mathbf{\Lambda}_n \rangle$ *of* **G***-diagrams there exists another minimal full diagram* $\mathscr{F}' \in \mathbf{Prob}\langle \mathbf{G}, \mathbf{\Lambda}_n \rangle$ *with the same sink entries and a reduction* $\mu \colon \mathscr{F} \to \mathscr{F}'$*, such that* μ *restricts to an isomorphism on sink entries of* $\mathscr{F}$*. Moreover,* $\mathscr{F}'$ *is unique upto isomorphism.*

2.6 Constant diagrams

Suppose X is a probability space and **G** is a poset category. One may form a *constant* **G***-diagram* by considering a functor that maps all objects in **G** to X and all the morphisms to the identity morphism $X \xrightarrow{\mathrm{Id}} X$. We denote such a constant diagram by $X^{\mathbf{G}}$ or simply by X, when **G** is clear from the context. Any constant diagram is automatically minimal.

If $\mathscr{Y} = \{Y_i;\, f_{ij}\}$ is another **G**-diagram, then a reduction $\rho \colon \mathscr{Y} \to X^{\mathbf{G}}$ (which we write sometimes simply as $\rho \colon \mathscr{Y} \to X$) is a collection of reductions $\rho_i \colon Y_i \to X$, such that

$$f_{ij} \circ \rho_i = \rho_j .$$

Let $\mathscr{X} = \{X_i;\, f_{ij}\}$ be a complete diagram with the initial space X_0. Then there is a canonical reduction

$$\rho \colon X_0^{\mathbf{G}} \to \mathscr{X}. \tag{4}$$

with components

$$\rho_i := f_{0i}.$$

By $\{\bullet\}$ we denote a one-point probability space. The constant **G**-diagram $\{\bullet\}^{\mathbf{G}}$ is a unit with respect to the product in $\mathbf{Prob}\langle \mathbf{G} \rangle$.

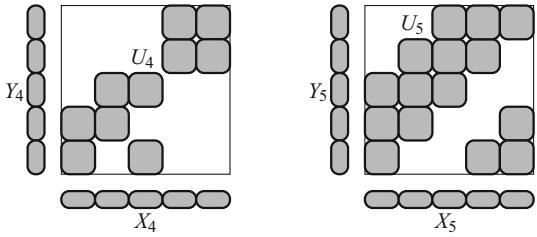


Fig. 4 Non-homogeneous two-fans consisting of uniform spaces

2.7 Homogeneous diagrams

A diagram $\mathscr{X} \in \mathbf{Prob}\langle\mathbf{G}\rangle$ indexed by some poset category $\mathbf{G}$ is called *homogeneous* if its automorphism group $\mathrm{Aut}(\mathscr{X})$ acts transitively on every probability space in $\mathscr{X}$. Three examples of homogeneous diagrams were given in the introduction. The subcategory of all homogeneous diagrams indexed by $\mathbf{G}$ will be denoted $\mathbf{Prob}\langle\mathbf{G}\rangle_{\mathbf{h}}$.

In fact, for $\mathscr{X}$ to be homogeneous it is sufficient that the $\mathrm{Aut}(\mathscr{X})$ acts transitively on every source space in $\mathscr{X}$. Thus, if $\mathscr{X}$ is complete with initial space X_0, to check homogeneity it is sufficient to check the transitivity of the action of the symmetries of $\mathscr{X}$ on X_0.

Any subdiagram of a homogeneous diagram is also homogeneous. In particular, all the individual spaces of a homogeneous diagram are homogeneous

$$\mathbf{Prob}\langle\mathbf{G}\rangle_{\mathbf{h}} \subset \mathbf{Prob}_{\mathbf{h}}\langle\mathbf{G}\rangle.$$

However homogeneity of the whole of the diagram is a stronger property than homogeneity of the individual spaces in the diagram, thus in general

$$\mathbf{Prob}\langle\mathbf{G}\rangle_{\mathbf{h}} \neq \mathbf{Prob}_{\mathbf{h}}\langle\mathbf{G}\rangle.$$

Two examples of non-homogeneous two-fans are shown in Fig. 4. The pictures are to be interpreted in the same way as the pictures in Fig. 1.

A single probability space is homogeneous if and only if there is a representative in its isomorphism class with uniform measure and the same holds true for chain diagrams, for the co-fan or any other diagram that does not contain a two-fan. However, for more complex diagrams, for example for two-fans, no such simple description is available.

2.7.1 Universal construction of homogeneous diagrams

Examples of homogeneous diagrams could be constructed in the following manner. Suppose Γ is a finite group and $\{H_i\}$ is a collection of subgroups. Consider a collection of sets $\underline{X}_i := \Gamma/H_i$ and consider a natural surjection $f_{ij} : \underline{X}_i \to \underline{X}_j$ whenever H_i is a subgroup of H_j. Equipping each $\underline{X}_i$ with the uniform distribution one can turn the diagram of sets $\{\underline{X}_i; f_{ij}\}$ into a homogeneous diagram of probability spaces. It will be complete if there is a smallest subgroup (under inclusion) among H_i's.

Such a diagram will be complete and minimal, if together with any pair of groups H_i and H_j in the collection, their intersection $H_i \cap H_j$ also belongs to the collection $\{H_i\}$.

In fact, any homogeneous diagram arises this way. Suppose diagram $\mathscr{X} = \{X_i;\, f_{ij}\}$ is homogeneous, then we set $\Gamma = \mathrm{Aut}(\mathscr{X})$ and choose a collection of points $x_i \in X_i$ such that $f_{ij}(x_i) = x_j$ and denote by $H_i := \mathrm{Stab}(x_i) \subset \Gamma$. Then, if one applies the construction of the previous paragraph to Γ, with the collection of subgroups $\{H_i\}$, one recovers the original diagram $\mathscr{X}$ upto isomorphism.

2.8 Conditioning

Suppose a diagram $\mathscr{X}$ contains a fan

$$\mathscr{F} = \left(X \xleftarrow{f} Z \xrightarrow{g} Y \right).$$

Given a point $x \in X$ with a non-zero weight one may consider *conditional probability distributions* $p_Z(\,\cdot\,\lfloor x)$ on $\underline{Z}$, and $p_Y(\,\cdot\,\lfloor x)$ on $\underline{Y}$. The distribution $p_Z(\,\cdot\,\lfloor x)$ is supported on $f^{-1}(x)$ and is defined by the property that for any function $f : Z \to \mathbb{R}$ holds

$$\int_Z f \,\mathrm{d}\, p_Z = \int_X \left[\int_Z f(z) \mathrm{d}\, p_Z(z\lfloor x) \right] \mathrm{d}\, p_X(x)$$

and is given by

$$p_Z(z\lfloor x) = \frac{p_Z(z)}{p_X(x)}.$$

The distribution $p_Y(\,\cdot\,\lfloor x)$ is the pushforward of $p_Z(\,\cdot\,\lfloor x)$ under g

$$p_Y(\,\cdot\,\lfloor x) = g_* p_Z(\,\cdot\,\lfloor x).$$

Recall that if $\mathscr{F}$ is minimal, the underlying set of Z can be assumed to be the product $\underline{X} \times \underline{Y}$. In that case

$$p_Y(y\lfloor x) = \frac{p_Z(x, y)}{p_X(x)}.$$

We denote the corresponding space $Y\lfloor x := \big(\underline{Y},\, p_Y(\,\cdot\,\lfloor x)\big)$.

Under some assumptions it is possible to condition a whole sub-diagram of $\mathscr{X}$. More specifically, if a diagram $\mathscr{X}$ contains a sub-diagram $\mathscr{Y}$ and a probability space X satisfying the condition that *there exists a space Z in $\mathscr{X}$ that reduces to all the spaces in $\mathscr{Y}$ and to X*, then we may condition the whole of $\mathscr{Y}$ on $x \in X$ given that $p_X(x) > 0$.

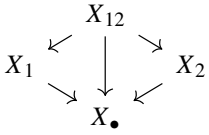
For $x \in X$ with positive weight we denote by $\mathscr{Y}\lfloor x$ the diagram of spaces in $\mathscr{Y}$ conditioned on $x \in X$. The diagram $\mathscr{Y}\lfloor x$ has the same combinatorial type as $\mathscr{Y}$ and will be called the *slice* of $\mathscr{Y}$ over $x \in X$. Note that the space X itself may or may not belong to $\mathscr{Y}$. The conditioning $\mathscr{Y}\lfloor x$ may depend on the choice of a fan between $\mathscr{Y}$ and X, however when $\mathscr{X}$ is complete the conditioning $\mathscr{Y}\lfloor x$ is well-defined and is independent of the choice of fans.

Suppose now that there are two subdiagram $\mathscr{Y}$ and $\mathscr{Z}$ in $\mathscr{X}$ and in addition $\mathscr{Z}$ is a constant diagram, $\mathscr{Z} = Z^{\mathbf{G}'}$ for some poset category $\mathbf{G}'$. Let $z \in \underline{Z}$, then $\mathscr{Y}\lfloor z$ is well defined and is independent of the choice of the space in $\mathscr{Z}$, the element of which z is to be considered.

If $\mathscr{X}$ is homogeneous, then $\mathscr{Y}\lfloor x$ is also homogeneous and its isomorphism class does not depend on the choice of $x \in \underline{X}$.

2.9 Entropy

We define entropy by the limit in Eq. 2. Entropy satisfies the so-called Shannon inequality, see for example [8]. Namely for any minimal diamond diagram



the following inequality holds,

$$\mathsf{Ent}(X_1) + \mathsf{Ent}(X_2) \geq \mathsf{Ent}(X_{12}) + \mathsf{Ent}(X_\bullet). \tag{5}$$

Furthermore, entropy is additive with respect to the tensor product, that is, for a pair of probability spaces $X, Y \in \mathbf{Prob}$ holds

$$\mathsf{Ent}(X \otimes Y) = \mathsf{Ent}(X) + \mathsf{Ent}(Y). \tag{6}$$

Conditional entropy $\mathsf{Ent}(X\lfloor Y)$ is defined for a pair X, Y of probability spaces included in a minimal two-fan $(X \leftarrow Z \rightarrow Y)$ as

$$\mathsf{Ent}(X\lfloor Y) := \mathsf{Ent}(Z) - \mathsf{Ent}(X).$$

The above quantity is always non-negative in view of Shannon inequality (5). Moreover, the following identity holds, see [8]

$$\mathsf{Ent}(X\lfloor Y) = \int_Y \mathsf{Ent}(X\lfloor y)\mathrm{d}\, p_Y(y). \tag{7}$$

For a $\mathbf{G}$-diagram $\mathscr{X} = \{X_i;\, f_{ij}\}$ define the entropy homomorphism

$$\mathsf{Ent}_*\colon\ \mathbf{Prob}\langle\mathbf{G}\rangle \to \mathbb{R}^{[\![G]\!]}, \quad \{X_i;\, f_{ij}\} \mapsto \big(\mathsf{Ent}(X_i)\big) \in \mathbb{R}^{[\![G]\!]}.$$

It will be convenient for us to equip the target $\mathbb{R}^{[\![\mathbf{G}]\!]}$ with the ℓ^1-norm. Thus

$$|\mathsf{Ent}_*(\mathscr{X})|_1 = \sum_{i=1}^{[\![\mathbf{G}]\!]} \mathsf{Ent}(X_i).$$

If $\mathscr{X}$ is a complete $\mathbf{G}$-diagram with initial space X_0, then by Shannon inequality (5) there is an obvious estimate

$$\mathsf{Ent}(X_0) \leq |\mathsf{Ent}_*(\mathscr{X})|_1 \leq [\![\mathscr{X}]\!] \cdot \mathsf{Ent}(X_0).$$

3 The entropy distance

We turn the space of diagrams into a pseudo-metric space by introducing the intrinsic entropy distance and asymptotic entropy distance. The intrinsic entropy distance is obtained by taking an infimum of the entropy distance over all possible joint distributions on two probability spaces.

3.1 Entropy distance and asymptotic entropy distance

3.1.1 Entropy distance in the case of single probability spaces

For a two-fan $\mathscr{F} = (X \leftarrow Z \rightarrow Y)$ define a "distance" $\mathrm{kd}(\mathscr{F})$ between probability spaces X and Y with respect to $\mathscr{F}$ by

$$\begin{aligned} \mathrm{kd}(\mathscr{F}) &:= \mathsf{Ent}(Z\lfloor Y) + \mathsf{Ent}(Z\lfloor X) \\ &= 2\mathsf{Ent}(Z) - \mathsf{Ent}(X) - \mathsf{Ent}(Y). \end{aligned}$$

If a two-fan $\mathscr{F}$ satisfies $\mathrm{kd}(\mathscr{F}) = 0$, then both reductions in $\mathscr{F}$ are isomorphisms. Thus, essentially $\mathrm{kd}(\mathscr{F})$ is some measure of the deviation of the statistical map defined by $\mathscr{F}$ from being a deterministic bijection between X and Y.

The minimal reduction $\mathscr{F}'$ of $\mathscr{F}$ satisfies

$$\mathrm{kd}(\mathscr{F}') \leq \mathrm{kd}(\mathscr{F}). \tag{8}$$

For a pair of probability spaces X, Y define the *intrinsic entropy distance* as

$$\mathbf{k}(X, Y) := \inf \left\{ \mathrm{kd}(\mathscr{F}) : \mathscr{F} = (X \leftarrow Z \rightarrow Y) \text{ is a two-fan} \right\}. \tag{9}$$

The optimization takes place over all two-fans with sink spaces X and Y. In view of inequality (8) one could as well optimize over the space of *minimal* two-fans, which we will also refer to as *couplings* between X and Y. The tensor product of X and Y trivially provides a coupling and the set of couplings is compact, therefore an optimum is always achieved and it is finite.

The bivariate function $\mathbf{k}\colon \mathbf{Prob} \times \mathbf{Prob} \to \mathbb{R}_{\geq 0}$ defines a notion of pseudo-distance and it vanishes exactly on pairs of isomorphic probability spaces. This follows directly from the Shannon inequality (5), and a more general statement will be proven in Proposition 3.1 below.

3.1.2 Entropy distance for complete diagrams

The definition of entropy distance for complete diagrams repeats almost literally the definition for single spaces. We fix a complete poset category $\mathbf{G}$ and will be considering diagrams from $\mathbf{Prob}\langle\mathbf{G}\rangle$.

Consider three such diagrams $\mathscr{X} = \{X_i, f_{ij}\}$, $\mathscr{Y} = \{Y_i, g_{ij}\}$ and $\mathscr{Z} = \{Z_i, h_{ij}\}$ from $\mathbf{Prob}\langle\mathbf{G}\rangle$. Recall that a two-fan $\mathscr{F} = (\mathscr{X} \leftarrow \mathscr{Z} \rightarrow \mathscr{Y})$ can also be viewed as a $\mathbf{G}$-diagram of two-fans

$$\mathscr{F}_i = (X_i \leftarrow Z_i \rightarrow Y_i).$$

Define

$$\begin{aligned} \mathrm{kd}(\mathscr{F}) &:= \sum_i \mathrm{kd}(\mathscr{F}_i) \\ &= \sum_i \big(2\mathsf{Ent}(Z_i) - \mathsf{Ent}(X_i) - \mathsf{Ent}(Y_i)\big). \end{aligned}$$

The quantity $\mathrm{kd}(\mathscr{F})$ vanishes if and only if the fan $\mathscr{F}$ provides isomorphisms between all individual spaces in $\mathscr{X}$ and $\mathscr{Y}$ that commute with the inner structure of the diagrams, that is, it provides an isomorphism between $\mathscr{X}$ and $\mathscr{Y}$ in $\mathbf{Prob}\langle\mathbf{G}\rangle$.

The *intrinsic entropy distance between diagrams* is defined in analogy with the case of single probability spaces

$$\mathbf{k}(\mathscr{X}, \mathscr{Y}) := \inf\left\{\mathrm{kd}(\mathscr{F}) : \mathscr{F} = (\mathscr{X} \leftarrow \mathscr{Z} \rightarrow \mathscr{Y})\right\},$$

where the infimum is over all two-fans of $\mathbf{G}$-diagrams with sink vertices $\mathscr{X}$ and $\mathscr{Y}$.

The following proposition records that the intrinsic entropy distance is in fact a pseudo-distance on $\mathbf{Prob}\langle\mathbf{G}\rangle$, provided $\mathbf{G}$ is a complete poset category (that is when $\mathbf{G}$ has a unique initial space).

Proposition 3.1 *Let $\mathbf{G}$ be a complete poset category. Then the bivariate function*

$$\boldsymbol{k}\colon \mathbf{Prob}\langle\mathbf{G}\rangle \times \mathbf{Prob}\langle\mathbf{G}\rangle \to \mathbb{R}$$

is a pseudo-distance on $\mathbf{Prob}\langle\mathbf{G}\rangle$.

Moreover, two diagrams $\mathscr{X}, \mathscr{Y} \in \mathbf{Prob}\langle\mathbf{G}\rangle$ satisfy $\boldsymbol{k}(\mathscr{X}, \mathscr{Y}) = 0$ if and only if $\mathscr{X}$ is isomorphic to $\mathscr{Y}$ in $\mathbf{Prob}\langle\mathbf{G}\rangle$.

The idea of the proof is very simple. In the case of single probability spaces X, Y, Z a coupling between X and Z can be constructed from a coupling between X and Y and

a coupling between Y and Z by *adhesion* on Y, see [16]. The triangle inequality then follows from Shannon inequality. However, since we are dealing with diagrams the combinatorial structure requires careful treatment. Therefore, we provide a detailed proof on page 276.

It is important to note, that the proof uses the fact that $\mathbf{G}$ is complete. In fact, even though the definition of $\mathbf{k}$ could be easily extended to some bivariate function on the space of diagrams of any fixed combinatorial type, it fails to satisfy the triangle inequality in general, because the composition of couplings requires completeness of $\mathbf{G}$.

3.1.3 The asymptotic entropy distance

Let $\mathbf{G}$ be a complete poset category. We will show in Corollary 3.5 below, that the sequence

$$n \mapsto \mathbf{k}(\mathscr{X}^n, \mathscr{Y}^n)$$

is sublinear and therefore the following limit exists.

$$\boldsymbol{\kappa}(\mathscr{X}, \mathscr{Y}) := \lim_{n\to\infty} \frac{1}{n}\mathbf{k}(\mathscr{X}^n, \mathscr{Y}^n). \tag{10}$$

We call it's value, $\boldsymbol{\kappa}(\mathscr{X}, \mathscr{Y})$, the *asymptotic entropy distance* between two diagrams $\mathscr{X}, \mathscr{Y} \in \mathbf{Prob}\langle\mathbf{G}\rangle$.

As a corollary of Proposition 3.1 and definition (10) we immediately obtain that the asymptotic entropy distance is a homogeneous pseudo-distance on $\mathbf{Prob}\langle\mathbf{G}\rangle$.

Corollary 3.2 *Let* $\mathbf{G}$ *be a complete poset category. Then the bivariate function*

$$\boldsymbol{\kappa}\colon\ \mathbf{Prob}\langle\mathbf{G}\rangle \times \mathbf{Prob}\langle\mathbf{G}\rangle \to \mathbb{R}$$

is a pseudo-distance on $\mathbf{Prob}\langle\mathbf{G}\rangle$ *satisfying the following properties*: *for any pair of diagrams* $\mathscr{X}, \mathscr{Y} \in \mathbf{Prob}\langle\mathbf{G}\rangle$

1. $\boldsymbol{\kappa}(\mathscr{X}, \mathscr{Y}) \leq \boldsymbol{k}(\mathscr{X}, \mathscr{Y})$
2. *for any* $n \in \mathbb{N}_0$ *holds* $\boldsymbol{\kappa}(\mathscr{X}^n, \mathscr{Y}^n) = n \cdot \boldsymbol{\kappa}(\mathscr{X}, \mathscr{Y})$.

We will see later that there are instances when $\boldsymbol{\kappa} < \mathbf{k}$, moreover there are pairs of non-isomorphic diagrams with vanishing asymptotic entropy distance between them.

In the next subsection we derive some elementary properties of the intrinsic entropy distance and the asymptotic entropy distance.

3.2 Properties of (asymptotic) entropy distance

3.2.1 Tensor product

We show that the tensor product on the space of diagrams is 1-Lipschitz. Later this will allow us to give a simple description of tropical diagrams, that is, of points in the

asymptotic cone of $\mathbf{Prob}\langle\mathbf{G}\rangle$, as limits of certain sequences of "classical" diagrams, as will be discussed in a subsequent article.

Proposition 3.3 *Let* $\mathbf{G}$ *be a complete poset category. Then with respect to the Kolmogorov distance on* $\mathbf{Prob}\langle\mathbf{G}\rangle$ *the tensor product*

$$\otimes : (\mathbf{Prob}\langle\mathbf{G}\rangle, \boldsymbol{k})^2 \to (\mathbf{Prob}\langle\mathbf{G}\rangle, \boldsymbol{k})$$

is 1*-Lipschitz in each variable, that is, for every triple* $\mathscr{X}, \mathscr{Y}, \mathscr{Y}' \in \mathbf{Prob}\langle\mathbf{G}\rangle$ *the following bound holds*

$$\boldsymbol{k}(\mathscr{X} \otimes \mathscr{Y}, \mathscr{X} \otimes \mathscr{Y}') \leq \boldsymbol{k}(\mathscr{Y}, \mathscr{Y}').$$

This statement is a direct consequence of additivity of entropy with respect to the tensor product. Details can be found on page 279.

It follows directly from definition (10) and Proposition 3.3, that the asymptotic entropy distance enjoys a similar property.

Corollary 3.4 *Let* $\mathbf{G}$ *be a complete poset category. Then with respect to the asymptotic entropy distance on* $\mathbf{Prob}\langle\mathbf{G}\rangle$ *the tensor product*

$$\otimes : (\mathbf{Prob}\langle\mathbf{G}\rangle, \boldsymbol{\kappa})^2 \to (\mathbf{Prob}\langle\mathbf{G}\rangle, \boldsymbol{\kappa})$$

is 1*-Lipschitz in each variable.*

As another corollary we obtain the subadditivity properties of the intrinsic entropy distance and asymptotic entropy distance.

Corollary 3.5 *Let* $\mathbf{G}$ *be a complete poset category and let* $\mathscr{X}, \mathscr{Y}, \mathscr{U}, \mathscr{V} \in \mathbf{Prob}\langle\mathbf{G}\rangle$*, then*

$$\boldsymbol{k}(\mathscr{X} \otimes \mathscr{U}, \mathscr{Y} \otimes \mathscr{V}) \leq \boldsymbol{k}(\mathscr{X}, \mathscr{Y}) + \boldsymbol{k}(\mathscr{U}, \mathscr{V})$$

and

$$\boldsymbol{\kappa}(\mathscr{X} \otimes \mathscr{U}, \mathscr{Y} \otimes \mathscr{V}) \leq \boldsymbol{\kappa}(\mathscr{X}, \mathscr{Y}) + \boldsymbol{\kappa}(\mathscr{U}, \mathscr{V}).$$

It implies in particular that shifts are non-expanding maps in $(\mathbf{Prob}\langle\mathbf{G}\rangle, \mathbf{k})$ or $(\mathbf{Prob}\langle\mathbf{G}\rangle, \boldsymbol{\kappa})$.

Corollary 3.6 *Let* $\mathbf{G}$ *be a complete poset category and* $\boldsymbol{\delta} = \boldsymbol{k}, \boldsymbol{\kappa}$ *be either intrinsic entropy distance or asymptotic entropy distance on* $\mathbf{Prob}\langle\mathbf{G}\rangle$*. Let* $\mathscr{U} \in \mathbf{Prob}\langle\mathbf{G}\rangle$*. Then the shift map*

$$\mathscr{U} \otimes \cdot : (\mathbf{Prob}\langle\mathbf{G}\rangle, \boldsymbol{\delta}) \to (\mathbf{Prob}\langle\mathbf{G}\rangle, \boldsymbol{\delta}), \quad \mathscr{X} \mapsto \mathscr{U} \otimes \mathscr{X}$$

is a non-expanding map with respect to either intrinsic entropy distance or asymptotic entropy distance.

Less obvious is the fact that $\boldsymbol{\kappa}$ is, in fact, translation invariant and in particular, $(\mathbf{Prob}\langle\mathbf{G}\rangle, \boldsymbol{\kappa})$ satisfies the cancellation property. This is the subject of Proposition 3.7 below, which was communicated to us by Tobias Fritz.

Proposition 3.7 *For any triple of diagrams $\mathscr{X}, \mathscr{Y}, \mathscr{U}$ holds*

$$\boldsymbol{\kappa}(\mathscr{X} \otimes \mathscr{U}, \mathscr{Y} \otimes \mathscr{U}) = \boldsymbol{\kappa}(\mathscr{X}, \mathscr{Y}).$$

The proof of the lemma can be found on page 280.

3.2.2 Entropy

Recall that we defined the entropy function

$$\mathsf{Ent}_*\colon \mathbf{Prob}\langle\mathbf{G}\rangle \to \mathbb{R}^{[\![\mathbf{G}]\!]}$$

by evaluating the entropy of all individual spaces in a $\mathbf{G}$-diagram. The target space $\mathbb{R}^{[\![\mathbf{G}]\!]}$ will be endowed with the ℓ^1-norm with respect to the natural coordinate system. With such a choice, the entropy function is 1-Lipschitz with respect to the Kolmogorov distance on $\mathbf{Prob}\langle\mathbf{G}\rangle$.

Proposition 3.8 *Suppose $\mathbf{G}$ is a complete poset category and $\boldsymbol{\delta} = \boldsymbol{k}$, $\boldsymbol{\kappa}$ is either intrinsic entropy distance or asymptotic entropy distance on $\mathbf{Prob}\langle\mathbf{G}\rangle$. Then the entropy function*

$$\mathit{Ent}_*\colon (\mathbf{Prob}\langle\mathbf{G}\rangle, \boldsymbol{\delta}) \to (\mathbb{R}^{[\![\mathbf{G}]\!]}, \ |\cdot|_1), \quad \mathscr{X} = \left\{X_i, f_{ij}\right\} \mapsto (\mathit{Ent}X_i)_i \in \mathbb{R}^{[\![\mathbf{G}]\!]}$$

is 1*-Lipschitz.*

Again, the proof of the proposition above is an application of Shannon's inequality, see page 281 for details.

3.3 The Slicing Lemma

The Slicing Lemma, Proposition 3.9 below, allows to estimate the intrinsic entropy distance between two diagrams with the integrated intrinsic entropy distance between "slices", which are diagrams obtained by conditioning on another probability space. It turned out to be a very powerful tool for estimation of the intrinsic entropy distance and will be used below on several occasions.

As described in Sect. 2.6, by a reduction of a diagram $\mathscr{X} = \left\{X_i, f_{ij}\right\}$ to a single space U we mean a collection of reductions $\left\{\rho_i : X_i \to U\right\}$ from the individual spaces in $\mathscr{X}$ to U, that commute with the reductions within $\mathscr{X}$

$$\rho_j \circ f_{ij} = \rho_i.$$

Alternatively, whenever a single probability space appears as a domain or a target of a morphism to or from a $\mathbf{G}$-diagram, it should be replaced by a constant $\mathbf{G}$-diagram.

Proposition 3.9 (Slicing Lemma) *Suppose* $\mathbf{G}$ *is a complete poset category and we are given* $\mathscr{X}, \hat{\mathscr{X}}, \mathscr{Y}, \hat{\mathscr{Y}} \in \hat{\mathbf{Prob}}\langle\mathbf{G}\rangle$*—four* $\mathbf{G}$*-diagrams and* $U, V, W \in \mathbf{Prob}$*—three probability spaces, that are included into the following three-tents diagram*

$$\begin{array}{ccccccccc} & & \hat{\mathscr{X}} & & & & \hat{\mathscr{Y}} & & \\ & \swarrow & & \searrow & & \swarrow & & \searrow & \\ \mathscr{X} & & & & U & & V & & \mathscr{Y} \end{array} \qquad W$$

such that the two-fan $(U \leftarrow W \rightarrow V)$ *is minimal. Then the following estimate holds*

$$\begin{aligned} \boldsymbol{k}(\mathscr{X}, \mathscr{Y}) \leq \int_W & \boldsymbol{k}(\mathscr{X}\lfloor u, \mathscr{Y}\lfloor v) d\, p_W(u, v) \\ & + [\![\mathbf{G}]\!] \cdot kd(U \leftarrow W \rightarrow V) \\ & + \sum_i \big[Ent(U\lfloor X_i) + Ent(V\lfloor Y_i) \big]. \end{aligned}$$

The idea of the proof of the Slicing Lemma (page 281) is as follows. For every pair $(u, v) \in \underline{W}$ we consider an optimal two-fan $\mathscr{G}_{uv}$ coupling $\mathscr{X}\lfloor u$ and $\mathscr{Y}\lfloor v$. These fans have the same underlying diagram of sets. Then we construct a coupling between $\mathscr{X}$ and $\mathscr{Y}$ as a convex combination of distributions of $\mathscr{G}_{uv}$'s weighted by $p_W(u, v)$. The estimates on the resulting two-fan then imply the proposition.

Various implications of the Slicing Lemma are summarized in the next corollary.

Corollary 3.10 *Let* $\mathbf{G}$ *be a complete poset category,* $\mathscr{X}, \mathscr{Y} \in \mathbf{Prob}\langle\mathbf{G}\rangle$ *and* $U \in \mathbf{Prob}$*.*

1. *Given a "two-tents" diagram*

$$\mathscr{X} \leftarrow \hat{\mathscr{X}} \rightarrow U \leftarrow \hat{\mathscr{Y}} \rightarrow \mathscr{Y}$$

the following inequality holds

$$\boldsymbol{k}(\mathscr{X}, \mathscr{Y}) \leq \int_U \boldsymbol{k}(\mathscr{X}\lfloor u, \mathscr{Y}\lfloor u) d\, p_U(u) + 2 \cdot [\![\mathbf{G}]\!] \cdot Ent(U).$$

2. *Given a fan*

$$\mathscr{X} \leftarrow \hat{\mathscr{X}} \rightarrow U$$

the following inequality holds

$$\boldsymbol{k}(\mathscr{X}, \mathscr{Y}) \leq \int_{UV} \boldsymbol{k}(\mathscr{X}\lfloor u, \mathscr{Y}) d\, p_U(u) + 2 \cdot [\![\mathbf{G}]\!] \cdot Ent(U).$$

3. *Let* $\mathscr{X} \rightarrow U$ *be a reduction, then*

$$\boldsymbol{k}(\mathscr{X}, \mathscr{Y}) \leq \int_U \boldsymbol{k}(\mathscr{X}\lfloor u, \mathscr{Y}) d\, p_U(u) + [\![\mathbf{G}]\!] \cdot Ent(U).$$

4. *For a co-fan* $\mathscr{X} \to U \leftarrow \mathscr{Y}$ *holds*

$$\boldsymbol{k}(\mathscr{X}, \mathscr{Y}) \leq \int_U \boldsymbol{k}(\mathscr{X}\lfloor u, \mathscr{Y}\lfloor u) d\, p_U(u).$$

4 Distributions and types

In this section we recall some elementary inequalities for (relative) entropies and the total variation distance for distributions on finite sets. Furthermore, we generalize the notion of a probability distribution on a set to a distribution on a diagram of sets. Finally, we give a perspective on the theory of types, and also introduce types in the context of complete diagrams.

4.1 Distributions

4.1.1 Single probability spaces

For a finite set S we denote by ΔS the collection of all probability distributions on S. It is a unit simplex in the real vector space $\mathbb{R}^S$. We often use the fact that it is a compact, convex set, whose interior points correspond to fully supported probability measures on S.

For $\pi_1, \pi_2 \in \Delta S$ denote by $|\pi_1 - \pi_2|_1$ the total variation of the signed measure $(\pi_1 - \pi_2)$ and define the entropy of the distribution π_1 by

$$h(\pi_1) := -\sum_{x \in X} \pi_1(x) \ln \pi_1(x). \tag{11}$$

If, in addition, π_2 lies in the interior of ΔS define the relative entropy by

$$D(\pi_1 \,||\, \pi_2) := \sum_{x \in X} \pi_1(x) \ln \frac{\pi_1(x)}{\pi_2(x)}.$$

The entropy of a probability space is often defined through formula (11). It is a standard fact, and can be verified with the help of Lemma 4.2 below, that for $\pi \in \Delta S$ holds

$$h(\pi) = \mathsf{Ent}(S, \pi) \tag{12}$$

which justifies the name "entropy" for the function $h\colon \Delta S \to \mathbb{R}$.

Define a *divergence ball* of radius $\varepsilon > 0$ centered at $\pi \in \mathrm{Interior}\Delta S$ as

$$B_\varepsilon(\pi) := \left\{\pi' \in \Delta S : D(\pi' \,||\, \pi) \leq \varepsilon\right\}. \tag{13}$$

For a fixed π and $\varepsilon \ll 1$ the ball $B_\varepsilon(\pi)$ also lies in the interior of ΔS.

The total variation norm and relative entropy are related by the following inequality.

Lemma 4.1 *Let S be a finite set, then for any $\pi_1, \pi_2 \in \Delta S$, Pinsker's inequality holds*

$$|\pi_1 - \pi_2|_1 \leq \sqrt{2D(\pi_1 \,||\, \pi_2)}.$$

The claim of the Lemma, Pinsker's inequality, is a well-known inequality in for instance information theory, and a proof can be found in [8].

4.1.2 Distributions on diagrams

A map $f\colon S \to S'$ between two finite sets induces an affine map $f_*\colon \Delta S \to \Delta S'$.

For a diagram of sets $\mathscr{S} = \{S_i;\, f_{ij}\}$ we define the *space of distributions on the diagram* $\mathscr{S}$ by

$$\Delta\mathscr{S} := \left\{ (\pi_i) \in \prod_i \Delta S_i \,:\, (f_{ij})_*\pi_i = \pi_j \right\}.$$

Essentially, an element of $\Delta\mathscr{S}$ is a collection of distributions on the sets S_i in $\mathscr{S}$ that is consistent with respect to the maps f_{ij}. The consistency conditions $(f_{ij})_*\pi_i = \pi_j$ form a collection of linear equations with integer coefficients with respect to the standard convex coordinates in $\prod \Delta S_i$. Thus, $\Delta\mathscr{S}$ is a rational affine subspace in the product of simplices. In particular, $\Delta\mathscr{S}$ has a convex structure.

If $\mathscr{S}$ is complete with initial set S_0, then specifying a distribution $\pi_0 \in \Delta S_0$ uniquely determines distributions on all of the S_i's by setting $\pi_i := (f_{0i})_*\pi_0$. In such a situation we have

$$\Delta\mathscr{S} \cong \Delta S_0.$$

If $\mathscr{S}$ is not complete and $S_0, \ldots, S_k$ is a collection of its source sets, then $\Delta\mathscr{S}$ is isomorphic to an affine subspace of the product $\Delta S_0 \times \cdots \times \Delta S_k$ cut out by linear equations with integer coefficients corresponding to co-fans in $\mathscr{S}$ with source sets among $S_0, \ldots, S_k$.

To simplify notation, for a probability space X or a diagram $\mathscr{X}$ we will write

$$\Delta X := \Delta\underline{X}$$
$$\Delta\mathscr{X} := \Delta\underline{\mathscr{X}}.$$

4.2 Types

We now discuss briefly the theory of types. Types are special subspaces of tensor powers that consist of sequences with the same "empirical distribution" as explained in details below. For a more detailed discussion the reader is referred to [7,8]. We generalize the theory of types to complete diagrams of sets and complete diagrams of probability spaces.

The theory of types for diagrams, that are not complete, is more complex and will be addressed in a subsequent article.

4.2.1 Types for single probability spaces

Let S be a finite set. For $n \in \mathbb{N}$ denote also

$$\Delta^{(n)} S := \Delta S \cap \frac{1}{n}\mathbb{Z}^S$$

a collection of rational points in ΔS with denominator n. (We say that a rational number $r \in \mathbb{Q}$ has denominator $n \in \mathbb{N}$ if $r \cdot n \in \mathbb{Z}$)

Define the *empirical distribution map* $\mathbf{q} : S^n \to \Delta S$, that sends $(s_i)_{i=1}^n = \mathbf{s} \in S^n$ to the empirical distribution $\mathbf{q}(\mathbf{s}) \in \Delta S$ given by

$$\mathbf{q}(\mathbf{s})(a) = \frac{1}{n} \cdot \left| \{i : s_i = a\} \right| \quad \text{for any } a \in S.$$

Clearly the image of $\mathbf{q}$ lies in $\Delta^{(n)} S$.

For $\pi \in \Delta^{(n)} S$, the space $T^n_\pi S := \mathbf{q}^{-1}(\pi)$ equipped with the uniform measure is called a *type* over π. The symmetric group $\mathbb{S}_n$ acts on S^n by permuting the coordinates. This action leaves the empirical distribution invariant and therefore could be restricted to each type, where it acts transitively. Thus, for $\pi \in \Delta^{(n)} S$ the probability space $(T^n_\pi S, u)$ with u being a uniform ($\mathbb{S}_n$-invariant) distribution, is a homogeneous space.

Suppose $X = (\underline{X}, p)$ is a probability space. Let τ_n be the pushforward of $p^{\otimes n}$ under the empirical distribution map $\mathbf{q}\colon \underline{X}^n \to \Delta X$. Clearly $\operatorname{supp} \tau_n \subset \Delta^{(n)} X$, thus $(\Delta X, \tau_n)$ is a finite probability space. Therefore we have a reduction

$$\mathbf{q}\colon X^n \to (\Delta X, \tau_n) \tag{14}$$

which we call the *empirical reduction*. If $\pi \in \Delta^{(n)} X$ is such that $\tau_n(\pi) > 0$, then

$$T^n_\pi \underline{X} = X^n \lfloor \pi.$$

In particular, it follows that the right-hand side does not depend on the probability p on X as long as π is "compatible" to it.

The following lemma records some standard facts about types, which can be checked by elementary combinatorics and found in [8].

Lemma 4.2 *Let X be a probability space and $\mathbf{x} \in X^n$, then*

1. $|\Delta^{(n)} X| = \binom{n + |X|}{|X|} \le \mathbf{e}^{|X| \cdot \ln(n+1)} = \mathbf{e}^{O(|X| \cdot \ln n)}$
2. $p^{\otimes n}(\mathbf{x}) = \mathbf{e}^{-n\left[h(\mathbf{q}(\mathbf{x})) + D(\mathbf{q}(\mathbf{x}) \,||\, p)\right]}$
3. $\mathbf{e}^{n \cdot h(\pi) - |X| \cdot \ln(n+1)} \le |T^n_\pi \underline{X}| \le \mathbf{e}^{n \cdot h(\pi)}$ *or* $|T^n_\pi \underline{X}| = \mathbf{e}^{n \cdot h(\pi) + O(|X| \cdot \ln n)}$
4. $\mathbf{e}^{-n \cdot D(\pi \,||\, p) - |X| \cdot \ln(n+1)} \le \tau_n(\pi) = p^{\otimes n}(T^n_\pi \underline{X}) \le \mathbf{e}^{-n \cdot D(\pi \,||\, p)}$ *or* $\tau_n(\pi) = \mathbf{e}^{-n \cdot D(\pi \,||\, p) + O(|X| \cdot \ln n)}$

If $X = (\underline{X}, p_X)$ is a probability space with rational probability distribution with denominator n, then the type over p_X will be called the *true type* of X

$$T^n X := T^n_{p_X} \underline{X}.$$

As a corollary to Lemma 4.2 and equation (12) we obtain the following.

Corollary 4.3 *For a finite set S and $\pi \in \Delta^{(n)} S$ holds*

$$n \cdot h(\pi) - |S| \cdot \ln(n+1) \leq \mathit{Ent}(T^n_\pi S) \leq n \cdot h(\pi).$$

Also, for a finite probability space $X = (S, p)$ with a rational distribution p with denominator n holds

$$n \cdot \mathit{Ent}(X) - |X| \cdot \ln(n+1) \leq \mathit{Ent}(T^n X) \leq n \cdot \mathit{Ent}(X).$$

In particular,

$$\mathit{Ent}(T^n X) = n \cdot \mathit{Ent}(X) + o(|X| \cdot n).$$

The following important theorem is known as Sanov's theorem. It can be easily derived from Lemma 4.2 or a proof can be found in [8].

Theorem 4.4 (Sanov's Theorem) *Let $X = (S, p)$ be a finite probability space and let $\mathbf{q}\colon X^n \to (\Delta X, \tau_n)$ be the empirical reduction. Then for every $r > 0$,*

$$\tau_n(\Delta X \setminus B_r(p)) \leq \mathbf{e}^{-n \cdot r + |X| \cdot \ln(n+1)}$$

where $B_r(p)$ is the divergence ball (relative entropy ball) defined in (13).

Combining the estimate in Theorem 4.4 with the Pinsker's inequality in 4.1 we obtain the following corollary.

Corollary 4.5 *For a finite probability space $X = (S, p)$ holds*

$$\tau_n(\{\pi : |\pi - p|_1 \geq r\}) \leq \mathbf{e}^{-\frac{1}{2} n \cdot r^2 + O(|X| \cdot \ln n)}.$$

4.3 Types for complete diagrams

In this subsection we generalize the theory of types for diagrams indexed by a complete poset category. The theory for a non-complete diagrams is more complex and will be addressed in our future work. Before we describe our approach we need some preparatory material.

Suppose we have a reduction $f\colon X \to Y$ between a pair of probability spaces. Then for any $n \in \mathbb{N}$ there is an induced reduction $f_*\colon (\Delta X, \tau_n) \to (\Delta Y, \tau_n)$ that can be included in the following diamond diagram

$$\begin{array}{ccc} X^n & \xrightarrow{f^n} & Y^n \\ \downarrow{\scriptstyle \mathbf{q}} & & \downarrow{\scriptstyle \mathbf{q}} \\ (\Delta X, \tau_n) & \xrightarrow{f_*} & (\Delta Y, \tau_n) \end{array}$$

that satisfies certain special condition, namely, the sides of the diamond are independent conditioned on the bottom space

$$\Delta X \perp\!\!\!\perp Y^n \lfloor \Delta Y.$$

In particular, for any $\pi \in \Delta X$ with $\tau_n(\pi) > 0$ and $\pi' = f_*\pi \in \Delta Y$ holds

$$Y^n \lfloor \pi = Y^n \lfloor \pi' \tag{15}$$

and there is a well-defined reduction

$$Tf\colon T^n_\pi X \to T^n_{\pi'} Y$$

for any $\pi \in \Delta^{(n)} X$ and $\pi' = f_*\pi \in \Delta^{(n)} Y$.

Now we are ready to give the definitions of types. Let $\mathscr{X} \in \mathbf{Prob}\langle \mathbf{G} \rangle$ be a complete diagram, $\mathscr{X} = \{X_i;\, f_{ij}\}$ with initial space X_0 and let $\pi \in \Delta^{(n)} \mathscr{X}$.

Define the type $T^n_\pi \underline{\mathscr{X}}$ as the $\mathbf{G}$-diagram, whose individual spaces are types of the individual spaces of $\mathscr{X}$ over the corresponding push-forwards of π

$$T^n_\pi \underline{\mathscr{X}} := \left\{ T^n_{\pi_i} \underline{X}_i;\, Tf_{ij} \right\}.$$

Consider a symmetric group $\mathbb{S}_n$ acting on $\mathscr{X}^n$ by automorphisms permuting the coordinates. The action leaves the types $T^n_\pi \underline{\mathscr{X}}$ invariant and it is transitive on the initial space $T^n_\pi X_0$. Thus, each type $T^n_\pi \underline{\mathscr{X}}$ is a homogeneous diagram.

4.3.1 The empirical two-fan

Unlike in the cases of single probability spaces there is no empirical reduction from the power of $\mathscr{X}$ to $\Delta\mathscr{X}$. It will be convenient for us to see the types as the power of the diagram conditioned on a distribution. This is achieved by including the power of diagram into a *empirical two-fan*.

Given a $\mathbf{G}$-diagram $\mathscr{X}$ with initial space X_0 we construct the associated empirical two-fan with sink vertices $\mathscr{X}^n$ and $(\Delta\mathscr{X}, \tau_n)^{\mathbf{G}}$ as the "composition" of the canonical reduction $(X_0)^{\mathbf{G}} \longrightarrow \mathscr{X}$, Eq. (4) in Sect. 2.6, and the empirical reduction $X_0^n \longrightarrow \Delta X_0 \cong \Delta\mathscr{X}$ in Eq. (14).

$$\mathscr{Q}_n(\mathscr{X}) = \left(\begin{array}{ccc} & (X_0^{\mathbf{G}})^{(n)} & \\ {\scriptstyle f_{0*}^n}\swarrow & & \searrow{\scriptstyle \mathbf{q}^{\mathbf{G}}} \\ \mathscr{X}^n & & (\Delta^{(n)}\mathscr{X}, \tau_n)^{\mathbf{G}} \end{array} \right) \tag{16}$$

The two-fan $\mathscr{Q}_n$ is not necessarily minimal, but its minimal reduction can be constructed using Lemma 2.2 on page 252.

Let $\pi_0 \in (\Delta\mathscr{X}, \tau_n)$ with $\tau_n(\pi) > 0$ and $\pi_i = f_{0i}\pi_0$. Then within $\mathscr{Q}_n$ holds

$$T_\pi^n X_i = X_i^n \lfloor \pi_i = X_i^n \lfloor \pi_0$$

by Eq. (15) and therefore

$$\mathscr{X}^n \lfloor \pi = T_\pi^n \mathscr{X}.$$

For every $n \in \mathbb{N}$ and $\pi \in \Delta^{(n)}\underline{X}_0$ the type $T_\pi^n\mathscr{X}$ is a homogeneous diagram. Suppose that a complete diagram $\mathscr{X}$ is such that the probability distribution p_0 on the initial set is rational with the denominator n, then we call $T_p^n\underline{\mathscr{X}}$ the *true type* of $\mathscr{X}$ and denote

$$T^n\mathscr{X} := T_{p_0}^n\underline{\mathscr{X}}.$$

5 Distance between types

Our goal in this section is to estimate the intrinsic entropy distance between two types over two different distributions $\pi_1, \pi_2 \in \Delta^{(n)}\mathscr{S}$ in terms of the total variation distance $|\pi_1 - \pi_2|_1$.

For this purpose we use a "lagging" technique which is explained below. Practically, we couple different types by randomly removing and inserting the appropriate amount of symbols to pass from a trajectory of the one type to a trajectory of the other.

5.1 The lagging trick

Let Λ_α be a binary probability space,

$$\Lambda_\alpha := \big(\{\square, \blacksquare\}\,;\, p_{\Lambda_\alpha}(\blacksquare) = \alpha \big)$$

and let $\mathscr{X} = \{(\underline{X}_i, p_i); f_{ij}\}$, $\mathscr{Z} = \{(\underline{Z}_i, q_i); g_{ij}\}$ be two diagrams indexed by a complete poset category $\mathbf{G}$ and included in a minimal two-fan, i.e a coupling,

$$(\Lambda_\alpha)^{\mathbf{G}} \xleftarrow{\lambda} \mathscr{Z} \xrightarrow{\rho} \mathscr{X}.$$

Assume further that the distribution q on $\mathscr{Z}$ is rational with denominator $n \in \mathbb{N}$, that is $q \in \Delta^{(n)}\underline{\mathscr{Z}}$. It follows that p and p_{Λ_α} are also rational with the same denominator n.

We construct a *lagging two-fan*

$$\mathscr{L} := \left(T^{(1-\alpha)n}(\mathscr{X}\lfloor\square) \xleftarrow{l} T^n\mathscr{Z} \xrightarrow{T\rho} T^n\mathscr{X}\right) \tag{17}$$

as follows. The right leg $T\rho$ of $\mathscr{L}$ is induced by the right leg ρ of the original two-fan. The left leg

$$l\colon T^n\mathscr{Z} \to T^{(1-\alpha)n}(\mathscr{X}\lfloor\square)$$

is obtained by erasing symbols that reduce to $\blacksquare$ and applying ρ to the remaining symbols. The target space for the reduction l is the true type of $\mathscr{X}\lfloor\square$ which is "lagging" behind $T^n\mathscr{Z}$ by a factor of $(1-\alpha)$. More specifically, the reduction l is constructed as follows.

Let $\lambda_j\colon Z_j \to \Lambda_\alpha$ be the components of the reduction $\lambda\colon \mathscr{Z} \to \Lambda_\alpha$. Given $\bar{z} = (z_i)_{i=1}^n \in T^n Z_j$ define the subset of indices

$$I_{\bar{z}} := \{i : \lambda_j(z_i) = \square\}$$

and define the jth component of l by

$$l_j\left((z_i)_{i=1}^n\right) := \left(\rho(z_i)\right)_{i\in I_{\bar{z}}}.$$

By equivariance each l_j is a reduction of homogeneous spaces, since the inverse image of any point has the same cardinality. Moreover the reductions l_j commute with the reductions in $T^n\mathscr{Z}$ as explained in Sect. 4.3 and therefore l is a reduction of diagrams.

The next lemma uses the lagging two-fan to estimate the intrinsic entropy distance between its sink diagrams.

Lemma 5.1 *Let $\mathscr{X}, \mathscr{Z} \in \mathbf{Prob}\langle\mathbf{G}\rangle$ be two diagrams indexed by a complete poset category $\mathbf{G}$ and included in a minimal two-fan*

$$\Lambda_\alpha \xleftarrow{\lambda} \mathscr{Z} \xrightarrow{\rho} \mathscr{X}$$

where distribution on $\mathscr{Z}$ is rational with denominator $n \in \mathbb{N}$. Then

$$\boldsymbol{k}\left(T^{(1-\alpha)n}(\mathscr{X}\lfloor\square),\ T^n\mathscr{X}\right)$$

$$\leq n \cdot [\![G]\!] \cdot [2Ent(\Lambda_\alpha) + \alpha \cdot \ln |X_0|] + 2 \cdot [\![\mathbf{G}]\!] \cdot |X_0| \cdot \ln(n+1)$$
$$= n \cdot [\![G]\!] \cdot \left[2Ent(\Lambda_\alpha) + \alpha \cdot \ln |X_0|\right] + O\left(|X_0| \cdot \ln n\right).$$

It is an immediate consequence of the Slicing Lemma, in particular Corollary 3.10 part (2) that

$$\mathbf{k}\Big(\mathscr{X}\lfloor\square\,,\ \mathscr{X}\Big) \leq [\![G]\!] \cdot [2\mathsf{Ent}(\Lambda_\alpha) + \alpha \cdot \ln |X_0|]\,.$$

By the subadditivity of the intrinsic entropy distance,

$$\mathbf{k}\Big((\mathscr{X}\lfloor\square)^{\otimes n}\,,\ \mathscr{X}^{\otimes n}\Big) \leq n \cdot [\![G]\!] \cdot [2\mathsf{Ent}(\Lambda_\alpha) + \alpha \cdot \ln |X_0|]\,.$$

This bound is almost the estimate in Lemma 5.1, except Lemma 5.1 estimates the distance between types rather than tensor powers. We will soon see that tensor powers and types are very close in the intrinsic entropy distance. However, for the purpose of the proof of Lemma 5.1, it suffices to know that their entropies are close, an estimate that is provided by Corollary 4.3.

Proof of Lemma 5.1 We will use the lagging two-fan constructed in Eq. (17), namely

$$\mathscr{L} := \left(T^{(1-\alpha)n}(\mathscr{X}\lfloor\square) \xleftarrow{l} T^n\mathscr{Z} \xrightarrow{T\rho} T^n\mathscr{X}\right)$$

as a coupling to estimate the intrinsic entropy distance

$$\mathbf{k}\big(T^{(1-\alpha)n}(\mathscr{X}\lfloor\square)\,,\ T^n\mathscr{X}\big) \leq \mathrm{kd}(\mathscr{L}).$$

Recall that by Corollary 4.3 for a probability space X with a rational distribution we have

$$n \cdot \mathsf{Ent}(X) - |X| \cdot \ln(n+1) \leq \mathsf{Ent}(T^n X) \leq n \cdot \mathsf{Ent}(X).$$

Thus we can estimate $\mathrm{kd}(\mathscr{L})$ as follows

$$\begin{aligned}\mathrm{kd}(\mathscr{L}) &= \sum_i \Big[\big(\mathsf{Ent}(T^n Z_i) - \mathsf{Ent}(T^n X_i)\big) \\ &\quad + (\mathsf{Ent}(T^n Z_i) - \mathsf{Ent}(T^{(1-\alpha)n}(X_i\lfloor\square)))\Big] \\ &\leq n \cdot \sum_i \Big[(\mathsf{Ent}(Z_i) - \mathsf{Ent}(X_i)) + (\mathsf{Ent}(Z_i) - (1-\alpha)\mathsf{Ent}(X_i\lfloor\square))\Big] \\ &\quad + 2 \cdot [\![\mathbf{G}]\!] \cdot |X_0| \cdot \ln(n+1).\end{aligned}$$

By minimality of the original two-fan and Shannon inequality (5) we have a bound

$$\mathsf{Ent}(Z_i) - \mathsf{Ent}(X_i) \leq \mathsf{Ent}(\Lambda_\alpha).$$

The second part in the sum can be estimated using relation (7) as follows

$$\begin{aligned}\mathsf{Ent}(Z_i) - (1-\alpha)\mathsf{Ent}(X_i\lfloor\square) &= \mathsf{Ent}(\Lambda_\alpha) + \mathsf{Ent}(X_i\lfloor\Lambda_\alpha) - (1-\alpha)\mathsf{Ent}(X_i\lfloor\square)\\ &= \mathsf{Ent}(\Lambda_\alpha) + (1-\alpha)\mathsf{Ent}(X_i\lfloor\square) + \alpha\mathsf{Ent}(X_i\lfloor\blacksquare) - (1-\alpha)\mathsf{Ent}(X_i\lfloor\square)\\ &\leq \mathsf{Ent}(\Lambda_\alpha) + \alpha\cdot\ln|X_i|.\end{aligned}$$

Combining all of the above we obtain the estimate in the conclusion of the lemma. □

5.2 Distance between types

In this section we use the lagging trick as described above to estimate the distance between types over two different distributions in $\Delta\mathscr{S}$ where $\mathscr{S}$ is a complete diagram of sets.

Proposition 5.2 *Suppose $\mathscr{S}$ is a complete* $\mathbf{G}$*-diagram of sets with initial set S_0. Suppose $p, q \in \Delta^{(n)}\mathscr{S}$ and let $\alpha = \frac{1}{2}|p_0 - q_0|_1$. Then*

$$\begin{aligned}\boldsymbol{k}(T_p^n\mathscr{S}, T_q^n\mathscr{S}) &\leq 2n\cdot[\![\mathbf{G}]\!]\cdot[\alpha\cdot\ln|S_0| + 2\mathit{Ent}(\Lambda_\alpha)] + 4[\![\mathbf{G}]\!]\cdot|X_0|\cdot\ln(n+1)\\ &= 2n\cdot[\![\mathbf{G}]\!]\cdot[\alpha\cdot\ln|S_0| + 2\mathit{Ent}(\Lambda_\alpha)] + O(|X_0|\cdot\ln n).\end{aligned}$$

The idea of the proof is to write p and q as a convex combination of a common distribution $\hat{p}$ and "small amounts" of p^+ and q^+, respectively. Then we use the lagging trick to estimate distances between types over p and $\hat{p}$, as well as between types over q and $\hat{p}$. We now present details of the proof.

Proof of Proposition 5.2 Recall that for a complete diagram $\mathscr{S}$ with initial set S_0 we have

$$\Delta\mathscr{S} \cong \Delta S_0. \tag{18}$$

Our goal now is to write p and q as the convex combination of three other distributions $\hat{p}$, p^+ and q^+ as in

$$\begin{aligned}p &= (1-\alpha)\cdot\hat{p} + \alpha\cdot p^+\\ q &= (1-\alpha)\cdot\hat{p} + \alpha\cdot q^+.\end{aligned}$$

We could do it the following way. Let $\alpha := \frac{1}{2}|p_0 - q_0|_1$. If $\alpha = 1$ then the proposition follows trivially by constructing a tensor-product fan, so from now on we assume that $\alpha < 1$. Define three probability distributions $\hat{p}_0$, p_0^+ and q_0^+ on S_0 by setting for every $x \in S_0$

$$\begin{aligned}\hat{p}_0(x) &:= \frac{1}{1-\alpha}\min\{p_0(x), q_0(x)\}\\ p_0^+ &:= \frac{1}{\alpha}(p_0 - (1-\alpha)\hat{p}_0)\end{aligned}$$

$$q_0^+ := \frac{1}{\alpha}(q_0 - (1-\alpha)\hat{p}_0).$$

Denote by $\hat{p}, p^+, q^+ \in \Delta\mathscr{S}$ the distributions corresponding to $\hat{p}_0, p_0^+, q_0^+ \in \Delta S_0$ under the affine isomorphism (18). Thus we have

$$\begin{aligned} p &= (1-\alpha)\cdot\hat{p} + \alpha\cdot p^+ \\ q &= (1-\alpha)\cdot\hat{p} + \alpha\cdot q^+. \end{aligned}$$

Now we construct a pair of two-fans of **G**-diagrams

$$\Lambda_\alpha \longleftarrow \tilde{\mathscr{X}} \longrightarrow \mathscr{X} \quad \text{and} \quad \Lambda_\alpha \longleftarrow \tilde{\mathscr{Y}} \longrightarrow \mathscr{Y} \tag{19}$$

by setting

$$\begin{aligned} \mathscr{X} &:= (\mathscr{S}, p) \\ \mathscr{Y} &:= (\mathscr{S}, q) \\ \tilde{X}_i &:= \Big(S_i \times \underline{\Lambda}_\alpha;\ \tilde{p}_i(s,\square) = (1-\alpha)\hat{p}_i(s),\ \tilde{p}_i(s,\blacksquare) = \alpha\cdot p_i^+(s)\Big) \\ \tilde{Y}_i &:= \Big(S_i \times \underline{\Lambda}_\alpha;\ \tilde{q}_i(s,\square) = (1-\alpha)\hat{p}_i(s),\ \tilde{q}_i(s,\blacksquare) = \alpha\cdot q_i^+(s)\Big) \\ \tilde{\mathscr{X}} &:= \Big\{\tilde{X}_i;\ f_{ij}\times \mathrm{Id}\Big\} \\ \tilde{\mathscr{Y}} &:= \Big\{\tilde{Y}_i;\ f_{ij}\times \mathrm{Id}\Big\}. \end{aligned}$$

The reductions in (19) are given by coordinate projections. We have the following isomorphisms

$$\mathscr{X}\lfloor\square \cong \mathscr{Y}\lfloor\square \cong (\mathscr{S}, \hat{p}).$$

To estimate the distance between types we now apply Lemma 5.1 to the fans in (19)

$$\begin{aligned} \mathbf{k}(T_p^n\mathscr{S}, T_q^n\mathscr{S}) &= \mathbf{k}(T^n\mathscr{X}, T^n\mathscr{Y}) \\ &\le \mathbf{k}\big(T^n\mathscr{X}, T^{(1-\alpha)n}(\mathscr{X}\lfloor\square)\big) + \mathbf{k}\big(T^{(1-\alpha)n}(\mathscr{Y}\lfloor\square), T^n\mathscr{Y}\big) \\ &\le 2n\cdot[\![\mathbf{G}]\!]\cdot[\alpha\cdot\ln|S_0| + 2\mathsf{Ent}(\Lambda_\alpha)] + 4[\![\mathbf{G}]\!]\cdot|X_0|\cdot\ln(n+1). \end{aligned}$$

□

6 Asymptotic equipartition property for diagrams

Below we prove that any Bernoulli sequence of complete diagrams can be approximated by a sequence of homogeneous diagrams. This is essentially the *Asymptotic Equipartition Theorem for diagrams*.

Theorem 6.1 *Suppose $\mathscr{X} \in \mathbf{Prob}\langle \mathbf{G} \rangle$ is a complete diagram of probability spaces. Then there exists a sequence $\bar{\mathscr{H}} = (\mathscr{H}_n)_{n=0}^{\infty}$ of homogeneous diagrams of the same combinatorial type as $\mathscr{X}$ such that for all $n \geq |X_0|$*

$$\frac{1}{n}\boldsymbol{k}(\mathscr{X}^{\otimes n}, \mathscr{H}_n) \leq C(|X_0|, [\![\mathbf{G}]\!]) \cdot \sqrt{\frac{\ln^3 n}{n}} \tag{20}$$

where X_0 is the initial space of $\mathscr{X}$ and $C(|X_0|, [\![\mathbf{G}]\!])$ is a constant only depending on $|X_0|$ and $[\![\mathbf{G}]\!]$.

Proof Denote by $\mathscr{S} = \underline{\mathscr{X}}$ the underlying diagram of sets and by $p_{\mathscr{X}}$ the true distribution on $\mathscr{S}$, such that

$$\mathscr{X} = (\mathscr{S}, p_X).$$

We will construct the approximating homogeneous sequence by taking types over rational approximations of $p_{\mathscr{X}}$ in $\Delta\mathscr{S}$, that converge sufficiently fast to the true distribution $p_{\mathscr{X}}$.

More specifically, we select rational distributions $p_n \in \Delta^{(n)}\mathscr{S}$ such that

$$|p_n - p_{\mathscr{X}}|_1 \leq \frac{|S_0|}{n}.$$

As homogeneous spaces $\mathscr{H}_n$ we set $\mathscr{H}_n = T^n_{p_n}\mathscr{S}$. We will show that the intrinsic entropy distance between $\mathscr{H}_n$ and $\mathscr{X}^n$ satisfies the required estimate (20).

First we apply slicing along the empirical two-fan

$$\mathscr{Q}_n(\mathscr{X}) = \left(\mathscr{X}^n \leftarrow \tilde{\mathscr{X}}^{(n)} \rightarrow (\Delta\mathscr{S}, \tau_n)^{\mathbf{G}}\right)$$

defined in Sect. 4.3, Eq. (16) on page 267.

For the estimate below we use the fact that

$$\mathsf{Ent}(\Delta\mathscr{S}, \tau_n) \leq \ln|\Delta^{(n)}\mathscr{S}| \leq |S_0| \cdot \ln(n+1).$$

By slicing (see Corollary 3.10(2)) along the empirical two-fan we have

$$\begin{aligned}\mathbf{k}(T^n_{p_n}\mathscr{S}, \mathscr{X}^{\otimes n}) &\leq 2 \cdot [\![\mathbf{G}]\!] \cdot \mathsf{Ent}(\Delta\mathscr{S}, \tau_n) + \int_{\Delta\mathscr{S}} \mathbf{k}(T^n_{p_n}\mathscr{S}, T^n_{\pi}\mathscr{S})\mathrm{d}\,\tau_n(\pi) \\ &\leq 2 \cdot [\![\mathbf{G}]\!] \cdot |S_0| \cdot \ln(n+1) + \int_{\Delta\mathscr{S}} \mathbf{k}(T^n_{p_n}\mathscr{S}, T^n_{\pi}\mathscr{S})\mathrm{d}\,\tau_n(\pi).\end{aligned}$$

To estimate the integral we split the domain into a small divergence ball $B_{\varepsilon_n} = B_{\varepsilon_n}(p_{\mathscr{X}})$ around the "true" distribution and its complement

$$\int_{\Delta\mathscr{S}} \mathbf{k}(T^n_{p_n}\mathscr{S}, T^n_{\pi}\mathscr{S})\mathrm{d}\,\tau_n(\pi) = \int_{\Delta\mathscr{S}\setminus B_{\varepsilon_n}} \mathbf{k}(T^n_{p_n}\mathscr{S}, T^n_{\pi}\mathscr{S})\mathrm{d}\,\tau_n(\pi) \tag{21}$$

$$+\int_{B_{\varepsilon_n}} \mathbf{k}(T^n_{p_n}\mathscr{S}, T^n_{\pi}\mathscr{S})\mathrm{d}\,\tau_n(\pi) \tag{22}$$

and we set the radius ε_n equal to

$$\varepsilon_n := (|S_0|+1)\frac{\ln(n+1)}{n}.$$

To estimate the first integral on the right-hand side of equality (22) note that the distance between two types over the same diagram of sets can always be crudely estimated by

$$2\cdot\ln|S_0|\cdot[\![\mathbf{G}]\!]\cdot n.$$

Moreover, by Sanov's theorem, Theorem 4.1, we can estimate the empirical measure of the complement of the divergence ball

$$\tau_n(\Delta\mathscr{S}\setminus B_{\varepsilon_n}) \le \mathbf{e}^{-n\cdot\varepsilon_n+|S_0|\cdot\ln(n+1)} \le \frac{1}{n}$$

where we used the definition of ε_n to conclude the last inequality. Therefore we obtain

$$\begin{aligned}\int_{\Delta\mathscr{S}\setminus B_{\varepsilon_n}} \mathbf{k}(T^n_{p_n}\mathscr{S}, T^n_{\pi}\mathscr{S})\mathrm{d}\,\tau_n(\pi) &\le 2\cdot\ln|S_0|\cdot[\![\mathbf{G}]\!]\cdot n\cdot\tau_n(\Delta\mathscr{S}\setminus B_{\varepsilon_n})\\ &\le 2\cdot\ln|S_0|\cdot[\![\mathbf{G}]\!].\end{aligned}$$

Define

$$\alpha_n = \frac{|S_0|}{n}+\sqrt{2\varepsilon_n}$$

if the right-hand side is smaller than 1 and set $\alpha_n = 1$ otherwise. Then every $\pi \in B_{\varepsilon_n}(p_{\mathscr{X}})$ satisfies $|p_n-\pi| \le 2\alpha_n$ by Pinsker's inequality (Lemma 4.1), and the triangle inequality. Consequently, by the estimate on the distance between types in Proposition 5.2

$$\begin{aligned}&\int_{B_{\varepsilon_n}} \mathbf{k}(T^n_{p_n}\mathscr{S}, T^n_{\pi}\mathscr{S})\mathrm{d}\,\tau_n(\pi)\\ &\le 2n\cdot[\![\mathbf{G}]\!]\cdot\left(\alpha_n\ln|S_0|+2\mathsf{Ent}(\Lambda_{\alpha_n})\right)+4\cdot[\![\mathbf{G}]\!]\cdot|S_0|\cdot\ln(n+1).\end{aligned}$$

Using the definition of α_n and ε_n we find that

$$\int_{B_{\varepsilon_n}} \mathbf{k}(T^n_{p_n}\mathscr{S}, T^n_{\pi}\mathscr{S})\mathrm{d}\,\tau_n(\pi) = \mathsf{O}\left(\sqrt{n\cdot\ln^3 n}\right)$$

and hence combining the above estimates

$$\frac{1}{n}\mathbf{k}(T^n_{p_n}\mathscr{S}, \mathscr{X}^{\otimes n}) = \mathsf{O}\left(\sqrt{\frac{\ln^3 n}{n}}\right).$$

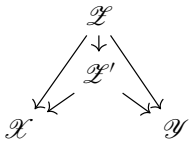
A more precise check shows that for $n \geq |S_0|$, the constants appearing in O only depend on $|S_0|$ and $[\![\mathbf{G}]\!]$. □

7 Technical proofs

This section contains some proofs that did not make it into the main text. The numbering of the claims in this section coincides with the numbering in the main text. Lemma that first appear in this section are numbered within section.

Proposition 2.1 *Let* $\mathbf{G}$ *be a poset category, and let* $\mathscr{X} = \{X_i; a_{ij}\}$, $\mathscr{Y} = \{Y_i; b_{ij}\}$ *and* $\mathscr{Z} = \{Z_i; c_{ij}\}$ *be three* $\mathbf{G}$*-diagrams. Then*

1. *A two-fan* $\mathscr{F} = (\mathscr{X} \leftarrow \mathscr{Z} \rightarrow \mathscr{Y}) \in \mathbf{Prob}\langle \mathbf{G}, \boldsymbol{\Lambda}_2\rangle$ *of* $\mathbf{G}$*-diagrams is minimal if and only if the constituent two-fans of probability spaces* $\mathscr{F}_i = (X_i \leftarrow Z_i \rightarrow Y_i)$ *are all minimal.*
2. *For any two-fan* $\mathscr{F} = (\mathscr{X} \leftarrow \mathscr{Z} \rightarrow \mathscr{Y})$ *of* $\mathbf{G}$*-diagrams its minimal reduction exists, that is, there exists a minimal two-fan* $\mathscr{F}' = (\mathscr{X} \leftarrow \mathscr{Z}' \rightarrow \mathscr{Y})$ *included in the following diagram*



Before we go to the proof of Proposition 2.1, we will need the following lemma.

Lemma 7.1 *Suppose we are given a pair of two-fans of probability spaces*

$$\mathscr{F} = (X \xleftarrow{\alpha} Z \xrightarrow{\beta} Y)$$
$$\mathscr{F}'' = (X'' \xleftarrow{\alpha''} Z'' \xrightarrow{\beta''} Y'')$$

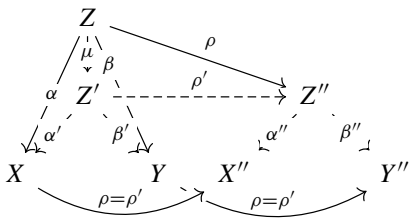
such that $\mathscr{F}''$ *is minimal. Let*

$$\mathscr{F} \xrightarrow{\mu} \mathscr{F}' = (X \xleftarrow{\alpha'} Z' \xrightarrow{\beta'} Y)$$

be a minimal reduction of $\mathscr{F}$*. Then for any reduction* $\rho : \mathscr{F} \rightarrow \mathscr{F}''$*, there exists a reduction* $\rho' : \mathscr{F}' \rightarrow \mathscr{F}''$ *such that* $\rho = \rho' \circ \mu$.

Proof of Lemma 7.1 We define ρ' on the sink spaces of $\mathscr{F}'$ to coincide with ρ.

To prove the lemma we just need to provide a dashed arrow that makes the following diagram commutative



The reduction ρ' is constructed by simple diagram chasing and by using the minimality of $\mathscr{F}''$. Suppose $z' \in \underline{Z'}$ and $z_1, z_2 \in \underline{Z}$ are such that $z' = \mu(z_1) = \mu(z_2)$. By commutativity of the solid arrows in the diagram above, we have

$$\alpha'' \circ \rho(z_1) = \rho \circ \alpha' \circ \mu(z_1) = \rho \circ \alpha' \circ \mu(z_2) = \alpha'' \circ \rho(z_2).$$

Similarly

$$\beta'' \circ \rho(z_1) = \beta'' \circ \rho(z_2).$$

Thus by minimality of $\mathscr{F}''$ it follows that $\rho(z_1) = \rho(z_2)$. Hence, ρ' can be constructed by setting $\rho'(z') = \rho(z_1)$. This finishes the proof of Lemma 7.1. □

Proof of Proposition 2.1 First we address claim (1) of the Proposition. Let $\mathbf{G} = \{O_i; m_{ij}\}$ be a poset category, $\mathscr{X}, \mathscr{Y}, \mathscr{Z} \in \mathbf{Prob}\langle\mathbf{G}\rangle$ be three $\mathbf{G}$-diagrams and $\mathscr{F} = (\mathscr{X} \leftarrow \mathscr{Z} \rightarrow \mathscr{Y})$ be a two-fan. Recall that it can also be considered as a $\mathbf{G}$-diagram of two-fans

$$\mathscr{F} = \{\mathscr{F}_i; f_{ij}\}.$$

Any minimizing reduction

$$\mathscr{F} = (\mathscr{X} \leftarrow \mathscr{Z} \rightarrow \mathscr{Y}) \longrightarrow \mathscr{F}' = (\mathscr{X} \leftarrow \mathscr{Z}' \rightarrow \mathscr{Y})$$

induces reductions

$$\mathscr{F}_i = (X_i \leftarrow Z_i \rightarrow Y_i) \longrightarrow \mathscr{F}_i = (X_i \leftarrow Z_i' \rightarrow Y_i)$$

for all i in the index set I. It follows that if all $\mathscr{F}_i$'s are minimal, then so is $\mathscr{F}$.

Now we prove the implication in the other direction. Suppose $\mathscr{F}$ is minimal. We have to show that all $\mathscr{F}_i$ are minimal as well. Suppose there exist a non-minimal fan among $\mathscr{F}_i$'s. For an index $i \in I$ let

$$\check{J}(i) := \{j \in I : \mathrm{Hom}_{\mathbf{G}}(O_j, O_i) \neq \emptyset\}$$

$$\hat{J}(i) := \left\{ j \in I \,:\, \mathrm{Hom}_{\mathbf{G}}(O_i, O_j) \neq \emptyset \right\}.$$

Choose an index i_0 such that

1. $\mathscr{F}_{i_0}$ is not minimal
2. for any $j \in \hat{J}(i_0)\backslash\{i_0\}$ the two-fan $\mathscr{F}_j$ is minimal.

Consider now the minimal reduction $\mu\colon \mathscr{F}_{i_0} \to \mathscr{F}'_{i_0}$ and construct a two-fan $\mathscr{G} = \left\{\mathscr{G}_i; g_{ij}\right\}$ of $\mathbf{G}$-diagrams by setting

$$\mathscr{G}_i := \begin{cases} \mathscr{F}'_i & \text{if } i = i_0 \\ \mathscr{F}_i & \text{otherwise} \end{cases}$$

and

$$g_{ij} := \begin{cases} \mu \circ f_{ij} & \text{if } j = i_0 \text{ and } i \in \check{J}(i_0) \\ f'_{ij} & \text{if } i = i_0 \text{ and } j \in \hat{J}(i_0) \\ f_{ij} & \text{otherwise} \end{cases}$$

where $f'_{i_0 j}$ is the reduction provided by the Lemma 7.1 applied to the diagram

$$\begin{array}{ccc} \mathscr{F}_{i_0} & & \\ \downarrow \mu & \searrow^{f_{i_0 j}} & \\ \mathscr{F}'_{i_0} & \overset{f'_{i_0 j}}{\dashrightarrow} & \mathscr{F}_j \end{array}$$

We thus constructed a non-trivial reduction $\mathscr{F} \to \mathscr{G}$ which is identity on the sink $\mathbf{G}$-diagrams $\mathscr{X}$ and $\mathscr{Y}$. This contradicts the minimality of $\mathscr{F}$.

To address the second assertion of the Lemma 2.1 observe that the argument above gives an algorithm for the construction of a minimal reduction of any two-fan of $\mathbf{G}$-diagrams. □

Proposition 3.1 *Let* $\mathbf{G}$ *be a complete poset category. Then the bivariate function*

$$\boldsymbol{k}\colon\ \mathbf{Prob}\langle\mathbf{G}\rangle \times \mathbf{Prob}\langle\mathbf{G}\rangle \to \mathbb{R}$$

is a pseudo-distance on $\mathbf{Prob}\langle\mathbf{G}\rangle$.

Moreover, two diagrams $\mathscr{X}, \mathscr{Y} \in \mathbf{Prob}\langle\mathbf{G}\rangle$ *satisfy* $\boldsymbol{k}(\mathscr{X}, \mathscr{Y}) = 0$ *if and only if* $\mathscr{X}$ *is isomorphic to* $\mathscr{Y}$ *in* $\mathbf{Prob}\langle\mathbf{G}\rangle$.

Proof The symmetry of $\mathbf{k}$ is immediate. The non-negativity of $\mathbf{k}$ follows from the fact that entropy of the target space of a reduction is not greater then the entropy of the domain, which is a particular instance of the Shannon inequality (5).

We proceed to prove the triangle inequality. We will make use of the following lemma

Lemma 7.2 *For a minimal full diagram of probability spaces*

$$\mathscr{Q} = \begin{pmatrix} & & Q & & \\ & \swarrow & \downarrow & \searrow & \\ U & & W & & V \\ \downarrow & \searrow\!\swarrow & & \swarrow\!\searrow & \downarrow \\ X & & Y & & Z \end{pmatrix}$$

holds

$$kd(X \leftarrow W \rightarrow Z) \leq kd(X \leftarrow U \rightarrow Y) + kd(Y \leftarrow V \rightarrow Z).$$

The Lemma 7.2 follows immediately from Shannon inequality.

Suppose for now that $\mathbf{G} = \bullet$ and we are given three probability spaces X, Y, Z together with the optimal couplings $\mathscr{U} = (X \leftarrow U \rightarrow Y)$ and $\mathscr{V} = (Y \leftarrow V \rightarrow Z)$ in the sense of optimization problem (9). Together they form a two-tents diagram $\mathscr{T} = (X \leftarrow U \rightarrow Y \leftarrow V \rightarrow Z)$. If we can extend $\mathscr{T}$ to a minimal full diagram $\mathscr{Q}$ as in the assumption of Lemma 7.2, the triangle inequality would follow. The diagram $\mathscr{Q} = \mathbf{ad}(\mathscr{T})$ can be constructed by the so called adhesion, as explained below.

As explained in Sect. 2.5.7, to construct a minimal full diagram with sink vertices X, Y and Z it is sufficient to provide a distribution on $\underline{Q} := \underline{X} \times \underline{Y} \times \underline{Z}$ with the correct push-forwards. We do this by setting

$$p_Q(x, y, z) := \frac{p_U(x, y) \cdot p_V(y, z)}{p_Y(y)}.$$

It is straightforward to check that the appropriate restriction of the full diagram defined in the above manner is indeed the original two-tents diagram. Essentially, to extend we need to provide a relationship (coupling) between spaces X and Z and we do it by declaring X and Z independent conditioned on Y. This is an instance of operation called *adhesion*, see [16]. Thus we have shown that $\mathbf{k}\colon \mathbf{Prob} \times \mathbf{Prob} \rightarrow \mathbb{R}$ is a pseudo-distance.

Assume now that $\mathbf{G}$ is an arbitrary complete poset category. Suppose $\mathscr{X} = \{X_i; f_{ij}\}$, $\mathscr{Y} = \{Y_i; g_{ij}\}$ and $\mathscr{Z} = \{Z_i; h_{ij}\}$ are $\mathbf{G}$-diagrams, with initial spaces being X_0, Y_0 and Z_0, respectively. Let

$$\hat{\mathscr{U}} = (\mathscr{X} \leftarrow \mathscr{U} \rightarrow \mathscr{Y}) \quad \text{and} \quad \hat{\mathscr{V}} = (\mathscr{Y} \leftarrow \mathscr{V} \rightarrow \mathscr{Z})$$

be two optimal minimal two-fans.

Recall that each two-fan of $\mathbf{G}$-diagrams is a $\mathbf{G}$-diagram of two-fans between the individual spaces, that is

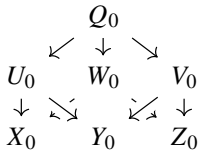
$$\mathscr{U} = \{\mathscr{U}_i = (X_i \leftarrow U_i \rightarrow Y_i)\}$$
$$\mathscr{V} = \{\mathscr{V}_i = (Y_i \leftarrow V_i \rightarrow Z_i)\}.$$

We construct a coupling $\hat{\mathscr{W}}$ between $\mathscr{X}$ and $\mathscr{Z}$ in the following manner. Starting with the two-tents diagram between the initial spaces, we use adhesion to extend it to a full diagram, thus constructing a coupling between X_0 and Z_0. This full diagram could then be "pushed down" and provides full extensions of two-tents on all lower levels. Thus we could "compose" couplings $\hat{\mathscr{U}}$ and $\hat{\mathscr{V}}$ and use a Shannon inequality to establish the triangle inequality for the intrinsic entropy distance. Details are as follows.

Consider a two-tents diagram

$$X_0 \leftarrow U_0 \rightarrow Y_0 \leftarrow V_0 \rightarrow Z_0$$

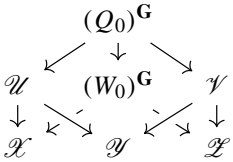
and extend it by adhesion, as described above, to a $\boldsymbol{\Lambda}_3$-diagram



Together with the reductions

$$(X_0)^{\mathbf{G}} \rightarrow \mathscr{X}, \qquad (Y_0)^{\mathbf{G}} \rightarrow \mathscr{Y}, \qquad (Z_0)^{\mathbf{G}} \rightarrow \mathscr{Z}$$
$$(U_0)^{\mathbf{G}} \rightarrow \mathscr{U}, \qquad (V_0)^{\mathbf{G}} \rightarrow \mathscr{V}$$

explained in Sect. 2.6, it gives rise to a $\boldsymbol{\Lambda}_3$-diagram of $\mathbf{G}$-diagrams



(23)

The diagram above is not necessarily minimal and we now consider the "minimization" of the $\boldsymbol{\Lambda}_3$-diagram (23), as provided by Lemma 2.2

$$\hat{\mathscr{Q}} = \left(\begin{array}{ccccc} & & \mathscr{Q} & & \\ & \swarrow & \downarrow & \searrow & \\ \mathscr{U} & & \mathscr{W} & & \mathscr{V} \\ \downarrow & \searrow & & \swarrow & \downarrow \\ \mathscr{X} & & \mathscr{Y} & & \mathscr{Z} \end{array} \right)$$

Applying Lemma 7.2 to each

$$\hat{\mathscr{Q}}i = \begin{pmatrix} & & Q_i & & \\ & \swarrow & \downarrow & \searrow & \\ U_i & & W_i & & V_i \\ \downarrow & \searrow\!\!\swarrow & & \swarrow\!\!\searrow & \downarrow \\ X_i & & Y_i & & Z_i \end{pmatrix}$$

which is minimal by Lemma 2.2(1), we obtain the required inequality, concluding the proof of the triangle inequality.

Finally, if $k(\mathscr{X}, \mathscr{Y}) = 0$, then there is a two-fan $\mathscr{F}$ of **G**-diagrams between $\mathscr{X}$ and $\mathscr{Y}$ with $\mathrm{kd}(\mathscr{F}) = 0$, from which it follows that $\mathscr{X}$ and $\mathscr{Y}$ are isomorphic. □

Proposition 3.3 *Let* **G** *be a complete poset category. Then with respect to the Kolmogorov distance on* $\mathbf{Prob}\langle\mathbf{G}\rangle$ *the tensor product*

$$\otimes\colon (\mathbf{Prob}\langle\mathbf{G}\rangle, \boldsymbol{k})^2 \to (\mathbf{Prob}\langle\mathbf{G}\rangle, \boldsymbol{k})$$

is 1-Lipschitz in each variable, that is, for every triple $\mathscr{X}, \mathscr{Y}, \mathscr{Y}' \in \mathbf{Prob}\langle\mathbf{G}\rangle$ *the following bound holds*

$$\boldsymbol{k}(\mathscr{X} \otimes \mathscr{Y}, \mathscr{X} \otimes \mathscr{Y}') \leq \boldsymbol{k}(\mathscr{Y}, \mathscr{Y}').$$

Proof The claim follows easily from the additivity of entropy in equation (6). Suppose that $\mathscr{X} = \{X_i; f_{ij}\}$, $\mathscr{Y} = \{Y_i; g_{ij}\}$ and $\mathscr{Y}' = \{Y_i'; g_{ij}'\}$ are three **G**-diagrams and

$$\mathscr{F} = (\mathscr{Y} \leftarrow \mathscr{Z} \rightarrow \mathscr{Y}')$$

is an optimal fan, so that

$$\mathbf{k}(\mathscr{Y}, \mathscr{Y}') = \sum_i \left[2\mathsf{Ent}(Z_i) - \mathsf{Ent}(Y_i) - \mathsf{Ent}(Y_i')\right].$$

Consider the fan

$$\mathscr{G} = (\mathscr{X} \otimes \mathscr{Y} \leftarrow \mathscr{X} \otimes \mathscr{Z} \rightarrow \mathscr{X} \otimes \mathscr{Y}').$$

Then, by additivity of entropy, in equation (6), we have

$$\begin{aligned} \mathrm{kd}(\mathscr{F}) &= \sum_i \left[2\mathsf{Ent}(X_i \otimes Z_i) - \mathsf{Ent}(X_i \otimes Y_i) - \mathsf{Ent}(X_i \otimes Y_i')\right] \\ &= \sum_i \left[2\mathsf{Ent}(Z_i) - \mathsf{Ent}(Y_i) - \mathsf{Ent}(Y_i')\right] \\ &= \mathrm{kd}(\mathscr{G}) \end{aligned}$$

and, therefore,

$$\mathbf{k}(\mathscr{X}\otimes\mathscr{Y},\,\mathscr{X}\otimes\mathscr{Y}')\le \mathrm{kd}(\mathscr{G})=\mathrm{kd}(\mathscr{F})=\mathbf{k}(\mathscr{Y},\mathscr{Y}').$$

Thus, the tensor product of probability spaces is 1-Lipschitz with respect to each argument. □

Proposition 3.7 *For any triple of diagrams $\mathscr{X},\mathscr{Y},\mathscr{U}$ holds*

$$\boldsymbol{\kappa}(\mathscr{X}\otimes\mathscr{U},\mathscr{Y}\otimes\mathscr{U})=\boldsymbol{\kappa}(\mathscr{X},\mathscr{Y}).$$

Proof Define a translation-invariant bivariate function $\boldsymbol{\kappa}'$ on $\mathbf{Prob}\langle\mathbf{G}\rangle$ by setting for a pair $\mathscr{A},\mathscr{B}\in\mathbf{Prob}\langle\mathbf{G}\rangle$

$$\boldsymbol{\kappa}'(\mathscr{A},\mathscr{B}):=\inf_{\mathscr{C}\in\mathbf{Prob}\langle\mathbf{G}\rangle}\boldsymbol{\kappa}(\mathscr{A}\otimes\mathscr{C},\mathscr{B}\otimes\mathscr{C}).$$

Clearly, function $\boldsymbol{\kappa}'$ is translation invariant and satisfies $\boldsymbol{\kappa}'\le\boldsymbol{\kappa}$. Our task now is to show that $\boldsymbol{\kappa}'\ge\boldsymbol{\kappa}$.

Fix $\mathscr{X},\mathscr{Y}\in\mathbf{Prob}\langle\mathbf{G}\rangle$. Choose some $\varepsilon>0$ and take $\mathscr{Z}\in\mathbf{Prob}\langle\mathbf{G}\rangle$ such that

$$\boldsymbol{\kappa}(\mathscr{X}\otimes\mathscr{Z},\mathscr{Y}\otimes\mathscr{Z})\le\boldsymbol{\kappa}'(\mathscr{X},\mathscr{Y})+\varepsilon.$$

Since translations are non-expanding, it holds also for any $\mathscr{U}\in\mathbf{Prob}\langle\mathbf{G}\rangle$

$$\boldsymbol{\kappa}(\mathscr{X}\otimes\mathscr{Z}\otimes\mathscr{U},\mathscr{Y}\otimes\mathscr{Z}\otimes\mathscr{U})\le\boldsymbol{\kappa}'(\mathscr{X},\mathscr{Y})+\varepsilon.$$

Then for any $n\in\mathbb{N}$ we can estimate

$$\begin{aligned}\boldsymbol{\kappa}(\mathscr{X},\mathscr{Y})&=\frac{1}{n}\boldsymbol{\kappa}(\mathscr{X}^n,\mathscr{Y}^n)\\&\le\frac{1}{n}\big[2\boldsymbol{\kappa}(\{\bullet\}^{\mathbf{G}},\mathscr{Z})+\boldsymbol{\kappa}(\mathscr{X}^n\otimes\mathscr{Z},\mathscr{Y}^n\otimes\mathscr{Z})\big].\end{aligned}$$

For $i=0,\dots,n$ we set $\mathscr{T}_i=\mathscr{X}^{n-i}\otimes\mathscr{Y}^i\otimes\mathscr{Z}$. Then we have $\mathscr{T}_0=\mathscr{X}^n\otimes\mathscr{Z}$ and $\mathscr{T}_n=\mathscr{Y}^n\otimes\mathscr{Z}$. Also for each $i=0,\dots,n-1$ the pair $(\mathscr{T}_i,\mathscr{T}_{i+1})$ is a translation of the pair $(\mathscr{X}\otimes\mathscr{Z},\mathscr{Y}\otimes\mathscr{Z})$ by $\mathscr{X}^{n-i-1}\otimes\mathscr{Y}^i$, therefore

$$\boldsymbol{\kappa}(\mathscr{T}_i,\mathscr{T}_{i+1})\le\boldsymbol{\kappa}(\mathscr{X}\otimes\mathscr{Z},\mathscr{Y}\otimes\mathscr{Z})\le\boldsymbol{\kappa}'(\mathscr{X},\mathscr{Y})+\varepsilon.$$

Now we continue the estimate

$$\begin{aligned}\boldsymbol{\kappa}(\mathscr{X},\mathscr{Y})&\le\frac{1}{n}\big[2\boldsymbol{\kappa}(\{\bullet\}^{\mathbf{G}},\mathscr{Z})+\boldsymbol{\kappa}(\mathscr{X}^n\otimes\mathscr{Z},\mathscr{Y}^n\otimes\mathscr{Z})\big]\\&\le\frac{1}{n}\Big[2\boldsymbol{\kappa}(\{\bullet\}^{\mathbf{G}},\mathscr{Z})+\sum_{i=0}^{n-1}\boldsymbol{\kappa}(\mathscr{T}_i,\mathscr{T}_{i+1})\Big]\end{aligned}$$

$$\leq \frac{2}{n}\boldsymbol{\kappa}(\{\bullet\}^{\mathbf{G}}, \mathscr{Z}) + \boldsymbol{\kappa}'(\mathscr{X}, \mathscr{Y}) + \varepsilon.$$

Since n is arbitrarily large, the first summand in the right-hand-side could be dropped and then by choosing $\varepsilon > 0$ arbitrarily small we obtain the required inequality. □

Proposition 3.8 *Suppose* $\mathbf{G}$ *is a complete poset category and* $\boldsymbol{\delta} = \boldsymbol{k}, \boldsymbol{\kappa}$ *is either intrinsic entropy distance or asymptotic entropy distance on* $\mathbf{Prob}\langle\mathbf{G}\rangle$. *Then the entropy function*

$$Ent_*\colon (\mathbf{Prob}\langle\mathbf{G}\rangle, \boldsymbol{\delta}) \to (\mathbb{R}^{[\![\mathbf{G}]\!]}, \ |\cdot|_1), \quad \mathscr{X} = \{X_i, f_{ij}\} \mapsto (EntX_i)_i \in \mathbb{R}^{[\![\mathbf{G}]\!]}$$

is 1*-Lipschitz.*

Proof Let $\mathscr{X}, \mathscr{Y} \in \mathbf{Prob}\langle\mathbf{G}\rangle$ and let

$$\mathscr{G} = (\mathscr{X} \leftarrow \mathscr{Z} \rightarrow \mathscr{Y})$$

be an optimal fan with components

$$\mathscr{G}_i = (X_i \leftarrow Z_i \rightarrow Y_i).$$

For a fixed index i we can estimate the difference of entropies

$$\mathsf{Ent}(X_i) - \mathsf{Ent}(Y_i) = 2\big(\mathsf{Ent}(X_i) - \mathsf{Ent}(Z_i)\big) + \mathrm{kd}(\mathscr{G}_i) \leq \mathrm{kd}(\mathscr{G}_i).$$

By symmetry we then have

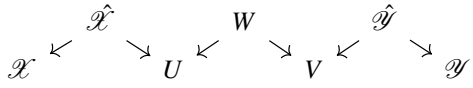
$$|\mathsf{Ent}(X_i) - \mathsf{Ent}(Y_i)| \leq \mathrm{kd}(\mathscr{G}_i).$$

Adding above inequalities for all i we have

$$|\mathsf{Ent}_*(\mathscr{X}) - \mathsf{Ent}_*(\mathscr{Y})|_1 \leq \mathrm{kd}(\mathscr{G}) = \mathbf{k}(\mathscr{X}, \mathscr{Y}).$$

By the additivity of entropy we also obtain the 1-Lipschitz property of the entropy function with respect to the asymptotic entropy distance $\boldsymbol{\kappa}$. □

Proposition 3.9 (Slicing Lemma) *Suppose* $\mathbf{G}$ *is a complete poset category and we are given* $\mathscr{X}, \hat{\mathscr{X}}, \mathscr{Y}, \hat{\mathscr{Y}} \in \hat{\mathbf{Prob}}\langle\mathbf{G}\rangle$*—four* $\mathbf{G}$*-diagrams and* $U, V, W \in \mathbf{Prob}$*—three probability spaces, that are included into the following three-tents diagram*



such that the two-fan $(U \leftarrow W \rightarrow V)$ *is minimal. Then the following estimate holds*

$$\boldsymbol{k}(\mathscr{X}, \mathscr{Y}) \leq \int_W \boldsymbol{k}(\mathscr{X}\lfloor u, \mathscr{Y}\lfloor v) d\, p_W(u, v)$$

$$+[\![\mathbf{G}]\!] \cdot kd(U \leftarrow W \rightarrow V)$$
$$+ \sum_i \left[Ent(U\lfloor X_i) + Ent(V\lfloor Y_i) \right].$$

Proof Since the two-fan $(U \leftarrow W \rightarrow V)$ is minimal the probability space W could be considered having underlying set to be a subset of the Cartesian product of the underlying sets of U and V. For any pair $(u, v) \in \underline{W}$ with a positive weight consider an optimal two-fan

$$\mathscr{G}_{uv} = \left(\mathscr{X}\lfloor u \xleftarrow{\pi_{\mathscr{X}}} \mathscr{Z}_{uv} \xrightarrow{\pi_Y} \mathscr{Y}\lfloor v \right) \tag{24}$$

where $\mathscr{Z}_{uv} = \left\{ Z_{uv,i}; \rho_{ij} \right\}$. Let $p_{uv,i}$ be the probability distributions on $Z_{uv,i}$—the individual spaces in the diagram $\mathscr{Z}_{uv}$. The next step is to take a convex combination of distributions $p_{uv,i}$ weighted by p_W to construct a coupling $\mathscr{X} \leftarrow \mathscr{Z} \rightarrow \mathscr{Y}$.

First we extend the 7-vertex diagram to a full $\boldsymbol{\Lambda}_4$-diagram of $\mathbf{G}$-diagrams, such that the top vertex has the distribution

$$p_i(x, y, u, v) := p_{uv,i}(x, y) p_W(u, v)$$

as described in the Sect. 2.5.7.

If we integrate over y, we obtain

$$\int_{y \in Y} \mathrm{d}\, p_i(x, u, v, y) = ((\pi_{\mathscr{X},i})_* p_{uv,i})(x) p_W(u, v).$$

Equation (24) implies that $(\pi_{\mathscr{X},i})_* p_{uv,i} = p_{X_i}(\cdot \lfloor u)$ and therefore

$$\int_{y \in Y} \mathrm{d}\, p_i(x, y, u, v) = p_{X_i}(x \lfloor u) p_W(u, v).$$

In the same way,

$$\int_{x \in X} \mathrm{d}\, p_i(x, y, u, v) = p_{Y_i}(y \lfloor v) p_W(u, v).$$

It follows that

$$\mathscr{X}\lfloor uv = \mathscr{X}\lfloor u \quad \text{and} \quad \mathscr{Y}\lfloor uv = \mathscr{Y}\lfloor v \tag{25}$$

and

$$\mathsf{Ent}(X_i \lfloor UV) = \mathsf{Ent}(X_i \lfloor U) \quad \text{and} \quad \mathsf{Ent}(Y_i \lfloor UV) = \mathsf{Ent}(Y_i \lfloor V). \tag{26}$$

The extended diagram contains a two-fan of diagrams $\mathscr{F} = (\mathscr{X} \leftarrow \mathscr{Z} \rightarrow \mathscr{Y})$ with sink vertices $\mathscr{X}$ and $\mathscr{Y}$. We call its initial vertex $\mathscr{Z} = \left\{ XY_i, f_{ij} \right\}$.

The following estimates conclude the proof the Slicing Lemma. First we use the definitions of intrinsic entropy distance $\mathbf{k}$ and of $\mathrm{kd}(\mathscr{F})$ to estimate

$$\begin{aligned}\mathbf{k}(\mathscr{X},\mathscr{Y}) &\le \mathrm{kd}(\mathscr{F})\\ &= \sum_i \mathrm{kd}(\mathscr{F}_i)\\ &= \sum_i \big[2\mathsf{Ent}(XY_i) - \mathsf{Ent}(X_i) - \mathsf{Ent}(Y_i)\big].\end{aligned}$$

Next, we apply the definition of the conditional entropy to rewrite the right-hand side

$$\begin{aligned}\mathbf{k}(\mathscr{X},\mathscr{Y}) \le \sum_i \big[&2\mathsf{Ent}(XY_i\lfloor UV) + 2\mathsf{Ent}(UV) - 2\mathsf{Ent}(UV\lfloor XY_i)\\ &- \mathsf{Ent}(X_i\lfloor U) - \mathsf{Ent}(U) + \mathsf{Ent}(U\lfloor X_i)\\ &- \mathsf{Ent}(Y_i\lfloor V) - \mathsf{Ent}(V) + \mathsf{Ent}(V\lfloor Y_i)\big].\end{aligned}$$

We now use (26) and rearrange terms to obtain

$$\begin{aligned}\mathbf{k}(\mathscr{X},\mathscr{Y}) \le \sum_i \big[&2\mathsf{Ent}(XY_i\lfloor UV) - \mathsf{Ent}(X_i\lfloor UV) - \mathsf{Ent}(Y_i\lfloor UV)\\ &+ 2\mathsf{Ent}(UV) - \mathsf{Ent}(U) - \mathsf{Ent}(V)\\ &- 2\mathsf{Ent}(UV\lfloor XY_i) + \mathsf{Ent}(U\lfloor X_i) + \mathsf{Ent}(V\lfloor Y_i)\big].\end{aligned}$$

By the integral formula for conditional entropy (7) applied to the first three terms we get

$$\begin{aligned}&\sum_i \big[2\mathsf{Ent}(XY_i\lfloor UV) - \mathsf{Ent}(X_i\lfloor UV) - \mathsf{Ent}(Y_i\lfloor UV)\big]\\ &\quad= \int_{UV} \mathbf{k}(\mathscr{X}\lfloor uv, \mathscr{Y}\lfloor uv)\mathrm{d}\, p_W(u,v)\end{aligned}$$

However, because of (25) this simplifies to

$$\int_{UV} \mathbf{k}(\mathscr{X}\lfloor uv, \mathscr{Y}\lfloor uv)\mathrm{d}\, p_W(u,v) = \int_{UV} \mathbf{k}(\mathscr{X}\lfloor u, \mathscr{Y}\lfloor v)\mathrm{d}\, p_W(u,v).$$

Therefore,

$$\begin{aligned}\mathbf{k}(\mathscr{X},\mathscr{Y}) \le \int_{UV} &\mathbf{k}(\mathscr{X}\lfloor u, \mathscr{Y}\lfloor v)\mathrm{d}\, p_W(u,v) + [\![\mathbf{G}]\!]\cdot \mathrm{kd}(U \leftarrow W \rightarrow V)\\ &+ \sum_i \big[\mathsf{Ent}(U\lfloor X_i) + \mathsf{Ent}(V\lfloor Y_i)\big].\end{aligned}$$

□

Acknowledgements Open access funding provided by Max Planck Society. We would like to thank Tobias Fritz, František Matúš, Misha Movshev and Johannes Rauh for inspiring discussions. We are grateful to the participants of the Wednesday Morning Session at the CASA group at the Eindhoven University of Technology for valuable feedback on the introduction of the article. Finally, we thank the Max Planck Institute for Mathematics in the Sciences, Leipzig, for its hospitality.

References

1. Ay, N., Bertschinger, N., Der, R., Güttler, F., Olbrich, E.: Predictive information and explorative behavior of autonomous robots. Eur. Phys. J. B **63**(3), 329–339 (2008)
2. Baez, J.C., Fritz, T., Leinster, T.: A characterization of entropy in terms of information loss. Entropy **13**(11), 1945–1957 (2011)
3. Boltzmann, L.: über die mechanische Bedeutung des zweiten Hauptsatzes der Wärmetheorie. Wiener Berichte **53**, 195–220 (1866)
4. Boltzmann, L.: Vorlesungen über Gastheorie, vols. I, II. J.A. Barth, Leipzig (1896)
5. Bertschinger, N., Rauh, J., Olbrich, E., Jost, J., Ay, N.: Quantifying unique information. Entropy **16**(4), 2161–2183 (2014)
6. Cicalese, F., Gargano, L., Vaccaro, U.: How to find a joint probability distribution of minimum entropy (almost) given the marginals. In: 2017 IEEE International Symposium on Information Theory (ISIT), pp. 2173–2177. IEEE, New York (2017)
7. Csiszár, I.: The method of types. IEEE Trans. Inf. Theory, 44(6):2505–2523 (1998) **(Information theory: 1948–1998)**
8. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley Series in Telecommunications. A Wiley-Interscience Publication, Wiley, New York (1991)
9. Friston, K.: The free-energy principle: a rough guide to the brain? Trends Cogn. Sci. **13**(7), 293–301 (2009)
10. Gromov, M.: In a search for a structure, part 1: on entropy (2012). Preprint. https://www.ihes.fr/~gromov/wp-content/uploads/2018/08/structre-serch-entropy-july5-2012.pdf and https://math.mit.edu/~dspivak/teaching/sp13/gromov--EntropyViaCT.pdf. Accessed 08 Oct 2018
11. Kocaoglu, M., Dimakis, A.G., Vishwanath, S., Hassibi, B.: Entropic causality and greedy minimum entropy coupling (2017). arXiv preprint. arXiv:1701.08254
12. Kolmogorov, A.N.: New metric invariant of transitive dynamical systems and endomorphisms of Lebesgue spaces. Dokl. Russ. Acad. Sci. **119**(5), 861–864 (1958)
13. Kolmogorov, A.N.: New metric invariant of transitive dynamical systems and endomorphisms of Lebesgue spaces. Dokl. Russ. Acad. Sci. **124**, 754–755 (1959)
14. Kovacevic, M., Stanojevic, I., Senk, V.: On the hardness of entropy minimization and related problems. In: Information Theory Workshop (ITW). 2012 IEEE, pp. 512–516. IEEE, New York (2012)
15. MacLane, S.: Categories for the Working Mathematician. Graduate Texts in Mathematics, vol. 5. Springer, New York (1971)
16. Matus, F.: Infinitely many information inequalities. In: IEEE International Symposium on Information Theory, 2007. ISIT 2007, pp. 41–44. IEEE, New York (2007)
17. Ornstein, D.: Bernoulli shifts with the same entropy are isomorphic. Adv. Math. **4**, 337–352 (1970)
18. Steudel, B., Ay, N.: Information-theoretic inference of common ancestors. Entropy **17**(4), 2304–2327 (2015)
19. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948)
20. Shannon, C.: The lattice theory of information. Trans. IRE Prof. Group Inf. Theory **1**(1), 105–107 (1953)
21. Sinai, Ya G.: On the notion of entropy of a dynamical system. Dokl. Russ. Acad. Sci. **124**, 768–771 (1959)

22. Sinai, Y.G.: Introduction to Ergodic Theory. Princeton University Press, Princeton (1976) **(Translated by V. Scheffer Mathematical Notes, vol. 18)**
23. Van Dijk, S.G., Polani, D.: Informational constraints-driven organization in goal-directed behavior. Adv. Complex Syst. **16**(02n03), 1350016 (2013)
24. Vidyasagar, M.: A metric between probability distributions on finite sets of different cardinalities and applications to order reduction. IEEE Trans. Autom. Control **57**(10), 2464–2477 (2012)
25. Yeung, R.W.: A First Course in Information Theory. Springer Science & Business Media, Berlin (2012)