Compositionality
compositionality.episciences.org
ISSN 2631-4444
Volume 8, Issue 1 (2026)

# Abstract Markov Random Fields

Leon Lang[1], Clélia de Mulatier[1,2], Rick Quax[1], and Patrick Forré[3]

[1]Informatics Institute, University of Amsterdam
[2]Institute for Theoretical Physics, University of Amsterdam
[3]Korteweg-de Vries Institute for Mathematics, University of Amsterdam



Markov random fields are known to be fully characterized by properties of their information diagrams, or $I$-diagrams. In particular, for Markov random fields, regions in the $I$-diagram corresponding to disconnected vertex sets in the graph vanish. Recently, $I$-diagrams have been generalized to $F$-diagrams, for a larger class of functions $F$ satisfying the chain rule beyond Shannon entropy, such as Kullback-Leibler divergence and cross-entropy. In this work, we generalize the notion and characterization of Markov random fields to this larger class of functions $F$ and investigate preliminary applications.

We define $F$-independences, $F$-mutual independences, and $F$-Markov random fields and characterize them by their $F$-diagram. In the process, we also define $F$-dual total correlation and prove that its vanishing is equivalent to $F$-mutual independence. We then apply our results to information functions $F$ that are applied to probability mass functions. We show that if the probability distributions of a set of random variables are Markov random fields for the same graph, then we formally recover the notion of an $F$-Markov random field for that graph. We then study the Kullback-Leibler diagrams on specific Markov chains, leading to a visual representation of the second law of thermodynamics and a simple explicit derivation of the decomposition of the evidence lower bound for diffusion models.

## Contents



Leon Lang: l.lang@uva.nl, 0000-0002-1950-2831. Main contributing author.
Clélia de Mulatier: c.m.c.demulatier@uva.nl, 0000-0003-3578-5453
Rick Quax: r.quax@uva.nl, 0000-0002-0299-0074
Patrick Forré: p.d.forre@uva.nl, 0000-0003-4663-3842

Digital object identifier: 10.46298/compositionality-8-1. Received: 2024-10-23. Accepted: 2025-12-20.

# 1 Introduction

Entropy, mutual information, and higher-order information terms between several random variables can be visualized in information diagrams, also known as $I$-diagrams. This is known as Hu's Theorem [Hu, 1962, Yeung, 1991]. These diagrams become especially interesting when the visualized variables obey conditional independences, which then implies that some regions in these diagrams vanish. A series of papers [Kawabata and Yeung, 1992, Yeung et al., 2002, 2019] exploited this idea to study Markov random fields, which according to the global Markov property satisfy a set of conditional independences determined by an underlying graph [Hammersley and Clifford, 1971, Preston, 1976, Spitzer, 1971]. This then implies that all intersections in the $I$-diagram corresponding to *disconnected vertex sets* in the graph vanish, a result we visualize for simple graphs in Figure 3.

Recently, Hu's theorem has been generalized from entropy $I$ to more general functions $F$ on commutative, idempotent monoids satisfying a chain rule [Lang et al., 2025]. Examples for $F$ are Kullback-Leibler divergence, cross-entropy, Tsallis entropy, and even Kolmogorov complexity. The resulting $F$-diagrams show structurally the same relations as $I$-diagrams, and thus allow higher-order $F$-terms to be visualized and reasoned about in a unified way — see Figure 2.

Functions such as the cross-entropy or Kullback-Leibler divergence are important in the context of statistical modeling of multivariate data, in which one aims to find a probabilistic model able to reproduce the information structure of the data. For instance, the $F$-diagram for cross-entropy allows us to visualize how the cross-entropy between a model probability distribution and the data distribution is decomposed into higher-order terms. Cocco and Monasson [2012] used these higher-order terms (which they called cluster (cross)-entropies) in their adaptive cluster expansion approach to statistical modeling of data with Ising models. Kullback-Leibler divergence has been studied in the context of decompositions of joint entropy and information [Amari, 2001] and is ubiquitous in machine learning. In other contexts, for example Kolmogorov complexity, the precise meaning of the higher-order terms is not yet clear.

In this work, we take the generalization of $I$-diagrams to $F$-diagrams as a motivation to generalize the results from Kawabata and Yeung [1992], Yeung et al. [2002, 2019]. We define $F$-independences by vanishing $F$-terms of degree two, similar to how the probabilistic independence of random variables is characterized by vanishing mutual information. We then define $F$-mutual independences and $F$-dual total correlation. We show that an $F$-mutual independence is characterized by a vanishing $F$-dual total correlation. We then define $F$-Markov random fields by the global Markov property and fully characterize them in terms of the $F$-diagram: The global Markov property holds if and only if regions corresponding to disconnected vertex sets in the graph vanish.

We then apply this theory to the case where $F$ is a function like entropy, Kullback-Leibler divergence, or cross-entropy that is applied to probability mass functions. We show that when applying $F$ to sets of probability distributions that form a Markov random field with respect to the same graph, then the underlying random variables form an $F$-Markov random field. In particular, when applying our $F$-diagram characterization of $F$-Markov random fields, this implies that regions in the $F$-diagram corresponding to disconnected vertex sets disappear. We then apply this result to the case of two joint distributions that form a Markov chain with equal transition probabilities from one time-step to the next, and the specific case that $F$ is Kullback-Leibler divergence. This leads to a degeneracy of the Kullback-Leibler diagram in which the Kullback-Leibler divergence progressively shrinks "over time" — a diagrammatic visualization of a weak version of the second law of thermodynamics. Finally, we study the loss function of diffusion models — the evidence lower bound — and show how its explicit decomposition can be derived using a Kullback-Leibler diagram over a Markov chain.

### Notation

For $i, k \in \mathbb{N}$, we set $[i:k] := \{i, i+1, \dots, k\}$ if $i \leq k$ and $[i:k] = \emptyset$, else. As a special case, we set $[k] := [1:k]$. If $I$ is a set and $i \in I$, we often write $I \setminus i$ for $I \setminus \{i\}$. For a set $\Sigma$, we denote by $2^\Sigma$ its power set, i.e., the set of its subsets. If $W_i$ are sets indexed with $i \in I$, then $W_I$ denotes $\bigcup_{i \in I} W_i$. If $X_i, i \in I$ are elements of a commutative monoid and $A \subseteq I$, then we set $X_A := \prod_{a \in A} X_a$. If $X_1, \dots, X_n$ are elements of a separoid that form a Markov chain, then we will write $X_1 \to X_2 \to \cdots \to X_n$. We will denote the trivial measurable space by $*$, which contains precisely one element denoted $* \in *$. The Shannon entropy of a random variable $X$ is denoted by $I(X)$ or $I_1(X)$, deviating from the typical notation of $H(X)$; The aim is to emphasize more strongly how entropy is embedded in the collection of (higher-order) Shannon information functions like mutual information and interaction information.

## 2 Background and Outline

In this section, we introduce important background in multivariate information theory and its abstract generalizations, precisely state Yeung's characterization of Markov random fields in terms of $I$-diagrams and our generalizations of those results, and outline the rest of the paper. The aim is for this section to be self-contained and to provide sufficient context to appreciate the general results that then follow.

In Section 2.1, we introduce Shannon entropy, mutual information, and interaction information in the abstract setting from Lang et al. [2025], highlighting the structure of monoids acting on abelian groups. In Section 2.2, we then state the generalized Hu theorem from Lang et al. [2025] and explain how it gives rise to the well-known $I$-diagrams from Yeung [1991] when specializing to the case of Shannon entropy. In Section 2.3 we introduce some graph terminology necessary in the theory of Markov random fields. In Section 2.4, we introduce Markov random fields in separoids and show Yeung's characterization of those in the probabilistic context in terms of the $I$-diagram [Yeung et al., 2002]. In Section 2.5, we motivate and state our generalization of this characterization to $F$-diagrams, making use of the generalized Hu theorem from Lang et al. [2025], and then outline the rest of the work, which will contain the proofs and applications of this result, in Section 2.6.

## 2.1 Entropy, Mutual Information, and Interaction Information

In this section, we introduce the well-known notions of entropy, mutual information, and interaction information of higher degrees in the precise formal framework from Lang et al. [2025] that reveals the monoid structure that we make use of in our work. Let $\Omega$ be a countable discrete measurable space. We do *not* fix a probability mass function on $\Omega$, which is useful for obtaining a monoid action later in Definition 2.3. When we speak of *random variables*, then we mean functions $X : \Omega \to E_X$ with a *finite and discrete* value space $E_X$.

We write the space of probability mass functions $P : \Omega \to [0, 1]$ as $\Delta(\Omega)$. We equip $\Delta(\Omega)$ with the smallest $\sigma$-algebra that makes all evaluation maps

$$\mathrm{ev}_x : \Delta(\Omega) \to \mathbb{R}, \quad P \mapsto \mathrm{ev}_x(P) \coloneqq P(x)$$

for all elements $x \in \Omega$ measurable. In the finite case, this equals the Borel $\sigma$-algebra under the embedding $\Delta(\Omega) \subseteq \mathbb{R}^\Omega \cong \mathbb{R}^{|\Omega|}$. This measurable structure is standard in the context of the Giry monad [Giry, 1982]. Measurability is not strictly necessary in our discrete setting, but would become important if one were to generalize our results to a non-discrete domain.

For a probability mass function $P \in \Delta(\Omega)$, we write the distributional law with respect to a random variable $X : \Omega \to E_X$ as

$$P_X \in \Delta(E_X), \quad P_X(x) \coloneqq P\big(X^{-1}(x)\big) = \sum_{\omega \in X^{-1}(x)} P(\omega).$$

This is also a probability mass function. Furthermore, for $x \in E_X$ with $P_X(x) \neq 0$, we define the conditional probability mass function $P|_{X=x} \in \Delta(\Omega)$ by

$$P|_{X=x} \in \Delta(\Omega), \quad P|_{X=x}(\omega) \coloneqq \frac{P\big(\{\omega\} \cap X^{-1}(x)\big)}{P_X(x)}.$$

If $X : \Omega \to E_X$ and $Y : \Omega \to E_Y$ are two random variables, then their joint variable is given by

$$XY : \Omega \to E_X \times E_Y, \quad \omega \mapsto \big(X(\omega), Y(\omega)\big).$$

If the random variable is clear from the context, we write $P(x)$ for $P_X(x)$. Similarly, we may write $P(x \mid y)$ for $\big(P|_{Y=y}\big)_X(x)$ and $P(x, y)$ for $P_{XY}(x, y)$.

We want to impose the structure of a monoid on collections of random variables. Recall that a commutative, idempotent monoid $M = (M, \cdot, \mathbf{1})$ consists of a set $M$ together with a multiplication rule $\cdot : M \times M \to M$ that is associative, commutative, has $\mathbf{1}$ as its neutral element, and is idempotent: $X \cdot X = X$ for all $X \in M$. For $X, Y \in M$, write $X \precsim Y$ if $X \cdot Y = Y$. With this definition, $(M, \precsim)$ becomes a *join-semilattice*, which is an equivalent description of a commutative, idempotent monoid. Intuitively, it is often useful to think of $\precsim$ as the inclusion of sets, and of the product of elements in $M$ as a union, which will be made precise in Theorem 2.8. Join-semilattices were used in the development of the theory of separoids in Dawid [2001], but we will not make use of that viewpoint and will throughout use the structure of monoids. We also note that we do *not* have a meet operator in our framework, neither in the context of random variables, nor in the generalization to monoids.

To impose the structure of a monoid on collections of random variables, we need to identify *equivalent* random variables. For two random variables $X, Y$ on $\Omega$, we define $X \precsim Y$ if $X$ is a deterministic function of $Y$, meaning there exists a function $f_{XY} : E_Y \to E_X$ such that $X = f_{XY} \circ Y$, i.e., $X(\omega) = f_{XY}(Y(\omega))$ for all $\omega \in \Omega$. We write $X \sim Y$ if $X \precsim Y$ and $Y \precsim X$, which is an equivalence relation.

In the following, we will then write $X$ for both the random variable and its equivalence class, and we denote by $*$ the trivial measurable space that contains precisely one element $* \in *$. One obtains the following:

**Proposition 2.1** (The Monoid of Random Variables)**.** *Equivalence classes of random variables, together with the multiplication given by the join operation $X \cdot Y \coloneqq XY$ and the neutral element given by $\mathbf{1} : \Omega \to *$, form a commutative, idempotent monoid.*

*Proof.* For a proof in this precise framework, see Lang et al. [2025, Section 2.3]. In the framework of lattices, this was formulated by [Dawid, 2001]. We note that the collection of equivalence classes of random variables on $\Omega$ is indeed a *set* instead of a proper class, as an equivalence class $[X]$ can be identified with the partition $\left\{X^{-1}(x) \mid x \in E_X\right\}$ on $\Omega$. □

Clearly, any subset of equivalence classes of random variables — if it is closed under multiplication and contains a constant random variable — then also forms a commutative, idempotent monoid. It is useful to work with monoids of random variables since they have useful structure, and since the information functions we are concerned with do not depend on the representative of an equivalence class. Recall that an abelian group $G = (G, +, 0)$ consists of a set $G$ together with an addition rule $+ : G \times G \to G$ that is associative and commutative, has 0 as its neutral element, and has "negative" elements: $g + (-g) = (-g) + g = 0$. Now, define $\mathrm{Meas}(\Delta(\Omega), \mathbb{R})$ as the abelian group of measurable functions from $\Delta(\Omega)$ to $\mathbb{R}$, where we define $(F + F')(P) \coloneqq F(P) + F'(P)$.

**Definition 2.2** (Shannon Entropy)**.** *Let* $\ln : (0, \infty) \to \mathbb{R}$ *be the natural logarithm and* $X : \Omega \to E_X$ *be a random variable. The* Shannon entropy of $X$ with respect to $P \in \Delta(\Omega)$ *is given by*

$$I(X; P) \coloneqq I(P_X) = -\sum_{x \in E_X} P_X(x) \ln P_X(x) \in \mathbb{R}.$$

*The* entropy function *or* Shannon entropy *of* $X$ *is the measurable function*

$$I(X) \in \mathrm{Meas}(\Delta(\Omega), \mathbb{R}), \quad \big[I(X)\big](P) \coloneqq I(X; P)$$

*defined on probability mass functions. This function does not depend on the representative of the equivalence class* $X$*.*

**Definition 2.3** (Averaged Conditioning)**.** *Let* $F \in \mathrm{Meas}(\Delta(\Omega), \mathbb{R})$*. For a random variable* $X : \Omega \to E_X$*, define the* averaged conditioning *of* $F$ *by* $X$ *as*

$$(X.F)(P) \coloneqq \sum_{x \in E_X} P_X(x) F(P|_{X=x}). \tag{1}$$

*Then* $X.F \in \mathrm{Meas}(\Delta(\Omega), \mathbb{R})$*. This definition does not depend on the representative of the equivalence class of* $X$*.*

Let $M$ be a monoid and $G$ an abelian group. Then an additive monoid action (or monoid action for short) is a function $. : M \times G \to G$ for which $\mathbf{1} \in M$ acts trivially ($\mathbf{1}.g = g$), which is associative ($X.(Y.g) = (X \cdot Y).g$), and which is additive ($X.(g + h) = X.g + X.h$) — which also implies $X.0 = 0$. Additive monoid actions generalize the conditioning operation of information functions on random variables, as the following proposition, whose proof we leave to the readers, shows:

**Proposition 2.4.** *Let* $M$ *be a monoid of equivalence classes of random variables and* $G = \mathrm{Meas}(\Delta(\Omega), \mathbb{R})$*. The averaged conditioning* $. : M \times G \to G$ *from Definition 2.3 is a monoid action.*

This viewpoint of the averaged conditioning was perhaps first studied in the context of information cohomology [Baudot and Bennequin, 2015, Baudot et al., 2019, Vigneaux, 2019]. The proof of the following well-known chain rule of Shannon entropy is also left to the reader to prove:

**Proposition 2.5** (Chain Rule)**.** *The following chain rule*

$$I(XY) = I(X) + X.I(Y)$$

*holds for arbitrary random variables* $X : \Omega \to E_X$ *and* $Y : \Omega \to E_Y$*.*

We remark that $X.I(Y)$ is typically written as $I(Y \mid X)$ in other literature on information theory. In our context, the notation $X.I(Y)$ is more natural, as it emphasizes that it comes from a monoid action. For the following definition, set $I_1 \coloneqq I$, which then embeds Shannon entropy into the set of higher-order information functions:

**Definition 2.6** (Mutual Information, Interaction Information)**.** *Let $q \in \mathbb{N}$ and assume that $I_{q-1}$ is already defined. Assume also that $Y_1, \dots, Y_q$ are $q$ random variables on $\Omega$. Then we define $I_q(Y_1; \dots; Y_q) \in \mathrm{Meas}(\Delta(\Omega), \mathbb{R})$, the* interaction information of degree $q$*, as the function*

$$I_q(Y_1; \dots; Y_q) \coloneqq I_{q-1}(Y_1; \dots; Y_{q-1}) - Y_q.I_{q-1}(Y_1; \dots; Y_{q-1}).$$

*$I_2$ is also called mutual information. This definition does not depend on the representatives of the equivalence classes of $Y_1, \dots, Y_q$. For a specific probability mass function $P \in \Delta(\Omega)$, we set $I_q(Y_1; \dots; Y_q; P) \coloneqq \big[I_q(Y_1; \dots; Y_q)\big](P) \in \mathbb{R}$, and, when there is another random variable $X$ on $\Omega$ on which we condition, $X.I_q(Y_1; \dots; Y_q; P) \coloneqq \big[X.I_q(Y_1; \dots Y_q)\big](P)$.*

We now summarize the abstract properties of interaction information $I_q$. Let $M$ be a commutative, idempotent monoid of (equivalence classes of) random variables as in Proposition 2.1. By abuse of notation, we do not distinguish between random variables and their equivalence classes, i.e., we write $Y$ instead of $[Y]$. Denote by $G \coloneqq \mathrm{Meas}\big(\Delta(\Omega), \mathbb{R}\big)$ the abelian group of measurable functions from $\Delta(\Omega)$ to $\mathbb{R}$. Averaged conditioning $. : M \times G \to G$ is a well-defined monoid action.

We can view $I_q$ as a function $I_q : M^q \to G$ that is defined on tuples of *equivalence classes* of random variables. By Proposition 2.5, entropy $I_1$ satisfies the equation

$$I_1(XY) = I_1(X) + X.I_1(Y)$$

for all $X, Y \in M$, where $X.I_1(Y)$ is the result of the action of $X \in M$ on $I_1(Y) \in G$ via averaged conditioning. Finally, by Definition 2.6, for all $q \geq 2$ and all $Y_1, \dots, Y_q \in M$, one has

$$I_q(Y_1; \dots; Y_q) = I_{q-1}(Y_1; \dots; Y_{q-1}) - Y_q.I_{q-1}(Y_1; \dots; Y_{q-1}).$$

## 2.2 The Generalized Hu Theorem and $F$-Diagrams

We now work towards a presentation of the generalized Hu theorem from Lang et al. [2025], which generalizes Hu [1962] and the $I$-diagrams from Yeung [1991]. These diagrams show in one overview the (higher-order) information functions of a set of variables and how they additively compose each other. Fix an abelian group $G$ (generalizing $\mathrm{Meas}(\Delta(\Omega), \mathbb{R})$ from above) and a commutative, idempotent monoid $M$ (generalizing a monoid of equivalence classes of random variables). We also fix an additive monoid action $. : M \times G \to G$, generalizing the averaged conditioning. For a set $\Sigma$, denote by $2^\Sigma$ its power set, i.e., the set of its subsets.

**Definition 2.7** ($(G$-Valued) Measure)**.** *Let $G$ be an abelian group and $\Sigma$ a set.*[1] *A* $G$-valued measure *is a function $\mu : 2^\Sigma \to G$ with the property*

$$\mu(A_1 \cup A_2) = \mu(A_1) + \mu(A_2)$$

*for all disjoint $A_1, A_2 \subseteq \Sigma$. One automatically obtains $\mu(\emptyset) = 0$, and $\mu$ turns arbitrary finite disjoint unions into the corresponding finite sums.*

We now fix elements $X_1, \dots, X_n \in M$, $n \geq 0$. These are the elements for which we want to obtain an $F$-diagram later on. Since $M$ is commutative, every product of these elements (of arbitrary order and multiplicity) can be reordered such that all $X_i$ with the same index $i$ are next to each other. Then, since $M$ is idempotent, we can reduce the product further until each $X_i$ appears maximally once. This means that general products of the $X_i$ are of the form

$$X_I \coloneqq \prod_{i \in I} X_i \coloneqq X_{i_1} X_{i_2} \cdots X_{i_q} \tag{2}$$

for some possibly empty subset $I = \{i_1 < i_2 < \dots < i_q\} \subseteq [n] = \{1, \dots, n\}$. Furthermore, we have $X_I X_J \coloneqq X_I \cdot X_J = X_{I \cup J}$.

[1] We only make use of the case $\Sigma = \widetilde{X}$ as defined below in Equation (3).

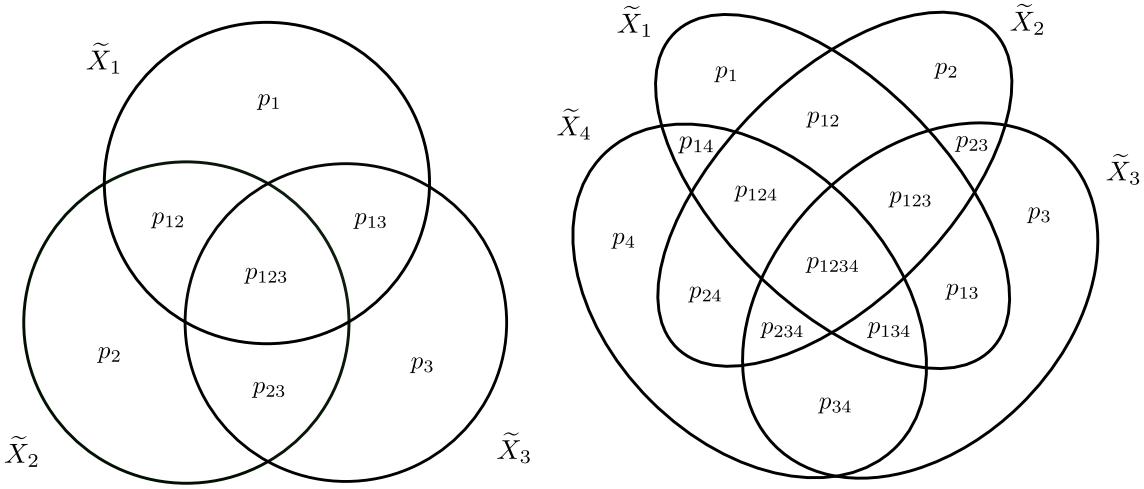



**Figure 1:** A depiction of $\widetilde{X} = \widetilde{X}(n)$ for $n = 3$ and $n = 4$, which will later be used to represent all the (higher-order) information functions.

**Definition of $\widetilde{X}$** We set

$$\widetilde{X} \coloneqq \widetilde{X}(n) \coloneqq 2^{[n]} \setminus \{\emptyset\}.$$

For $\emptyset \neq I \subseteq [n]$, we will write $p_I \coloneqq I$. We can then also write

$$\widetilde{X} = \Big\{ p_I \;\Big|\; \emptyset \neq I \subseteq [n] \Big\}. \tag{3}$$

For $i \in [n]$, we denote by $\widetilde{X}_i \coloneqq \big\{ p_I \in \widetilde{X} \;\big|\; i \in I \big\}$ a set which we can imagine to be depicted by a disk corresponding to the element $X_i$. We visualize this in Figure 1. This is actually the simplest construction that leads to the $\widetilde{X}_i$ being in general position, meaning that for each choice of a nonempty set of disks indexed by $i \in I$, there is a single point inside all of them and not in any of the others:

$$\bigcap_{i \in I} \widetilde{X}_i \;\setminus\; \bigcup_{j \in [n] \setminus I} \widetilde{X}_j = \{p_I\}. \tag{4}$$

Consequently, we call $p_I$ also the *atom* corresponding to $I$, as it is an atomic part of a diagram of intersecting disks.

With $\widetilde{X}_I \coloneqq \bigcup_{i \in I} \widetilde{X}_i$ we denote the union of the disks corresponding to the joint variable $X_I$. Clearly, we have $\widetilde{X} = \widetilde{X}_{[n]}$. In the following, we will also be flexible with our notation. For example, if we have a commutative, idempotent monoid $M$ and fixed elements $X, Y, Z, W \in M$, then we can also define

$$\widetilde{XYZW} = \widetilde{X} \cup \widetilde{Y} \cup \widetilde{Z} \cup \widetilde{W}.$$

The atom $p_{XZ}$ would then be characterized by

$$\{p_{XZ}\} = \Big( \widetilde{X} \cap \widetilde{Z} \Big) \setminus \Big( \widetilde{Y} \cup \widetilde{W} \Big).$$

The measure constructed in the proof of the following theorem is essentially constructed using a Möbius inversion formula:

**Theorem 2.8** (Generalized Hu Theorem, Lang et al. [2025])**.** *Let $M$ be a commutative, idempotent monoid, $G$ an abelian group, and $. : M \times G \to G$ an additive monoid action.*

*Assume $F_1 : M \to G$ is a function that satisfies the following chain rule: for all $X, Y \in M$, one has*

$$F_1(XY) = F_1(X) + X.F_1(Y). \tag{5}$$

*Construct $F_q : M^q \to G$ for $q \geq 2$ inductively by*

$$F_q(Y_1; \ldots; Y_q) \coloneqq F_{q-1}(Y_1; \ldots; Y_{q-1}) - Y_q.F_{q-1}(Y_1; \ldots; Y_{q-1}) \tag{6}$$

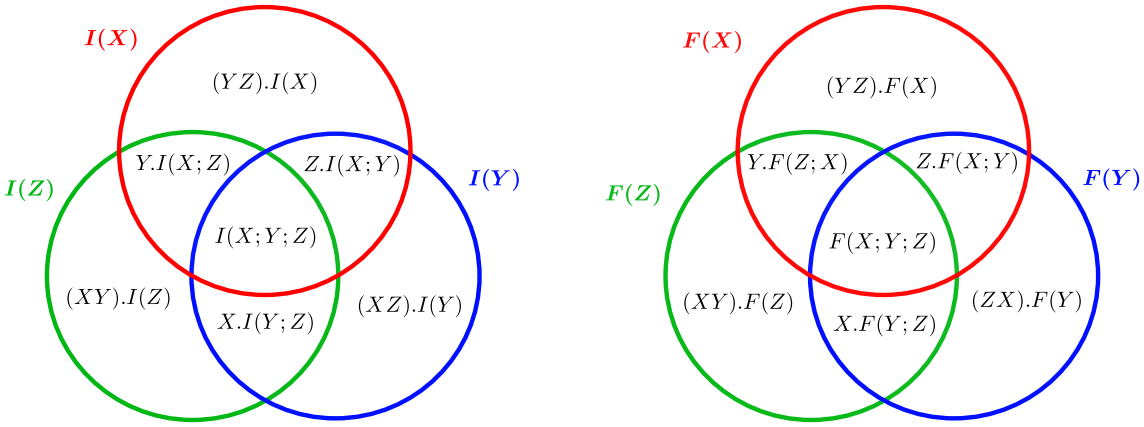



**Figure 2:** A depiction of the $I$-diagram and the $F$-diagram from the (generalized) Hu theorem. On the left, it shows the interplay of (conditional) Shannon entropy, mutual information, and interaction information for three random variables $X, Y, Z$. On the right, $X, Y$ and $Z$ are elements of a commutative, idempotent monoid, generalizing the collection of equivalence classes of random variables together with their joint operation, and the higher-order terms are all derived from a function $F$ satisfying the chain rule $F(XY) = F(X) + X.F(Y)$.

*for all* $Y_1, \dots, Y_q \in M$.

*Fix elements* $X_1, \dots, X_n \in M$, $n \geq 0$. *Set* $\widetilde{X} = \widetilde{X}(n)$ *as in Equation* (3). *Then there exists a* $G$*-valued measure* $\widetilde{F} : 2^{\widetilde{X}} \to G$ *such that for all* $q \geq 1$ *and* $J, L_1, \dots, L_q \subseteq [n]$, *the following identity holds:*

$$X_J.F_q(X_{L_1}; \dots; X_{L_q}) = \widetilde{F}\left( \bigcap_{k=1}^{q} \widetilde{X}_{L_k} \setminus \widetilde{X}_J \right). \tag{7}$$

Note that in the preceding theorem, the $G$-valued measure $\widetilde{F}$ depends on the fixed elements $X_1, \dots, X_n$. For ease of notation, we write every $F_q$ simply as $F$. Thus, in an expression of the form $F(X; Y; Z)$, $F$ is necessarily $F_3$. Since Shannon entropy $I$ and its higher order generalizations satisfy all the assumptions from Theorem 2.8, as explained in Section 2.1, we recover the $I$-diagrams from Yeung [1991] as a special case. We refer back to Figure 2 for a depiction of the $I$-diagram or $F$-diagram that results from the (generalized) Hu theorem. For figures that explain how to use Hu's theorem to visualize additive identities involving $F$, we refer to Lang et al. [2025]. As an example, we note that Figure 2 shows that

$$I(X) = (YZ).I(X) + Y.I(X;Z) + Z.I(X;Y) + I(X;Y;Z),$$

as we can see by observing how the disk corresponding to $I(X)$ decomposes. Formally, such identities follow from the fact that $\widetilde{F}$ (or $\widetilde{I}$ in the special case of Shannon's information functions) is a *measure*, which means it is additive over disjoint unions in $\widetilde{X}$.

## 2.3 Graph Terminology

Before introducing Markov random fields, we need to discuss some necessary graph terminology.

**Definition 2.9** (Graph)**.** *A* graph *is a tuple* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, *where*

- $\mathcal{V}$ *is a finite set, called the set of* vertices*;*
- $\mathcal{E} \subseteq \big\{\{i, j\} \subseteq \mathcal{V} \mid i \neq j\big\}$, *which we call the set of* edges.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph. Instead of $\{i, j\} \in \mathcal{E}$, we also write $i — j$, or $i — j \in \mathcal{E}$. Note that edges are non-oriented, as indicated by them being sets instead of tuples. Furthermore, we have $i \neq j$ for every edge $i — j$, which means that there are no *loops* (i.e., edges from a vertex to itself) in the graph. However, there can be *cycles*, i.e., walks from a vertex to itself passing through other vertices.

For a subset $\mathcal{U} \subseteq \mathcal{V}$, we can define $\mathcal{G}^{\setminus \mathcal{U}}$ as the graph obtained from $\mathcal{G}$ by removing all the vertices in $\mathcal{U}$ and all the edges that have at least one endpoint in $\mathcal{U}$. Formally, one can write

$$\mathcal{G}^{\setminus \mathcal{U}} := (\mathcal{V} \setminus \mathcal{U}, \mathcal{E}^{\setminus \mathcal{U}}), \quad \text{where} \quad \mathcal{E}^{\setminus \mathcal{U}} := \big\{\{i,j\} \in \mathcal{E} \mid i,j \in \mathcal{V} \setminus \mathcal{U}\big\}.$$

A *walk* in $\mathcal{G}$ from $i$ to $j$ is a sequence of edges $i\text{—}i_1, i_1\text{—}i_2, \dots, i_n\text{—}j$ in $\mathcal{E}$. We write such a walk as $i\text{—}i_1\text{—}\dots\text{—}i_n\text{—}j$. The empty sequence is considered to be a walk from a vertex $i$ to itself, and walks are allowed to have the same vertices twice. A vertex set $\mathcal{U} \subseteq \mathcal{V}$ is called *connected in* $\mathcal{G}$ if for all $i \neq j \in \mathcal{U}$ there is a walk in $\mathcal{G}$ from $i$ to $j$.[2] A set $\mathcal{U} \subseteq \mathcal{V}$ is called a *component* if it is a maximal connected set, that is:

- $\mathcal{U}$ is connected in $\mathcal{G}$;
- if $\mathcal{U} \subseteq \mathcal{W}$ and $\mathcal{W}$ is connected in $\mathcal{G}$, then $\mathcal{U} = \mathcal{W}$.

It is easy to see that the set of components of $\mathcal{G}$ forms a partition of $\mathcal{V}$: formally, the components are the equivalence classes under the equivalence relation $\sim$ on $\mathcal{V}$ that is defined by $i \sim j$ if there is a walk from $i$ to $j$ in $\mathcal{G}$. For a vertex set $\mathcal{U} \subseteq \mathcal{V}$, we define $s(\mathcal{U})$ as the number of components in the graph $\mathcal{G}^{\setminus \mathcal{U}}$. In general, we denote by $\mathcal{V}_1(\mathcal{U}), \dots, \mathcal{V}_{s(U)}(\mathcal{U})$ these components. We call $\mathcal{U}$ a *cutset* if $s(\mathcal{U}) > 1$. Note that cutsets do not necessarily need to "cut" anything: if $\mathcal{G}$ already contains several components, then $\mathcal{U} = \emptyset$ is a cutset. Finally, for disjoint subsets $\mathcal{A}, \mathcal{B}, \mathcal{C} \subseteq \mathcal{V}$, we say that $\mathcal{C}$ *separates* $\mathcal{A}$ from $\mathcal{B}$ if *every* walk from *any* vertex in $\mathcal{A}$ to *any* vertex in $\mathcal{B}$ passes through *some* vertex in $\mathcal{C}$. Note: if $\mathcal{A} \neq \emptyset \neq \mathcal{B}$ in this definition, then $\mathcal{C}$ is automatically a cutset.

We now define the notions of connected and disconnected atoms with respect to a graph $\mathcal{G}$, which builds a bridge between the graph and information diagrams. They are identical to the notions of type I and type II atoms defined in Yeung et al. [2002].

**Definition 2.10** (Connected and Disconnected Atoms)**.** *Let $n \in \mathbb{N}$ and $\mathcal{G}$ a graph with vertex set $\mathcal{V} = [n]$, and let $\widetilde{X} = \widetilde{X}(n)$ be defined as in the previous subsection. For $\mathcal{W} \subseteq [n]$, the corresponding atom $p_{\mathcal{W}} \in \widetilde{X}$ is called* disconnected *with respect to $\mathcal{G}$ if $\mathcal{V} \setminus \mathcal{W} = [n] \setminus \mathcal{W}$ is a cutset, i.e., if $\mathcal{W}$ is disconnected as a vertex set in $\mathcal{G}^{\setminus \left([n] \setminus \mathcal{W}\right)}$. Otherwise, $p_{\mathcal{W}}$ is called* connected *(with respect to $\mathcal{G}$).*

## 2.4 Yeung's Characterization of Markov Random Fields via $I$-diagrams

Markov random fields encode independence relations of random variables that correspond to separations in a corresponding graph. As such, we first need to define:

**Definition 2.11** (Probabilistic Conditional Independence)**.** *Let $X, Y, Z$ be random variables on $\Omega$ and $P \in \Delta(\Omega)$ a probability mass function. Then $X$ and $Y$ are said to be* $P$-independent *given $Z$, written*

$$X \underset{P}{\perp\!\!\!\perp} Y \mid Z,$$

*if for all $(x, y, z) \in E_X \times E_Y \times E_Z$, the joint distribution decomposes as follows:*[3]

$$P(x, y, z) = P(x \mid z) \cdot P(y, z). \tag{8}$$

*Equivalently, for all $z \in E_Z$ with $P(z) \neq 0$ one has*

$$P(x, y \mid z) = P(x \mid z) \cdot P(y \mid z).$$

*Equivalently, for all $y, z \in E_Y \times E_Z$ with $P(y, z) \neq 0$ one has*

$$P(x \mid y, z) = P(x \mid z).$$

[2] Note that in Figure 3, we will also have a notion of connectedness. That notion, however, is about connectedness in $\mathcal{G}^{\setminus \left(\mathcal{V} \setminus \mathcal{U}\right)}$, see Definition 2.10.

[3] If $P(z) = 0$, then $P(x \mid z)$ is not defined. But then $P(x, y, z) = 0$ and $P(y, z) = 0$, so we can simply define the right-hand-side of the equation as zero.

If $X' \sim X$, $Y' \sim Y$, $Z' \sim Z$, and $P \in \Delta(\Omega)$, then

$$X \underset{P}{\perp\!\!\!\perp} Y \mid Z \quad \iff \quad X' \underset{P}{\perp\!\!\!\perp} Y' \mid Z',$$

see, for example, Dawid [2001], Section 6.2. Thus, also in this context, it is suitable to identify a random variable with its equivalence class. This definition gives rise to a separoid, which is a useful abstract structure that applies to all settings that we study in this paper. We use the slightly adapted but equivalent version of the separoid axioms from Forré [2021], Appendix A.4:

**Definition 2.12** (Separoid, Separoid Axioms)**.** *Let $M$ be a commutative, idempotent monoid. Let $\perp\!\!\!\perp$ be a relation on $M^2 \times M$, written for $X, Y, Z \in M$ by*

$$X \perp\!\!\!\perp Y \mid Z.$$

*The tuple $(M, \perp\!\!\!\perp)$ is called a* separoid *if $\perp\!\!\!\perp$ satisfies the following* separoid axioms *for all $X, Y, Z, W \in M$:*

(S1) symmetry : $X \perp\!\!\!\perp Y \mid Z \implies Y \perp\!\!\!\perp X \mid Z$;
(S2) redundancy : $W \precsim Z \implies W \perp\!\!\!\perp Y \mid Z$;
(S3) decomposition : $WX \perp\!\!\!\perp Y \mid Z \implies X \perp\!\!\!\perp Y \mid Z$;
(S4) weak union : $WX \perp\!\!\!\perp Y \mid Z \implies W \perp\!\!\!\perp Y \mid XZ$;
(S5) contraction : $\big(W \perp\!\!\!\perp Y \mid XZ$ and $X \perp\!\!\!\perp Y \mid Z\big) \implies WX \perp\!\!\!\perp Y \mid Z$.

For a relationship to the more familiar notion of a (semi-)graphoid, see [Dawid, 2001, Section 3.2]. In this paper, if a relation $X \perp\!\!\!\perp Y \mid Z$ holds, then we also say "$X$ and $Y$ are independent given $Z$", or "$X$ and $Y$ are independent conditioned on $Z$", and often append "with respect to $\perp\!\!\!\perp$" to clarify the independence relation.

Note that properties (S2)–(S5) have obvious "right-handed versions" as well due to symmetry (S1). When we refer to one of the properties (S2)–(S5), then we mean one of the two versions depending on the context. We obtain:

**Proposition 2.13.** *Let $M$ be a monoid of random variables as in Proposition 2.1 and $P \in \Delta(\Omega)$ a probability mass function. Then $(M, \perp\!\!\!\perp_P)$ is a separoid.*

*Proof.* This is well-known and appeared originally in this formulation in Dawid [2001]. It also follows from the generalization given in Forré [2021], Theorem 3.11. □

We are now ready to define Markov random fields via the global Markov property, in general separoids and also in the specific probabilistic case:

**Definition 2.14** (Global Markov Property, Markov Random Field)**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with $\mathcal{V} = [n]$ and $(M, \perp\!\!\!\perp)$ a separoid. A sequence of elements $X_1, \dots, X_n \in M$ is called a* Markov random field *with respect to $\perp\!\!\!\perp$ and $\mathcal{G}$ if it satisfies the* global Markov property*, that is: for all disjoint $\mathcal{A}, \mathcal{B}, \mathcal{C} \subseteq \mathcal{V}$ such that $\mathcal{C}$ separates $\mathcal{A}$ from $\mathcal{B}$, we have*

$$X_{\mathcal{A}} \perp\!\!\!\perp X_{\mathcal{B}} \mid X_{\mathcal{C}}.^{4}$$

*Finally, in the probabilistic context, if random variables $X_1, \dots, X_n$ on $\Omega$ form a Markov random field with respect to $\mathcal{G}$ and $\perp\!\!\!\perp_P$, then we say that $X_1, \dots, X_n$ form a $P$-Markov random field with respect to $\mathcal{G}$.*

[4] We could have also defined the separation relation $A \perp B \mid C$ on $\mathcal{G}$, defined for all vertex sets $A, B, C$ in $\mathcal{G}$, and not only those that are disjoint as in our terminology. This relation would satisfy the separoid axioms. The global Markov property is then equivalent to the statement that $A \perp B \mid C$ implies $X_{\mathcal{A}} \perp\!\!\!\perp X_{\mathcal{B}} \mid X_{\mathcal{C}}$. This, in turn, can be phrased as the statement that $A \mapsto X_A$ constitutes a *separoid homomorphism* as defined in Dawid [2001, Definition 1.3].

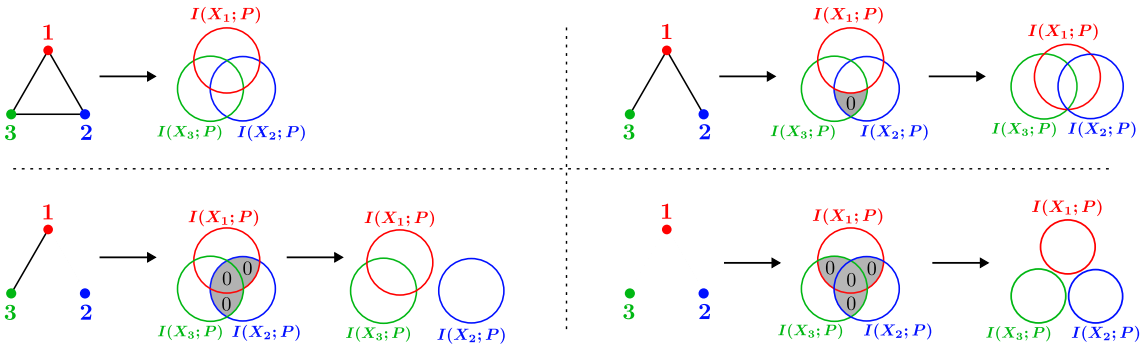



**Figure 3:** We show the effect of the graph structure of simple Markov random fields on the corresponding $I$-diagrams for a fixed probability mass function $P$. The $I$-diagram visualizes the relationships of the entropy, mutual information, and interaction information of the variables. The figure is based on Yeung's result Theorem 2.15, which shows that a set of random variables forms a Markov random field corresponding to a graph if and only if all atoms in the $I$-diagram corresponding to disconnected sets of vertices in the graph disappear.
Concretely, to a set of vertices $J$, the corresponding atom in the $I$-diagram is the intersection of all disks with indices $j \in J$, without any element in the union of all the other disks. For the lower left panel, the three sets of vertices $\{1,2\}$, $\{2,3\}$, and $\{1,2,3\}$ are disconnected, giving rise to three disappearing atoms in the $I$-diagram. Consequently, $I(X_2)$ can be drawn to not intersect with the other disks. However, the other four sets of vertices $\{1\},\{2\},\{3\},\{1,3\}$ are clearly connected, which means that we cannot infer their corresponding atoms to vanish. Similar reasoning applies to the other three panels.

The definition of $P$-Markov random fields is stated in terms of conditional $P$-independences. Now, one crucial observation is the well-known fact that such independences can be characterized using conditional mutual information:

$$X \underset{P}{\perp\!\!\!\perp} Y \mid Z \quad \Longleftrightarrow \quad Z.I(X;Y;P) = 0. \tag{9}$$

Yeung's insight was that this should make it possible to characterize $P$-Markov random fields by properties of their corresponding $I$-diagrams. Let $M$ be the monoid of (equivalence classes of) random variables on $\Omega$, and $I : M \to \mathrm{Meas}\big(\Delta(\Omega), \mathbb{R}\big)$ the Shannon entropy function, as defined in Definition 2.2. Let $X_1, \dots, X_n$ be fixed random variables on the sample space $\Omega$. Let $\widetilde{I} : 2^{\widetilde{X}} \to \mathrm{Meas}\big(\Delta(\Omega), \mathbb{R}\big)$ be the measure resulting from Hu's Theorem 2.8. For any probability mass function $P \in \Delta(\Omega)$, we then obtain the "slice"

$$\widetilde{I}^P : 2^{\widetilde{X}} \to \mathbb{R}, \quad A \mapsto \big[\widetilde{I}(A)\big](P). \tag{10}$$

This is a signed measure with values in the real numbers instead of functions, and it visualizes the interplay of the (higher-order) Shannon information functions *for the specific probability mass function* $P$.

This can be imagined as a "slice" of the original $I$-diagram: e.g., in the special case that $\Omega = \{0,1\}$, the mass functions $P \in \Delta(\Omega)$ are essentially numbers in $[0,1]$. For each $P \in [0,1]$, we then get an information diagram corresponding to $\widetilde{I}^P$. If we stack these "on top of each other", the full stack is the $I$-diagram of $\widetilde{I}$, with the "slices" given by the diagrams for $\widetilde{I}^P$. Yeung et al. [2002] then showed:

**Theorem 2.15** (Characterization of $P$-Markov Random Fields)**.** *Let $I : M \to \mathrm{Meas}\big(\Delta(\Omega), \mathbb{R}\big)$ be the Shannon entropy function. Let $X_1, \dots, X_n$ be random variables on $\Omega$ and $P \in \Delta(\Omega)$ a probability mass function, giving rise to $\widetilde{I}^P : 2^{\widetilde{X}} \to \mathbb{R}$ by Equation (10). Let $\mathcal{G}$ be a graph with vertex set $[n]$. Then the following two statements are equivalent:*

- *$X_1, \dots, X_n$ form a $P$-Markov random field with respect to $\mathcal{G}$;*
- *$\widetilde{I}^P(p_{\mathcal{W}}) = 0$ for all atoms $p_{\mathcal{W}}$ that are disconnected with respect to $\mathcal{G}$.*

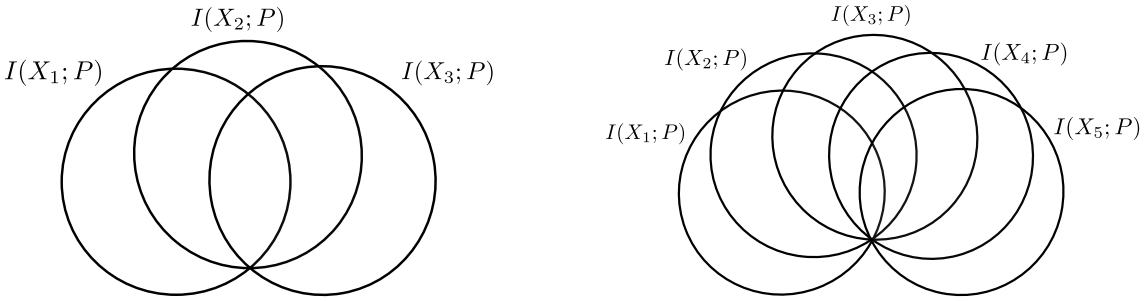



**Figure 4:** If random variables $X_1, \dots, X_n$ form a $P$-Markov chain, then many atoms in the $I$-diagram with respect to $P$ disappear by Corollary 2.17. The only atoms that remain are those corresponding to "intervals" in $[n]$. This leads to a fan-like structure of the $I$-diagram with respect to $P$, as visualized here for $n = 3$ and $n = 5$.

We visualize this result in Figure 3, which shows the intuitive relation between the graph of a Markov random field and vanishing regions of atoms that are "disconnected" according to the graph. One can then specialize the theorem to Markov chains:

**Definition 2.16** (Markov Chain)**.** *Let $n \geq 0$ and $(M, \perp\!\!\!\perp)$ a separoid. A sequence of elements $X_1, \dots, X_n$ in $M$ is called a* Markov chain *(with respect to $\perp\!\!\!\perp$) if for all $2 \leq i \leq n$, the following independence holds:*

$$X_i \perp\!\!\!\perp X_{[i-2]} \mid X_{i-1}.$$

*We also write $X_1 \to X_2 \to \cdots \to X_n$ to indicate that the sequence $X_1, \dots, X_n$ forms a Markov chain.*

*Finally, in the probabilistic context, a collection of random variables $X_1, \dots, X_n$ on $\Omega$ are said to form a $P$-Markov chain if they form a Markov chain with respect to $\perp\!\!\!\perp_P$.*

For $n = 0, 1$ or 2, all sequences form a Markov chain. The first interesting case is $n = 3$. The only non-vacuous condition for a Markov chain is then $X_3 \perp\!\!\!\perp X_1 \mid X_2$. The following corollary of Theorem 2.15 was already shown in Kawabata and Yeung [1992]:

**Corollary 2.17.** *With all notation as above, the following two statements are equivalent:*

- *$X_1, \dots, X_n$ form a $P$-Markov chain.*
- *$\widetilde{I}^P(p_{\mathcal{W}}) = 0$ for all $\mathcal{W} \subseteq [n]$ that do* not *only contain consecutive numbers.*

We visualize this result in Figure 4. As an example, this structure of the $I$-diagram can be used to prove the well-known data processing inequality:

**Corollary 2.18** (Data Processing Inequality)**.** *Let $X_1 \to X_2 \to X_3$ be a $P$-Markov chain of random variables on $\Omega$. Then $I(X_1; X_3; P) \leq I(X_1; X_2; P)$.*

*Proof.* Using Hu's theorem, Theorem 2.8, Equation (10), and the Venn diagram of the Markov chain from Figure 4, we obtain:

$$\begin{aligned} I(X_1; X_2; P) &= \widetilde{I}^P(\widetilde{X}_1 \cap \widetilde{X}_2) \\ &= \widetilde{I}^P(\widetilde{X}_1 \cap \widetilde{X}_3) + \widetilde{I}^P(\widetilde{X}_1 \cap \widetilde{X}_2 \setminus \widetilde{X}_3) \\ &= I(X_1; X_3; P) + X_3.I(X_1; X_2; P) \\ &\geq I(X_1; X_3; P). \end{aligned}$$

In the last step, we used the well-known property of conditional mutual information to be non-negative. □

## 2.5 An Outline of our Generalizations of Yeung's Results

We now explain how we aim to generalize Yeung's results on the $I$-diagram characterization of Markov random fields. The motivation is that we would like to study implications of Markov random field structures on more general information diagrams. For example, assume that $X_1, \dots, X_n$ are random variables on $\Omega$ that form a Markov random field with respect to a graph $\mathcal{G}$ and two different probability mass functions $P$ and $Q$. Can we then say anything about how the Kullback-Leibler divergence $D(X_1 \cdots X_n; P \| Q)$ of $P$ and $Q$ over the whole joint variable $X_1 \cdots X_n$ distributes into components of different variables in this structure? Such questions are commonplace for example in machine learning [Bishop, 2007], where it is sometimes assumed that both the model- and data distribution follow a specific graphical form, and the loss function takes the form of a Kullback-Leibler divergence (or, equivalently up to a constant, cross-entropy) between the two. We will demonstrate this type of application by showing how our theory allows to find a conceptually simple derivation of a decomposition of the loss function of diffusion models in Section 5.5.

In this whole subsection, let $M$ be a commutative, idempotent monoid acting on an abelian group $G$, and $F : M \to G$ a function satisfying the chain rule, Equation (5). To generalize Yeung's characterization, we use the setting of $F$-diagrams as in Hu's theorem, Theorem 2.8. Motivated by Equation (9), which relates $P$-independence to mutual information, we generalize $P$-independence using $F$ itself.

**Definition 2.19** ($F$-independence)**.** *We define the relation $\perp\!\!\!\perp_F$ on $M^2 \times M$ by*

$$X \underset{F}{\perp\!\!\!\perp} Y \mid Z \quad :\Longleftrightarrow \quad Z.F(X;Y) = 0.$$

*If $X \underset{F}{\perp\!\!\!\perp} Y \mid Z$, then $X$ is called $F$-independent from $Y$ given $Z$.*

The definition is analogous to the conditional independence relation defined for general submodular information functions in Steudel et al. [2010]. In Proposition 4.8 we will show that $(M, \perp\!\!\!\perp_F)$ forms a separoid. We can then specialize the notion of a Markov random field and Markov chains from general separoids to the separoid $(M, \perp\!\!\!\perp_F)$:

**Terminology 2.20** ($F$-Markov Random Field, $F$-Markov Chain)**.** *Elements $X_1, \dots, X_n \in M$ are said to form an $F$-Markov random field with respect to $\mathcal{G}$ if they form a Markov random field with respect to $\perp\!\!\!\perp_F$ and $\mathcal{G}$, see Definition 2.14.*

*Similarly, $X_1, \dots, X_n$ are said to form an $F$-Markov chain if they form a Markov chain with respect to $\perp\!\!\!\perp_F$, see Definition 2.16.*

We will then prove the following theorem in Section 4.5:

**Theorem 2.21** ($F$-Markov Random Field Characterization)**.** *Let $M$ be a commutative, idempotent monoid acting additively on an abelian group $G$, and $F : M \to G$ a function satisfying the chain rule Equation (5). Additionally, fix elements $X_1, \dots, X_n$ giving rise to $\widetilde{F} : 2^{\widetilde{X}} \to G$, and a graph $\mathcal{G}$ with vertex set $[n]$. Then the following statements are equivalent:*

- *$X_1, \dots, X_n$ form an $F$-Markov random field with respect to $\mathcal{G}$;*
- *$\widetilde{F}(p_{\mathcal{W}}) = 0$ for all disconnected atoms $p_{\mathcal{W}} \in \widetilde{X}$.*

## 2.6 An Outline of the Coming Sections

In Section 3, we take a step back from independences induced by a function $F$ by studying independences in separoids, Definition 2.12, which generalizes both $P$-independence and $F$-independence. We also define conditional mutual independences $\perp\!\!\!\perp_{i=1}^{n} X_i \mid Y$ and prove a simple characterization. We study Markov random fields in separoids and characterize them equivalently by the *cutset property*; It states that if a vertex set $\mathcal{U}$ is a "cutset" that, when removed, cuts the rest of the graph into components $\mathcal{V}_1, \dots, \mathcal{V}_q$, then this leads to a corresponding mutual independence $\perp\!\!\!\perp_{i=1}^{q} X_{\mathcal{V}_i} \mid X_{\mathcal{U}}$. Finally, we characterize Markov chains in separoids.

In Section 4, we first study important consequences of Hu's theorem, Theorem 2.8, including the statement that $F$-independence gives rise to a separoid, Proposition 4.8. One crucial property

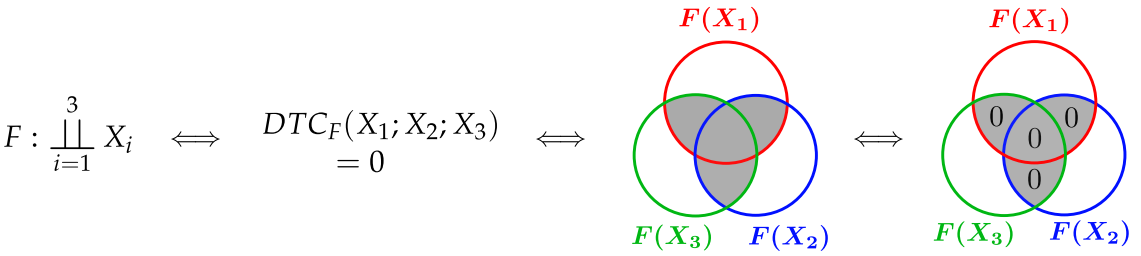



**Figure 5:** One key ingredient in the $F$-diagram characterization of $F$-Markov random fields is the characterization of $F$-mutual independences, Theorem 4.15, here visualized for three elements $X_1, X_2, X_3 \in M$. The characterization shows that the mutual independence is equivalent to the vanishing of $F$-dual total correlation $DTC_F(X_1; X_2; X_3)$, which, by Hu's theorem, Theorem 2.8, corresponds to the vanishing of a region of four atoms in the $F$-diagram, visualized as gray. Subset determination, Theorem 4.1, then allows to conclude that every *individual* atom in this region vanishes, as shown in the rightmost part of the figure. The implication from right to left again follows from Hu's theorem and the fact that $F$-diagrams visualize a *measure*, meaning that larger regions are the sum of their atoms.

that is used in its proof is what we call *subset determination*. It states that when the value of a region in an $F$-diagram is known, this fully determines the value of all subregions. In particular, if a region in an $F$-diagram vanishes, then all subsets vanish as well. A precise formulation is given in Theorem 4.1. Crucially, this replaces the repeated usage of inequalities in the proofs in [Yeung et al., 2002], which we could not make use of in our general context. As an aside, in Appendix A, we show that the basic idea of subset determination implies a characterization of all functions $F$ satisfying the chain rule, as long as $M$ contains a "top element" $\top$ such that $X \cdot \top = \top$ for all $X \in M$ — which is the case for finitely generated $M$. The functions $F$ then correspond to elements in $G$ that are *annihilated* by $\top$, by the simple mapping $F \mapsto F(\top)$. We also provide a cohomological interpretation of this result.

To avoid misunderstandings, we mention a subtlety with subset determination: it only applies to $F$-diagrams for functions $F : M \to G$ from a monoid to an abelian group *on which the monoid additively acts*. Formally, for information functions such as entropy or Kullback-Leibler divergence, it becomes invalid when *fixing the underlying probability mass functions*. For example, the joint entropy $I(XY; P)$ does not determine the mutual information $I(X; Y; P)$ even though the entire function $I(XY)$ determines the function $I(X; Y)$ via the monoid action. Nevertheless, we will manage to apply our results to classical information functions in Section 5 and can deduce Yeung's results for fixed probability mass functions, including Theorem 2.15, as we demonstrate in Appendix D.

In the later parts of Section 4, we generalize the results from Kawabata and Yeung [1992], Yeung et al. [2002]. We start by a simple characterization of pairwise $F$-independences via the $F$-diagram. Motivated by this special case, and generalizing the classical case [Han, 1978], we define $F$-dual total correlation by

$$DTC_F(X_1; \dots; X_n) \coloneqq F(X_{[n]}) - \sum_{i=1}^{n} X_{[n]\setminus i}.F(X_i),$$

where $X_{[n]} = X_1 \cdots X_n$. In Theorem 4.15, we then characterize conditional *mutual* $F$-independences $F : \coprod_{i=1}^{n} X_i \mid Y$ by the vanishing of (conditional) $F$-dual total correlation $Y.DTC_F(X_1; \dots; X_n)$. The theorem also uses Hu's theorem to characterize this by the vanishing of the corresponding atomic regions. This crucially uses subset determination, Theorem 4.1, as visualized in Figure 5. Notably, we also study $F$-total correlation $TC_F$, generalizing classical total correlation [Watanabe, 1960], and find that it also provides a characterization of mutual $F$-independences, albeit only if the abelian group $G$ is *torsion-free*; The reason is that the total correlation double-counts atoms. This is studied in Appendix B.

We then prove Theorem 2.21 in Section 4.5, which shows that $F$-Markov random fields (Terminology 2.20) are fully characterized by the vanishing of atoms in the $F$-diagram that correspond to disconnected vertex sets in the underlying graph. One ingredient of Theorem 2.21 will be Theorem 4.21, the characterization of so-called full conditional mutual $F$-independences in terms of the

$F$-diagram. Finally, in Appendix C, we explain which results from Yeung et al. [2019] generalize to our setting, in particular by reproving Theorem 2.15 with our results.

In Section 5, we then come back to the case where the information functions are applied to probability mass functions. In particular, $M$ is then a monoid of equivalence classes of random variables acting additively on the abelian group $G = \mathrm{Meas}(\Delta(\Omega)^{r+1}, \mathbb{R})$ of measurable functions from tuples of $r+1$ probability mass functions over a finite sample space $\Omega$ to the real numbers. $F : M \to G$ is from a specific class of functions such as entropy (for $r = 0$), cross-entropy, or Kullback-Leibler divergence (for $r = 1$). To give meaning to the resulting functions, we briefly explain how the higher-order cross-entropy and Kullback-Leibler divergence relate to the cluster cross-entropy from Cocco and Monasson [2012]. We then adapt $G$ to only contain functions on *subsets* of probability mass functions that are conditionally stable, which allows to restrict $F$ without losing the monoid action from $M$ to $G$. As we show, the property to form a $P$-Markov random field is conditionally stable, i.e., stable under conditioning $P$ on any evaluation $X = x$ of any random variable in the Markov random field, and so we can restrict information functions $F$ to this and similar properties.

In Theorem 5.12, we obtain an important result: When restricting $F$ to tuples of probability mass functions that satisfy a conditionally stable property that *implies* the Markov random field property, then the underlying random variables formally form an $F$-Markov random field as in Section 4.[5] In particular, using Theorem 2.21, this leads to the vanishing of regions in the $F$-diagram that correspond to disconnected vertex sets. We apply this to the case that $F$ is the Kullback-Leibler divergence restricted to tuples of joint probability mass functions on a Markov chain with equal transition probabilities from one time-step to the next. In Theorem 5.13 and Figure 9, we obtain a degeneracy of the Kullback-Leibler diagram that can be interpreted as a diagram-representation of a weak version of the second law of thermodynamics: The Kullback-Leibler divergence progressively "shrinks over time". Additionally, we apply Theorem 5.12 to obtain a simple explicit derivation of the evidence lower bound in diffusion models by decomposing the Kullback-Leibler divergence over a Markov chain.

Finally, in Section 6, we summarize our results and discuss possible extensions of the theory and open questions.

# 3 Markov Random Fields in Separoids

In this section, we study Markov random fields in general separoids as defined in Definition 2.12 to prepare for our generalizations of information characterizations of Markov random fields. In Section 3.1, we define conditional *mutual* independences in separoids and present three equivalent descriptions. They are used in the *cutset* description of Markov random fields, which we prove in Section 3.2 to be equivalent to the usual global Markov property. Finally, Section 3.3 defines Markov chains in separoids and characterizes them as Markov random fields corresponding to a path-shaped graph. All these results will apply to $F$-independence, as we will in the next section, in Proposition 4.8, show that it satisfies the separoid axioms.

## 3.1 Conditional Mutual Independences in Separoids

Fix a separoid $(M, \perp\!\!\!\perp)$. We remind again of our notation: For $i, k \in \mathbb{N}$, we set $[i:k] := \{i, i+1, \dots, k\}$ if $i \leq k$ and $[i:k] = \emptyset$, otherwise. As a special case, set $[k] := [1:k]$. For a set $I$, set $I \setminus i := I \setminus \{i\}$.

**Definition 3.1** (Conditional Mutual Independence)**.** *Let $X_1, \dots X_n, Y \in M$. We say $X_1, \dots, X_n$ are* mutually independent given $Y$*, written*

$$\coprod_{i=1}^{n} X_i \mid Y,$$

[5] Importantly, there is no reverse of this statement. E.g., if two probability mass functions $P$ and $Q$ give rise to a Kullback-Leibler diagram with vanishing regions corresponding to the graph of a Markov random field, it does *not* imply that $P$ and $Q$ factorize according to the graph. For example, whenever $P = Q$, the whole Kullback-Leibler diagram trivially vanishes, but we can't conclude anything nontrivial about $P$ or $Q$ from it.

*if for all $i = 1, \dots, n$, the following pairwise independence holds:*

$$X_i \perp\!\!\!\perp X_{[n]\setminus i} \mid Y.$$

*Recall that in this expression, we have $X_{[n]\setminus i} = \prod_{j\in[n]\setminus\{i\}} X_j$.*

Note: in some contexts we may also write

$$\left(X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp \dots \perp\!\!\!\perp X_n\right) \mid Y \quad \text{or} \quad X_1 \perp\!\!\!\perp \coprod_{i=2}^{n} X_i \mid Y$$

for the mutual independence of $X_1, \dots, X_n$ given $Y$.

**Proposition 3.2.** *For $X_1, \dots, X_n, Y \in M$, the following statements are equivalent:*

1. $\coprod_{i=1}^{n} X_i \mid Y$;
2. *For all $i = 1, \dots, n$, the following pairwise independence holds:*
$$X_i \perp\!\!\!\perp X_{[i-1]} \mid Y;$$
3. $\coprod_{i=1}^{n-1} X_i \mid Y$ *and* $X_n \perp\!\!\!\perp X_{[n-1]} \mid Y$;
4. *for all disjoint $I_1, I_2 \subseteq [n]$, one has $X_{I_1} \perp\!\!\!\perp X_{I_2} \mid Y$.*

*Proof.* 1 immediately implies 2 by using decomposition (S3). Assume 2 holds. We want to prove 1. Let $i \in [n]$. Assume by induction that we already know

$$X_i \perp\!\!\!\perp X_{[l]\setminus i} \mid Y \tag{11}$$

for some $l \geq i$, where the case $l = i$ holds by assumption. If we can show the same with $l+1$ replacing $l$, then induction shows the statement for $l = n$, which then results in mutual independence since $i$ was arbitrary. For the induction step, note that

$$X_{l+1} \perp\!\!\!\perp X_{[l]} \mid Y, \tag{12}$$

which, using symmetry (S1) and weak union (S4), implies

$$X_i \perp\!\!\!\perp X_{l+1} \mid Y X_{[l]\setminus i}. \tag{13}$$

Contraction (S5) applied to Equations (11) and (13) gives

$$X_i \perp\!\!\!\perp X_{[l+1]\setminus i} \mid Y,$$

which shows the induction step.

Knowing that 1 and 2 are equivalent then immediately implies that both are equivalent to 3.

It is also clear that 4 implies 1. Finally, we show that 1 implies 4: let $I_1, I_2 \subseteq [n]$ be disjoint. The case $I_1 = \emptyset$ is trivial by redundancy (S2). Thus, assume $\emptyset \neq I_1 = \{i_1, \dots, i_k\}$ with pairwise different elements $i_l$. Then from decomposition (S3) applied to the independence $X_{i_1} \perp\!\!\!\perp X_{[n]\setminus i_1} \mid Y$ we obtain

$$X_{i_1} \perp\!\!\!\perp X_{I_2} \mid Y.$$

By induction, we can assume

$$X_{I_1\setminus i_k} \perp\!\!\!\perp X_{I_2} \mid Y \tag{14}$$

and want to show that we can add $X_{i_k}$ to the left-hand-side. By decomposition (S3) applied to the independence $X_{i_k} \perp\!\!\!\perp X_{[n]\setminus i_k} \mid Y$ we obtain

$$X_{i_k} \perp\!\!\!\perp X_{I_1\cup I_2\setminus i_k} \mid Y,$$

and thus by weak union (S4)

$$X_{i_k} \perp\!\!\!\perp X_{I_2} \mid Y X_{I_1\setminus i_k}. \tag{15}$$

Contraction (S5) applied to Equations (14) and (15) shows

$$X_{I_1} \perp\!\!\!\perp X_{I_2} \mid Y,$$

which is the independence we wanted to show. □

## 3.2 Markov Random Fields in Separoids

Fix again a general separoid $(M, \perp\!\!\!\perp)$. In this part, we show that Markov random fields as defined in Definition 2.14 by the global Markov property are equivalently described by the cutset property.

**Definition 3.3** (Full Conditional Mutual Independence (FCMI))**.** *Let $X_1, \dots, X_n \in M$. A conditional mutual independence*

$$\coprod_{i=1}^{q} X_{L_i} \mid X_J,$$

*where $L_1, \dots, L_q, J \subseteq [n]$ are pairwise disjoint sets whose union is $[n]$, is called a* full conditional mutual independence *(with respect to $n$ and $X_1, \dots, X_n$) — or FCMI for short.*

The term "full" indicates that each of the variables $X_i$, $i = 1, \dots, n$, appears exactly once. Recall the notion of a cutset and the corresponding components from Section 2.3.

**Definition 3.4** (Cutset Property)**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ a graph with $\mathcal{V} = [n]$. A sequence of elements $X_1, \dots, X_n \in M$ is said to satisfy the* cutset property *with respect to $\perp\!\!\!\perp$ and $\mathcal{G}$ if for all cutsets $\mathcal{U} \subseteq \mathcal{V}$, the FCMI*

$$\coprod_{i=1}^{s(\mathcal{U})} X_{\mathcal{V}_i(\mathcal{U})} \mid X_{\mathcal{U}}$$

*holds.*

The following proposition shows that the global Markov property and cutset property are equivalent. For the special case of probabilistic independence, this was also stated in the introduction of Yeung et al. [2002].

**Proposition 3.5.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ a graph with $\mathcal{V} = [n]$. A sequence of elements $X_1, \dots, X_n \in M$ form a Markov random field if and only if they satisfy the cutset property (with respect to $\perp\!\!\!\perp$ and $\mathcal{G}$).*

*Proof.* Assume $X_1, \dots, X_n$ satisfy the cutset property with respect to $\perp\!\!\!\perp$ and $\mathcal{G}$. Let $\mathcal{A}, \mathcal{B}, \mathcal{C} \subseteq \mathcal{V}$ disjoint such that $\mathcal{C}$ separates $\mathcal{A}$ from $\mathcal{B}$. Without loss of generality, we can assume that $\mathcal{A}$ and $\mathcal{B}$ are both nonempty, since otherwise the independence $X_{\mathcal{A}} \perp\!\!\!\perp X_{\mathcal{B}} \mid X_{\mathcal{C}}$ follows from redundancy (S2) and we are done. This, together with the separation assumption, implies that $\mathcal{C}$ is a cutset. Let $\mathcal{V}_1(\mathcal{C}), \dots, \mathcal{V}_{s(\mathcal{C})}(\mathcal{C})$ be the components of $\mathcal{G}^{\setminus \mathcal{C}}$. Since $\mathcal{A}, \mathcal{B} \subseteq \mathcal{V} \setminus \mathcal{C} = \bigcup_{i=1}^{s(\mathcal{C})} \mathcal{V}_i(\mathcal{C})$, we have

$$\mathcal{A} = \bigcup_{i=1}^{s(\mathcal{C})} \mathcal{V}_i^{\mathcal{A}}(\mathcal{C}), \quad \mathcal{B} = \bigcup_{i=1}^{s(\mathcal{C})} \mathcal{V}_i^{\mathcal{B}}(\mathcal{C}),$$

where $\mathcal{V}_i^{\mathcal{A}}(\mathcal{C}) \coloneqq \mathcal{A} \cap \mathcal{V}_i(\mathcal{C})$ and $\mathcal{V}_i^{\mathcal{B}}(\mathcal{C}) \coloneqq \mathcal{B} \cap \mathcal{V}_i(\mathcal{C})$. Now, we claim that for all $i$, we have $\mathcal{V}_i^{\mathcal{A}}(\mathcal{C}) = \emptyset$ or $\mathcal{V}_i^{\mathcal{B}}(\mathcal{C}) = \emptyset$: indeed, if there were elements $a \in \mathcal{V}_i^{\mathcal{A}}(\mathcal{C})$ and $b \in \mathcal{V}_i^{\mathcal{B}}(\mathcal{C})$, then they would both be in $\mathcal{V}_i(\mathcal{C})$ and thus connected by a walk that lies completely within $\mathcal{V}_i(\mathcal{C}) \subseteq \mathcal{V} \setminus \mathcal{C}$, a contradiction to the assumption that $\mathcal{C}$ separates $\mathcal{A}$ from $\mathcal{B}$.

Thus, we can write

$$\mathcal{A} = \bigcup_{i \in I_1} \mathcal{V}_i^{\mathcal{A}}(\mathcal{C}), \quad \mathcal{B} = \bigcup_{i \in I_2} \mathcal{V}_i^{\mathcal{B}}(\mathcal{C}) \tag{16}$$

with $I_1 \cap I_2 = \emptyset$. Since $X_1, \dots, X_n$ satisfy the cutset property and $\mathcal{C}$ is a cutset, we obtain the FCMI $\coprod_{i=1}^{s(\mathcal{C})} X_{\mathcal{V}_i(\mathcal{C})} \mid X_{\mathcal{C}}$, from which, by the equivalence of parts 1 and 4 in Proposition 3.2, we obtain

$$X_{\left[\bigcup_{i \in I_1} \mathcal{V}_i(\mathcal{C})\right]} \perp\!\!\!\perp X_{\left[\bigcup_{i \in I_2} \mathcal{V}_i(\mathcal{C})\right]} \mid X_{\mathcal{C}}. \tag{17}$$

Equation (16) and decomposition (S3) applied to both the left and right side of Equation (17) implies $X_{\mathcal{A}} \perp\!\!\!\perp X_{\mathcal{B}} \mid X_{\mathcal{C}}$ and thus the global Markov property. Therefore, $X_1, \dots, X_n$ form a Markov random field with respect to $\perp\!\!\!\perp$ and $\mathcal{G}$.

For the other direction, assume the global Markov property holds. Let $\mathcal{U} \subseteq \mathcal{V}$ be a cutset and let $\mathcal{V}_1(\mathcal{U}), \dots, \mathcal{V}_{s(\mathcal{U})}(\mathcal{U})$ be the components of $\mathcal{G}^{\setminus \mathcal{U}}$. We know that $\mathcal{U}$ separates $\mathcal{V}_1(\mathcal{U})$ from $\mathcal{V}_2(\mathcal{U})$,

which by the global Markov property implies $X_{\mathcal{V}_1(\mathcal{U})} \perp\!\!\!\perp X_{\mathcal{V}_2(\mathcal{U})} \mid X_{\mathcal{U}}$. This can be interpreted as the conditional mutual independence $\coprod_{i=1}^{2} X_{\mathcal{V}_i(\mathcal{U})} \mid X_{\mathcal{U}}$. Assume by induction that we know

$$\coprod_{i=1}^{m} X_{\mathcal{V}_i(\mathcal{U})} \mid X_{\mathcal{U}}. \tag{18}$$

Note that $\mathcal{U}$ also separates $\bigcup_{i=1}^{m} \mathcal{V}_i(\mathcal{U})$ from $\mathcal{V}_{m+1}(\mathcal{U})$, which by the global Markov property implies

$$X_{\mathcal{V}_{m+1}(\mathcal{U})} \perp\!\!\!\perp \prod_{i=1}^{m} X_{\mathcal{V}_i(\mathcal{U})} \mid X_{\mathcal{U}}.$$

This, together with Equation (18) and the equivalence of parts 1 and 3 in Proposition 3.2, implies $\coprod_{i=1}^{m+1} X_{\mathcal{V}_i(\mathcal{U})} \mid X_{\mathcal{U}}$. By induction, this shows the FCMI

$$\coprod_{i=1}^{s(\mathcal{U})} X_{\mathcal{V}_i(\mathcal{U})} \mid X_{\mathcal{U}}.$$

Overall, this shows that $X_1, \dots, X_n$ satisfies the cutset property with respect to $\perp\!\!\!\perp$ and $\mathcal{G}$. □

## 3.3 Markov Chains in Separoids

Again, we fix a general separoid $(M, \perp\!\!\!\perp)$. Recall the definition of a Markov chain from Definition 2.16.

**Proposition 3.6** (Characterization of Markov Chains)**.** *Let $X_1, \dots, X_n \in M$. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be the graph with $\mathcal{V} = [n]$ and $\mathcal{E} = \big\{\{i, i+1\} \mid i = 1, \dots, n-1\big\}$. The following properties are equivalent:*

1. *$X_1 \to X_2 \to \cdots \to X_n$, i.e., the sequence forms a Markov chain;*
2. *$X_{[i-1]} \to X_i \to X_{[i+1:n]}$ for all $i = 1, \dots, n-1$;*
3. *$X_1, \dots, X_n$ form a Markov random field with respect to $\mathcal{G}$ and $\perp\!\!\!\perp$.*

*Proof.* Assume 1. To prove 2, let $i \in \{1, \dots, n-1\}$. We need to show the independence

$$X_{[i+1:n]} \perp\!\!\!\perp X_{[i-1]} \mid X_i. \tag{19}$$

We already know that the independence $X_{i+1} \perp\!\!\!\perp X_{[i-1]} \mid X_i$ holds. Assume by induction that

$$X_{[i+1:l]} \perp\!\!\!\perp X_{[i-1]} \mid X_i \tag{20}$$

for some $l \geq i+1$. If $l = n$, then we are done. Otherwise, note that the independence $X_{l+1} \perp\!\!\!\perp X_{[l-1]} \mid X_l$ gives us, by weak union (S4), the property

$$X_{l+1} \perp\!\!\!\perp X_{[i-1]} \mid X_i X_{[i+1:l]}. \tag{21}$$

Contraction (S5) applied to Equations (20) and (21) results in $X_{[i+1:l+1]} \perp\!\!\!\perp X_{[i-1]} \mid X_i$. By induction, we obtain the result, Equation (19).

Now assume 2. To prove 3, by Proposition 3.5 it is enough to show the cutset property. For this, let $\mathcal{U} \subseteq \mathcal{G}$ a cutset and $\mathcal{V}_1(\mathcal{U}), \dots, \mathcal{V}_{s(\mathcal{U})}(\mathcal{U})$ the corresponding components in $\mathcal{G}^{\setminus \mathcal{U}}$. A component in $\mathcal{G}$ necessarily consists of consecutive elements, and we can thus assume that they are ordered in such a way that $v_1 < \cdots < v_{s(\mathcal{U})}$ for all vertices $v_1 \in \mathcal{V}_1(\mathcal{U}), \dots, v_{s(U)} \in \mathcal{V}_{s(\mathcal{U})}(\mathcal{U})$.

Now, let $i \in \{2, \dots, s(\mathcal{U})\}$ be arbitrary. Let $u \in \mathcal{U}$ be any element such that $v < u < w$ for all $v \in \bigcup_{k=1}^{i-1} \mathcal{V}_k(\mathcal{U})$ and $w \in \bigcup_{k=i}^{s(\mathcal{U})} \mathcal{V}_k(\mathcal{U})$; this clearly exists since otherwise we could merge the components with indices $i-1$ and $i$. Let $\mathcal{U}_{u-}$ and $\mathcal{U}_{u+}$ be the sets of elements of $\mathcal{U}$ that are smaller and larger than $u$, respectively. Then by 2, we obtain

$$\big(X_{\mathcal{V}_i(\mathcal{U})} \cdots X_{\mathcal{V}_{s(\mathcal{U})}(\mathcal{U})} X_{\mathcal{U}_{u+}}\big) \perp\!\!\!\perp \big(X_{\mathcal{V}_1(\mathcal{U})} \cdots X_{\mathcal{V}_{i-1}(\mathcal{U})} X_{\mathcal{U}_{u-}}\big) \mid X_u.$$

With weak union (S4) applied to both sides, we obtain

$$\left(X_{\mathcal{V}_i(\mathcal{U})} \cdots X_{\mathcal{V}_{s(\mathcal{U})}(\mathcal{U})}\right) \perp\!\!\!\perp \left(X_{\mathcal{V}_1(\mathcal{U})} \cdots X_{\mathcal{V}_{i-1}(\mathcal{U})}\right) \mid X_{\mathcal{U}}.$$

Decomposition (S3) applied to the left side gives

$$X_{\mathcal{V}_i(\mathcal{U})} \perp\!\!\!\perp \left(X_{\mathcal{V}_1(\mathcal{U})} \cdots X_{\mathcal{V}_{i-1}(\mathcal{U})}\right) \mid X_{\mathcal{U}}.$$

Since $i$ was arbitrary, by the equivalence of 1 and 2 in Proposition 3.2, we obtain the FCMI $\perp\!\!\!\perp_{i=1}^{s(\mathcal{U})} X_{\mathcal{V}_i(\mathcal{U})} \mid X_{\mathcal{U}}$, showing the cutset property and thus 3.

Since $\{i-1\}$ separates $\{i\}$ from $\{1, \dots, i-2\}$ in $\mathcal{G}$, the global Markov property shows that 3 implies 1. □

# 4 Characterizing $F$-Independences and $F$-Markov Random Fields

In this section we generalize the work Yeung et al. [2002] on information characterizations of (full) conditional mutual independences and Markov random fields from Shannon entropy to general $F$. The reader may also compare with Yeung [2008], Chapter 12, which contains the same content as Yeung et al. [2002] with more explanations.

In Section 4.1, we state a main technique used in our proofs, Theorem 4.1, which we term *subset determination*. It states that the value of a region in an $F$-diagram always fully determines the value of all subregions, including the atomic parts. This result will replace the use of inequalities in Yeung et al. [2002]. In Section 4.2, we then show that the $F$-independence defined in Definition 2.19 satisfies the separoid axioms, and so all results from Section 3 apply.

In Section 4.3, we define conditional mutual $F$-independences and show in Theorem 4.15 a characterization by the vanishing of a conditional $F$–dual total correlation and its corresponding atoms. In Section 4.4, with Theorem 4.21, we then generalize this to a characterization of full conditional mutual $F$-independences. By Proposition 3.5, Markov random fields in a separoid are equivalently characterized by the cutset property, and thus by a set of full conditional mutual independences. Thus, the previous results lead to the proof for the characterization of $F$-Markov random fields in terms of the $F$-diagram, as stated in Theorem 2.21. We also specialize this to $F$-Markov chains. Finally, in Appendix C, we briefly look at Yeung et al. [2019], which builds on Yeung et al. [2002], and explain which of the results transfer to our generalized setting. We put this into the appendix since no later results build on this.

Let in this whole section $M$ be a commutative, idempotent monoid acting additively on an abelian group $G$ and $F : M \to G$ a function satisfying the chain rule Equation (5).

## 4.1 Subset Determination

The following theorem, which did not appear in Lang et al. [2025], highlights a property that we call *subset determination*. This crucial property lies at the heart of the proofs of the main results in this work.

**Theorem 4.1** (Subset Determination)**.** *Let $M$ be a commutative, idempotent monoid acting additively on an abelian group $G$ and $F : M \to G$ be a function satisfying the chain rule Equation (5). Fix elements $X_1, \dots, X_n \in M$ and let $\widetilde{X} := \widetilde{X}(n)$ be defined as in Equation (3), resulting in the $G$-valued measure $\widetilde{F} : 2^{\widetilde{X}} \to G$ from Theorem 2.8.*

*Then for any $A \subseteq \widetilde{X}$ and any atom $p_I \in A$, one has*

$$\widetilde{F}(p_I) = \sum_{K \subseteq I} (-1)^{|K|-|I|} \cdot X_{[n]\setminus K}.\widetilde{F}(A).$$

*In particular, if $\widetilde{F}(A) = 0$, then $\widetilde{F}(p_I) = 0$ for all atoms $p_I \in A$, and consequently $\widetilde{F}(B) = 0$ for all $B \subseteq A$.*

**Remark 4.2.** *It is important to note that this theorem is* not *true if we work in the setting of information functions that apply to probability mass functions and restrict our attention to a fixed*

*probability mass function $P$. For example, the total entropy $I(XY;P)$ of a joint variable $XY$ does not determine the mutual information $I(X;Y;P)$ between the two, even though $I(XY)$ does determine $I(X;Y)$. Nevertheless, we are able to apply our results also to fixed probability mass functions, as we explain in Appendix D on* slices *of $I$-diagrams.*

We first need more notation and several lemmas:

**Notation 4.3.** *For all $\emptyset \neq I = \{i_1 < \cdots < i_q\} \subseteq [n]$, we write*

$$F(\,;_{i\in I} X_i) \coloneqq F(X_{i_1};\ldots;X_{i_q}).$$

**Lemma 4.4.** *For all $\emptyset \neq I \subseteq [n]$, one has*

$$\widetilde{F}(p_I) = X_{[n]\setminus I}.F(\,;_{i\in I} X_i).$$

*Proof.* This follows directly from Equation (4) and Hu's Theorem. □

**Lemma 4.5.** *Assume $p_L$ is an atom and $K \subseteq [n]$. Then*

$$X_{[n]\setminus K}.\widetilde{F}(p_L) = \begin{cases} \widetilde{F}(p_L), & L \subseteq K, \\ 0, & \text{else.} \end{cases}$$

*Proof.* By Lemma 4.4, we have

$$X_{[n]\setminus K}.\widetilde{F}(p_L) = X_{[n]\setminus K}.\Big(X_{[n]\setminus L}.F(\,;_{l\in L} X_l)\Big) = X_{[n]\setminus(L\cap K)}.F(\,;_{l\in L} X_l).$$

From this, we immediately see the result for the case $L \subseteq K$. If $L \nsubseteq K$, then there is $l \in L \setminus K$. Then $l \in [n] \setminus (L \cap K)$ and consequently, with $Y \coloneqq X_{([n]\setminus l)\setminus(L\cap K)}$:

$$X_{[n]\setminus K}.\widetilde{F}(p_L) = Y.\Big(X_l.F(\,;_{l'\in L} X_{l'})\Big) = Y.\widetilde{F}\left(\bigcap_{l'\in L} \widetilde{X}_{l'} \setminus \widetilde{X}_l\right) = Y.\widetilde{F}(\emptyset) = 0,$$

where we have used Hu's Theorem 2.8 in the second step. That was to show. □

**Lemma 4.6.** *Let $L \subseteq I$ be two sets. Then*

$$\sum_{K:\ L\subseteq K\subseteq I} (-1)^{|K|} = (-1)^{|L|} \cdot \mathbb{1}_{I=L}.$$

*Proof.* We have:

$$\begin{aligned}
\sum_{K:\ L\subseteq K\subseteq I} (-1)^{|K|} &= \sum_{k=|L|}^{|I|} (-1)^k \cdot \Big|\Big\{K \mid L \subseteq K \subseteq I,\ |K| = k\Big\}\Big| \\
&= \sum_{k=|L|}^{|I|} (-1)^k \cdot \binom{|I|-|L|}{k-|L|} \\
&= (-1)^{|L|} \cdot \sum_{k=0}^{|I|-|L|} \binom{|I|-|L|}{k} \cdot 1^{|I|-|L|-k} \cdot (-1)^k \\
&\overset{(\star)}{=} (-1)^{|L|} \cdot (1-1)^{|I|-|L|} \\
&= (-1)^{|L|} \cdot \mathbb{1}_{I=L}.
\end{aligned}$$

In $(\star)$, we used the well-known binomial theorem and in the final step that $0^0 = 1$.[6] □

[6] If one is not comfortable with the definition $0^0 = 1$, one can also directly verify the overall result in the case $L = I$.

*Proof of Theorem 4.1.* We have

$$\begin{aligned}
\sum_{K\subseteq I}(-1)^{|K|-|I|}\cdot X_{[n]\setminus K}.\widetilde{F}(A) &= \sum_{K\subseteq I}(-1)^{|K|-|I|}\cdot X_{[n]\setminus K}.\left(\sum_{p_L\in A}\widetilde{F}(p_L)\right)\\
&= \sum_{p_L\in A}(-1)^{-|I|}\sum_{K\subseteq I}(-1)^{|K|}\cdot X_{[n]\setminus K}.\widetilde{F}(p_L)\\
&= \sum_{p_L\in A}(-1)^{-|I|}\cdot\left(\sum_{K\colon\, L\subseteq K\subseteq I}(-1)^{|K|}\right)\cdot\widetilde{F}(p_L) && \text{(Lemma 4.5)}\\
&= \sum_{p_L\in A}(-1)^{-|I|}\cdot(-1)^{|L|}\cdot\mathbb{1}_{L=I}\cdot\widetilde{F}(p_L) && \text{(Lemma 4.6)}\\
&= (-1)^{-|I|}\cdot(-1)^{|I|}\cdot\widetilde{F}(p_I) && (p_I\in A)\\
&= \widetilde{F}(p_I). && \square
\end{aligned}$$

**Remark 4.7.** *For the special case that $A=\widetilde{X}$, the theorem says that $F\big(X_{[n]}\big)$ determines $\widetilde{F}(p_I)$ for all atoms $p_I\in\widetilde{X}$. If $M$ is generated by the $X_1,\dots,X_n$, this means that $F\big(X_{[n]}\big)$ entirely determines $F$. In Appendix A, we generalize this observation and show that if $\top\in M$ is any so-called* top element*, then $F(\top)$ determines $F$ entirely. This leads to a one-to-one correspondence between elements of $G$ that are* annihilated *by $\top$ and functions $F:M\to G$ satisfying the chain rule. We also interpret this result as the vanishing of a cohomology group of degree one.*

## 4.2 $F$-Independence Satisfies the Separoid Axioms

Let $M$ again be a commutative, idempotent monoid acting on an abelian group $G$, and $F:M\to G$ a function satisfying the chain rule Equation (5). We now show that the $F$-independence introduced in Definition 2.19 satisfies the separoid axioms.

Our proof of the following proposition, which shows that $(M,\perp\!\!\!\perp_F)$ is a separoid, makes extensive use of Hu's Theorem 2.8 and the subset determination property, Theorem 4.1.

**Proposition 4.8.** *Let $M$ be a commutative, idempotent monoid acting on the abelian group $G$ and $F:M\to G$ a function satisfying the chain rule Equation (5). Let $\perp\!\!\!\perp_F$ be the $F$-independence relation on $M$ from Definition 2.19. Then $(M,\perp\!\!\!\perp_F)$ is a separoid.*

*Proof.* We need to prove the five separoid axioms from Definition 2.12. In proving each axiom, we will use the elements of $M$ appearing in the formulas as the elements $X_1,\dots,X_n$ in Hu's Theorem 2.8.

1. Symmetry: for any $X,Y,Z\in M$, we have

   $$Z.F(X;Y)=\widetilde{F}\big(\widetilde{X}\cap\widetilde{Y}\setminus\widetilde{Z}\big)=\widetilde{F}\big(\widetilde{Y}\cap\widetilde{X}\setminus\widetilde{Z}\big)=Z.F(Y;X),$$

   showing symmetry $X\underset{F}{\perp\!\!\!\perp}Y\mid Z\implies Y\underset{F}{\perp\!\!\!\perp}X\mid Z$.

2. Redundancy: assume $W\precsim Z$, i.e., $WZ=Z$. It follows

   $$\widetilde{F}\big(\widetilde{W}\setminus\widetilde{Z}\big)=Z.F(W)\overset{\text{Eq. (5)}}{=}F(ZW)-F(Z)=F(Z)-F(Z)=0.$$

   Since $\widetilde{W}\cap\widetilde{Y}\setminus\widetilde{Z}\subseteq\widetilde{W}\setminus\widetilde{Z}$, subset determination Theorem 4.1 shows

   $$Z.F(W;Y)=\widetilde{F}\big(\widetilde{W}\cap\widetilde{Y}\setminus\widetilde{Z}\big)=0$$

   and thus $W\underset{F}{\perp\!\!\!\perp}Y\mid Z$.

3. Decomposition and weak union: assume $WX\underset{F}{\perp\!\!\!\perp}Y\mid Z$. Then we have

   $$\widetilde{F}\big(\widetilde{WX}\cap\widetilde{Y}\setminus\widetilde{Z}\big)=Z.F(WX;Y)=0.$$

Note that

$$\widetilde{X} \cap \widetilde{Y} \setminus \widetilde{Z} \subseteq \widetilde{WX} \cap \widetilde{Y} \setminus \widetilde{Z} \supseteq \widetilde{W} \cap \widetilde{Y} \setminus \widetilde{XZ}.$$

Thus, subset determination Theorem 4.1 shows

$$\begin{aligned} Z.F(X;Y) &= \widetilde{F}\big(\widetilde{X} \cap \widetilde{Y} \setminus \widetilde{Z}\big) = 0, \\ (XZ).F(W;Y) &= \widetilde{F}\big(\widetilde{W} \cap \widetilde{Y} \setminus \widetilde{XZ}\big) = 0, \end{aligned}$$

which shows $X \underset{F}{\perp\!\!\!\perp} Y \mid Z$ and $W \underset{F}{\perp\!\!\!\perp} Y \mid XZ$ and thus both decomposition and weak union.

4. Contraction: assume $W \underset{F}{\perp\!\!\!\perp} Y \mid XZ$ and $X \underset{F}{\perp\!\!\!\perp} Y \mid Z$. We obtain from Hu's Theorem:

$$\begin{aligned} Z.F(WX;Y) &= \widetilde{F}\big(\widetilde{WX} \cap \widetilde{Y} \setminus \widetilde{Z}\big) \\ &= \widetilde{F}\big(\widetilde{W} \cap \widetilde{Y} \setminus \widetilde{XZ}\big) + \widetilde{F}\big(\widetilde{X} \cap \widetilde{Y} \setminus \widetilde{Z}\big) \\ &= (XZ).F(W;Y) + Z.F(X;Y) \\ &= 0 + 0 = 0. \end{aligned}$$

Thus, we have proven $WX \underset{F}{\perp\!\!\!\perp} Y \mid Z$. In the calculation, the second step uses that $\widetilde{F}$ is additive over disjoint unions.

□

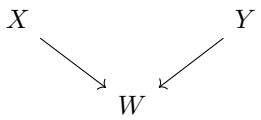
In the following proposition, we show that $F$-independences are preserved under conditioning:

**Proposition 4.9.** *Let $(M, \perp\!\!\!\perp_F)$ be the separoid from above. Let $X, Y, Z, W \in M$. Then the following implication holds:*

$$X \underset{F}{\perp\!\!\!\perp} Y \mid Z \quad \Longrightarrow \quad X \underset{F}{\perp\!\!\!\perp} Y \mid WZ.$$

*Proof.* The claim follows from subset determination Theorem 4.1 and Hu's Theorem 2.8 by using that $\widetilde{X} \cap \widetilde{Y} \setminus \widetilde{WZ} \subseteq \widetilde{X} \cap \widetilde{Y} \setminus \widetilde{Z}$. □

**Remark 4.10.** *Note that the preceding proposition is incorrect when working in the probabilistic setting and fixing the probability mass function. For example, the case of a collider*




*in a Bayesian network shows that it is possible that the probabilistic independence $X \perp\!\!\!\perp_P Y$ is true, while $X \perp\!\!\!\perp_P Y \mid W$ is not. Nevertheless, we will be able to apply our theory to the probabilistic case in Sections 5 and D.*

## 4.3 Conditional Mutual $F$-Independences and $F$-Dual Total Correlation

The following definition specializes conditional mutual independences, Definition 3.1, to the case that the independence relation is given by $\perp\!\!\!\perp_F$:

**Definition 4.11** (Conditional Mutual $F$-Independence)**.** *Let $X_1, \dots, X_n, Y \in M$. If $X_1, \dots, X_n \in M$ are mutually independent given $Y \in M$ with respect to $\perp\!\!\!\perp_F$, then we write*

$$F : \prod_{i=1}^{n} X_i \mid Y.$$

*We call this a* conditional mutual $F$-independence*, and we say $X_1, \dots, X_n$ are mutually $F$-independent given $Y$.*

As in Proposition 4.9, we get the following peculiar implication which we will later use:

**Proposition 4.12.** *Let $X_1, \dots, X_n, Y, W \in M$. Then we have the following implication:*

$$F : \coprod_{i=1}^{n} X_i \mid Y \quad \Longrightarrow \quad F : \coprod_{i=1}^{n} X_i \mid WY.$$

*Proof.* This follows immediately from Proposition 4.9 using that conditional mutual $F$-independences are defined as a combination of conditional pairwise $F$-independences. □

As in Remark 4.10, we mention that the preceding proposition does not hold in the probabilistic setting when *fixing* the underlying probability mass function.

### Warmup: Characterizing Pairwise Conditional $F$-Independence

**Proposition 4.13.** *Let $X_1, X_2, Y \in M$. Then the following are equivalent:*

1. $X_1 \underset{F}{\perp\!\!\!\perp} X_2 \mid Y$;
2. $Y.F(X_1X_2) = Y.F(X_1) + Y.F(X_2)$;
3. $Y.F(X_1X_2) = (YX_2).F(X_1) + (YX_1).F(X_2)$;
4. $Y.F(X_1) = (YX_2).F(X_1)$;
5. $Y.F(X_2) = (YX_1).F(X_2)$.

*Proof.* Using Hu's Theorem 2.8 with elements $X_1, X_2, Y$, we generally have

$$\begin{aligned} Y.F(X_1X_2) &= \widetilde{F}\Big(\widetilde{X}_1 \cup \widetilde{X}_2 \setminus \widetilde{Y}\Big) \\ &= \widetilde{F}\Big(\widetilde{X}_1 \setminus \widetilde{Y}\Big) + \widetilde{F}\Big(\widetilde{X}_2 \setminus \widetilde{Y}\Big) - \widetilde{F}\Big(\widetilde{X}_1 \cap \widetilde{X}_2 \setminus \widetilde{Y}\Big) \\ &= Y.F(X_1) + Y.F(X_2) - Y.F(X_1; X_2). \end{aligned}$$

This shows the equivalence of 1 and 2. Similarly, one can show the decomposition

$$Y.F(X_1X_2) = (YX_2).F(X_1) + (YX_1).F(X_2) + Y.F(X_1; X_2),$$

which shows the equivalence of 1 and 3. Finally, we similarly obtain decompositions

$$\begin{aligned} Y.F(X_1) &= (YX_2).F(X_1) + Y.F(X_1; X_2); \\ Y.F(X_2) &= (YX_1).F(X_2) + Y.F(X_1; X_2). \end{aligned}$$

The first decomposition shows the equivalence of 1 and 4, and the second the one of 1 and 5. □

Of interest to us are especially properties 2 and 3, which can equivalently be expressed as the following vanishing conditions:

$$\begin{aligned} Y.\Big[F(X_1) + F(X_2) - F(X_1X_2)\Big] &= 0; \\ Y.\Big[F(X_1X_2) - \big(X_2.F(X_1) + X_1.F(X_2)\big)\Big] &= 0. \end{aligned}$$

The concepts of $F$–total correlation and $F$–dual total correlations provide natural generalizations of the quantities at the left-hand-sides of these conditions. These generalize total correlation [Watanabe, 1960] and dual total correlation [Han, 1978] by replacing Shannon entropy $I$ in the defining expressions by $F$:

**Definition 4.14** ($F$–(Dual )Total Correlation)**.** *Let $X_1, \dots, X_n \in M$. Then their* $F$–total correlation *is given by*

$$TC_F(X_1; \dots; X_n) := \sum_{i=1}^{n} F(X_i) - F\big(X_{[n]}\big).$$

*Similarly, the* $F$–dual total correlation *is given by*

$$DTC_F(X_1;\dots;X_n) \coloneqq F\big(X_{[n]}\big) - \sum_{i=1}^{n} X_{[n]\setminus i}.F(X_i),$$

*where* $[n] \setminus i \coloneqq [n] \setminus \{i\}$. *If* $I = \{i_1 < \dots < i_q\} \subseteq [n]$, *then we also write*

$$TC_F\big(\,;_{i\in I} X_i\big) \coloneqq TC_F(X_{i_1};\dots;X_{i_q}),$$
$$DTC_F\big(\,;_{i\in I} X_i\big) \coloneqq DTC_F(X_{i_1};\dots;X_{i_q}).$$

Similarly to Proposition 4.13, we want to use $F$–total correlation and $F$–dual total correlation to characterize conditional *mutual* $F$-independences. We will focus on the case of $F$–dual total correlation, which is slightly easier and works in full generality. For the interested reader, we consider the case of $F$–total correlation in Appendix B. This only provides a valid characterization in the case that the group $G$ is *torsion-free*, as Example B.5 will demonstrate.

### Characterization using $F$-Dual Total Correlation

Let $X_1,\dots,X_n \in M$. For $I = \{i_1 < \dots < i_q\} \subseteq [n]$, we can then consider the interaction term $F(X_{i_1};\dots;X_{i_q})$. Recall the notation $F\big(\,;_{i\in I} X_i\big) \coloneqq F\big(X_{i_1};\dots;X_{i_q}\big)$.

**Theorem 4.15.** *Let* $M$ *be a commutative, idempotent monoid acting additively on an abelian group* $G$, *and* $F : M \to G$ *a function satisfying the chain rule Equation* (5). *Let* $X_1,\dots,X_n,Y \in M$. *Then the following properties are equivalent:*

1. $F : \coprod_{i=1}^{n} X_i \mid Y$;
2. $Y.DTC_F\big(\,;_{i\in[n]} X_i\big) = 0$;
3. $\big(YX_{[n]\setminus I}\big).F\big(\,;_{i\in I} X_i\big) = 0$ *for all* $I \subseteq [n]$ *with* $|I| \geq 2$;
4. $Y.F\big(X_i; X_{[n]\setminus i}\big) = 0$ *for all* $i = 1,\dots,n$.

*Proof.* Assume 1. We prove 2 by induction over $n$, with the case $n = 2$ corresponding to the equivalence of 1 and 3 in Proposition 4.13. Let $n \geq 3$. By Proposition 3.2 we have

$$F : \coprod_{i=1}^{n-1} X_i \mid Y, \quad X_n \underset{F}{\perp\!\!\!\perp} X_{[n-1]} \mid Y.$$

The first $F$-independence implies by Proposition 4.12 the following: $F : \coprod_{i=1}^{n-1} X_i \mid YX_n$. By induction, first using the pairwise and then the mutual case,[7] we then obtain:

$$\begin{aligned} Y.F\big(X_{[n]}\big) &= \big(YX_n\big).F\big(X_{[n-1]}\big) + \big(YX_{[n-1]}\big).F(X_n) \\ &= \sum_{i=1}^{n-1} \big(YX_nX_{[n-1]\setminus i}\big).F(X_i) + \big(YX_{[n-1]}\big).F(X_n) \\ &= Y.\Big(\sum_{i=1}^{n} X_{[n]\setminus i}.F(X_i)\Big). \end{aligned}$$

That shows 2.

For the rest of the proof, write $X_{n+1} \coloneqq Y$ and assume that $X_1,\dots,X_{n+1}$ are the elements in Hu's Theorem 2.8.

[7] Similarly to Proposition 4.13, we write the vanishing of the $F$–dual total correlation as a different equality by bringing a term to the other side.

Now assume 2. Note that

$$\begin{aligned} 0 = Y.DTC_F(\,;_{i\in[n]}X_i) &= X_{n+1}.F(X_{[n]}) - \sum_{i=1}^{n} X_{[n+1]\setminus i}.F(X_i) \\ &= \widetilde{F}\Big(\widetilde{X}_{[n]} \setminus \widetilde{X}_{n+1}\Big) - \sum_{i=1}^{n} \widetilde{F}(p_i) \\ &= \widetilde{F}\Big(\big\{p_I \mid I \subseteq [n], |I| \geq 2\big\}\Big), \end{aligned}$$

where we used Lemma 4.4 in the second step. Let $I \subseteq [n]$ with $|I| \geq 2$. Then, subset determination Theorem 4.1 and the same corollary again imply:

$$0 = \widetilde{F}(p_I) = X_{[n+1]\setminus I}.F(\,;_{i\in I}X_i) = (YX_{[n]\setminus I}).F(\,;_{i\in I}X_i).$$

That is precisely 3.

Now, assume 3. To prove 4, for symmetry reasons it is enough to show that $Y.F\big(X_n, X_{[n-1]}\big) = 0$. We have

$$\begin{aligned} Y.F\big(X_n; X_{[n-1]}\big) &= X_{n+1}.F\big(X_n; X_{[n-1]}\big) \\ &= \widetilde{F}\Big(\widetilde{X}_n \cap \widetilde{X}_{[n-1]} \setminus \widetilde{X}_{n+1}\Big) \\ &= \sum_{\substack{I\subseteq[n]:\ n\in I,\\ I\cap[n-1]\neq\emptyset}} \widetilde{F}(p_I) \\ &= \sum_{\substack{I\subseteq[n]:\ n\in I,\\ I\cap[n-1]\neq\emptyset}} X_{[n+1]\setminus I}.F(\,;_{i\in I}X_i) \\ &= \sum_{\substack{I\subseteq[n]:\ n\in I,\\ I\cap[n-1]\neq\emptyset}} (YX_{[n]\setminus I}).F(\,;_{i\in I}X_i) \\ &= 0, \end{aligned}$$

where the fourth step used Lemma 4.4. The last step follows since for all $I$ over which we sum, we necessarily have $|I| \geq 2$.

Assuming 4, 1 follows immediately by the definitions of conditional mutual and pairwise $F$-independences. □

## 4.4 Full Conditional Mutual $F$-Independences

We now build on Section 4.3. We will consider mutual $F$-independences of variables that are *themselves* products of several variables. The main result of this section, Theorem 4.21, will generalize the equivalence between properties 1 and 3 in Theorem 4.15.

Fix $n \geq 0$ and $X_1, \ldots, X_n \in M$. This gives rise to a set $\widetilde{X} = \widetilde{X}(n)$ of $2^n - 1$ atoms according to Equation (3) and a $G$-valued measure $\widetilde{F} : 2^{\widetilde{X}} \to G$ according to Hu's Theorem 2.8.

In the following, if $W_i$ are sets indexed with $i \in I$, then $W_I$ denotes $\bigcup_{i\in I} W_i$. We now define $F$-FCMIs, specializing the notion of FCMIs from Definition 3.3 to the setting with the independence relation given as $\perp\!\!\!\perp_F$. We will, however, first define the notion of a conditional partition, since this will prove valuable when studying the effect of $F$-FCMIs on $F$-diagrams:

**Definition 4.16** (Conditional Partition)**.** *Let $q \geq 1$, $L_i \subseteq [n]$ for all $i \in [q]$ and $J \subseteq [n]$. Set $L := L_{[q]} = \bigcup_{i=1}^{q} L_i$. Assume that the $L_i$ and $J$ are all pairwise disjoint and cover $[n]$. Then the family $K := \big(J, L_i, 1 \leq i \leq q\big)$ is called a* conditional partition *of $[n]$.*[8]

[8] We allow $J$ and the $L_i$ to be empty, different from usual partitions.

**Definition 4.17** (Full Conditional Mutual $F$-Independence ($F$-FCMI)). *Let $K = \big(J, L_i, 1 \leq i \leq q\big)$ be a conditional partition with $q \geq 2$. Then an FCMI of the form*

$$F : \coprod_{i=1}^{q} X_{L_i} \mid X_J \tag{22}$$

*is called a full conditional mutual $F$-independence (with respect to the previously fixed elements $X_1, \dots, X_n$) — or $F$-FCMI for short. If this $F$-FCMI holds, then we also say that the conditional partition $K = \big(J, L_i, 1 \leq i \leq q\big)$* induces *an $F$-FCMI (with respect to the previously fixed elements $X_1, \dots, X_n$).*

**Definition 4.18** (Image of a Conditional Partition). *Let $K = \big(J, L_i, 1 \leq i \leq q\big)$ be a conditional partition of $[n]$ with $q \geq 2$. Then its image is defined as*

$$\mathrm{Im}(K) \coloneqq \left\{ p_W \in \widetilde{X}(n) \;\middle|\; \exists I \subseteq [q], |I| \geq 2, \forall i \in I \;\; \exists \emptyset \neq W_i \subseteq L_i : W = W_I = \bigcup_{i \in I} W_i \right\},$$

*i.e., as a certain set of atoms in $\widetilde{X}(n)$.*

**Lemma 4.19.** *Let $q \geq 1$ and $K = \big(J, L_i, 1 \leq i \leq q\big)$ a conditional partition of $[n]$. Let $\emptyset \neq I \subseteq [q]$. Then the equality*

$$A_I \coloneqq \bigcap_{i \in I} \widetilde{X}_{L_i} \;\setminus\; \widetilde{X}_{J \cup L \setminus L_I} = \left\{ p_W \in \widetilde{X}(n) \;\middle|\; \forall i \in I \;\exists \emptyset \neq W_i \subseteq L_i : W = W_I = \bigcup_{i \in I} W_i \right\}$$

*holds. Furthermore, if $q \geq 2$, then we have $\mathrm{Im}(K) = \bigcup_{I \subseteq [q]:\ |I| \geq 2} A_I$.*

*Proof.* Let $p_W$ be an atom, where by definition $\emptyset \neq W \subseteq [n] = J \cup L$. Then we have

$$\begin{aligned} p_W \in A_I \quad &\iff \quad \forall i \in I \;\; \exists l \in L_i : p_W \in \widetilde{X}_l \quad \wedge \quad \forall l \in J \cup L \setminus L_I : p_W \notin \widetilde{X}_l \\ &\iff \quad \forall i \in I : \emptyset \neq W \cap L_i \quad \wedge \quad W \cap \big(J \cup L \setminus L_I\big) = \emptyset. \end{aligned}$$

Now, assume $p_W \in A_I$ and let $W_i \coloneqq W \cap L_i \neq \emptyset$ for $i \in I$. Then from the preceding characterization, we obtain:

$$W = W \cap (J \cup L) = W \cap L_I = W \cap \bigcup_{i \in I} L_i = \bigcup_{i \in I} W_i = W_I$$

and thus $p_W = p_{W_I}$ is in the set at the right-hand-side.

If, vice versa, $p_W = p_{W_I}$ is of the stated form of the set at the right-hand-side, then $W \cap L_i = W_i \neq \emptyset$ for all $i \in I$ and $W \cap \big(J \cup L \setminus L_I\big) = \emptyset$, and thus $p_W \in A_I$.

The last statement follows immediately from the definition of $\mathrm{Im}(K)$ and what we just showed. □

Before we state the following proposition, we remind of the fact that, by Lemma 4.4, we have

$$F\big(X_{[n]}\big) = \widetilde{F}\big(\widetilde{X}\big) = \sum_{p_I \in \widetilde{X}} \widetilde{F}(p_I) = \sum_{\emptyset \neq I \subseteq [n]} X_{[n] \setminus I} . F\big(\,;_{i \in I} X_i\big).$$

The following proposition generalizes this to the case that the left-hand-side is itself a "term of higher degree":

**Proposition 4.20.** *Let $q \geq 1$, and let $L_1, \dots, L_q \subseteq [n]$ be pairwise disjoint sets. For any $\emptyset \neq I \subseteq [q]$, we have*

$$F\big(\,;_{i \in I} X_{L_i}\big) = \sum_{\substack{(W_i)_{i \in I}: \\ \forall i:\ \emptyset \neq W_i \subseteq L_i}} X_{L_I \setminus W_I} . F\big(\,;_{w \in W_I} X_w\big).^{9}$$

[9] For mitigating confusion, note that each $W_i$ is itself a set, and so each $w \in W_I$ is an element contained in exactly one of the $W_i$.

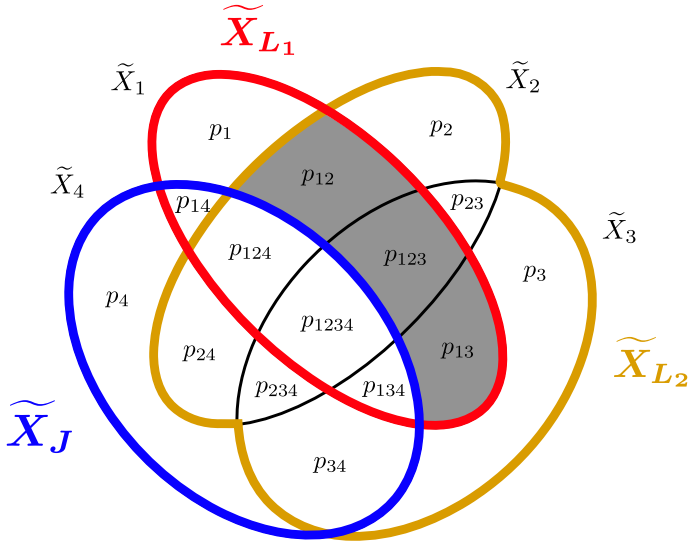



**Figure 6:** In this illustration, we visualize the full conditional mutual $F$-independence ($F$-FCMI) $X_{L_1} \overset{\perp\!\!\!\perp}{}_F X_{L_2} \mid X_J$ for the case $n = 4$, $J = \{4\}$, $L_1 = \{1\}$, and $L_2 = \{2, 3\}$. The independence is by Hu's Theorem 2.8 given by $\widetilde{F}\big(\widetilde{X}_{L_1} \cap \widetilde{X}_{L_2} \setminus \widetilde{X}_J\big) = X_J.F(X_{L_1}; X_{L_2}) = 0$. That is, $\widetilde{F}$ vanishes on the gray region. Subset determination Theorem 4.1 results in $\widetilde{F}$ vanishing even on all the atoms *within* that region — namely $p_{12}, p_{123}$, and $p_{13}$. Defining the conditional partition $K = \big(J, L_1, L_2\big)$, these atoms are precisely the elements in $\mathrm{Im}(K)$, thus confirming the characterization of $F$-FCMIs in terms of $\widetilde{F}$ given in Theorem 4.21.

*Proof.* Assume without loss of generality that $I = [q]$ and $L_I = [n]$. Then, applying Hu's Theorem 2.8 and Lemma 4.19, we obtain:

$$F\big(\,;_{i\in I} X_{L_i}\big) = \widetilde{F}\left(\bigcap_{i\in I} \widetilde{X}_{L_i}\right) = \sum_{\substack{\emptyset \neq W \subseteq L_I: \\ p_W \in \bigcap_{i\in I} \widetilde{X}_{L_i}}} \widetilde{F}(p_W) = \sum_{\substack{(W_i)_{i\in I}: \\ \forall i:\ \emptyset \neq W_i \subseteq L_i}} \widetilde{F}(p_{W_I}).$$

The result follows from $\widetilde{F}(p_{W_I}) = X_{L_I \setminus W_I}.F\big(\,;_{w\in W_I} X_w\big)$, see Lemma 4.4. □

The following Theorem generalizes Theorem 5 of Yeung et al. [2002]. We illustrate it in Figures 6 and 7. The reader may also compare with Figures 3 and 5 from the introduction, which illustrates the effect of simple $F$-FCMIs for the case that $F = I$ and general $F$, respectively. In the first of those figures, the FCMIs come from the graph-structure of a Markov random field.

**Theorem 4.21.** *Let $M$ be a commutative, idempotent monoid acting additively on an abelian group $G$ and $F : M \to G$ a function satisfying the chain rule Equation (5). We assume fixed elements $X_1, \dots, X_n \in M$, giving rise to a $G$-valued measure $\widetilde{F} : 2^{\widetilde{X}} \to G$ according to Hu's Theorem 2.8.*

*Let $K = \big(J, L_i, 1 \leq i \leq q\big)$ be a conditional partition of $[n]$ with $q \geq 2$. Set $L \coloneqq L_{[q]}$. Then the following properties are equivalent:*

1. *$K$ induces an $F$-FCMI with respect to $X_1, \dots, X_n$, i.e. $F : \overset{q}{\underset{i=1}{\perp\!\!\!\perp}} X_{L_i} \mid X_J$;*
2. *for all $I \subseteq [q]$ with $|I| \geq 2$: $\big(X_{J\cup L\setminus L_I}\big).F\big(\,;_{i\in I} X_{L_i}\big) = 0$;*
3. *for all $I \subseteq [q]$ with $|I| \geq 2$ and all $(W_i)_{i\in I}$ with $\emptyset \neq W_i \subseteq L_i$ for all $i \in I$, we have $X_{J\cup L\setminus W_I}.F\big(\,;_{w\in W_I} X_w\big) = 0$;*
4. *$\widetilde{F}(p_W) = 0$ for all $p_W \in \mathrm{Im}(K)$.*

*Proof.* That 1 and 2 are equivalent follows immediately from the equivalence of properties 1 and 3 in Theorem 4.15. In doing so, $q$ replaces $n$, $X_J$ replaces $Y$, and $X_{L_i}$ replaces $X_i$.

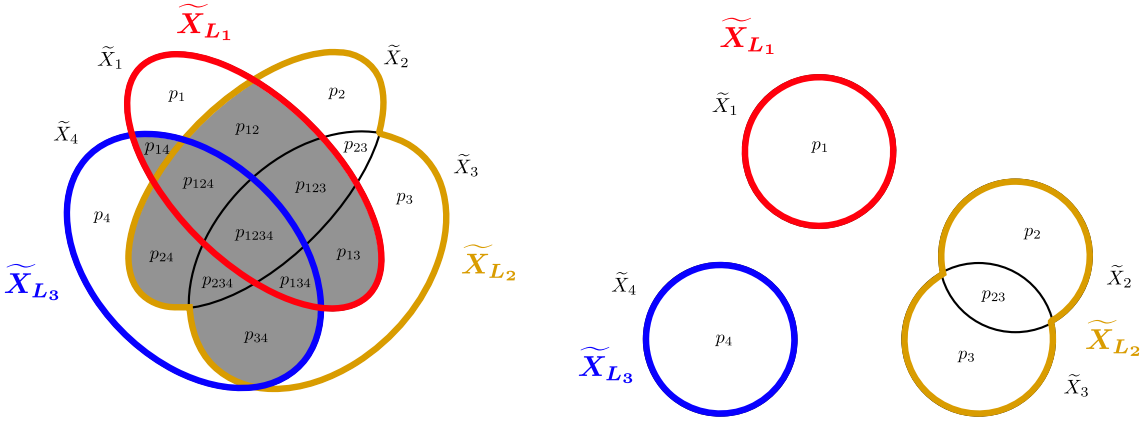



**Figure 7:** This figure visualizes the full conditional mutual $F$-independence ($F$-FCMI) of $X_{L_1}$, $X_{L_2}$ and $X_{L_3}$ for the case $n = 4$, $L_1 = \{1\}$, $L_2 = \{2, 3\}$, and $L_3 = \{4\}$. Here, the conditional variable is trivial, corresponding to $J = \emptyset$. By the characterization of $F$-FCMIs, Theorem 4.21, $\widetilde{F}$ vanishes on all atoms in the gray region. This leads to a degeneracy of the $F$-diagram, which can then be depicted on the small set of atoms at the right-hand-side of the figure.

3 immediately implies 2 by Proposition 4.20 and observing that $X_{J\cup L\setminus L_I} \cdot X_{L_I \setminus W_I} = X_{J\cup L\setminus W_I}$. 3 and 4 are clearly seen to be equivalent using the definition of $\mathrm{Im}(K)$ and the fact that

$$X_{J\cup L\setminus W_I}.F(\,;_{w\in W_I} X_w) = X_{[n]\setminus W_I}.F(\,;_{w\in W_I} X_w) = \widetilde{F}(p_{W_I}).$$

We used $J \cup L = [n]$ in the first step and Lemma 4.4 in the second.

Finally, we need to see that 2 implies 4. Let $p_W \in \mathrm{Im}(K)$. Then by Lemma 4.19, there is a set $I \subseteq [q]$ with $|I| \geq 2$ such that $p_W \in A_I = \bigcap_{i\in I} \widetilde{X}_{L_i} \setminus \widetilde{X}_{J\cup L\setminus L_I}$. Then, using Hu's Theorem 2.8 and property 2, we obtain $\widetilde{F}(A_I) = (X_{J\cup L\setminus L_I}).F(\,;_{i\in I} X_{L_i}) = 0$. By subset determination, Theorem 4.1, this results in particular in $\widetilde{F}(p_W) = 0$, and we are done. □

**Remark 4.22.** *If we replace property 1 in the preceding theorem by simply stating the independence relation — without naming this an $F$-FCMI — then properties 1, 2, and 3 are also equivalent without assuming $L \cup J = [n]$. This corresponds to conditional mutual $F$-independences that are not* full. *The reason is that $n$ does not even appear in those statements, and so we can always relabel the elements in $J \cup L$ to be equal to $[n']$ for some $n' < n$. However, for the equivalence to statement 4 we need the property $L \cup J = [n]$.*

## 4.5 $F$-Markov Random Fields and $F$-Markov Chains

In this section, we characterize $F$-Markov random fields with respect to a graph $\mathcal{G}$ in terms of the $F$-diagram. Then we specialize this to a characterization of $F$-Markov chains. These notions were defined in Terminology 2.20. As in the previous subsection, fix elements $X_1, \ldots, X_n \in M$, giving rise to a $G$-valued measure $\widetilde{F} : 2^{\widetilde{X}} \to G$ by Hu's Theorem 2.8. Additionally, we now fix a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V} = [n]$, see Definition 2.9.

**Lemma 4.23.** *Let $\mathcal{U}$ a cutset for $\mathcal{G}$. Let $\mathcal{V}_i(\mathcal{U})$, $1 \leq i \leq s(\mathcal{U})$ be the corresponding components and $K := (\mathcal{U}, \mathcal{V}_i(\mathcal{U}), 1 \leq i \leq s(\mathcal{U}))$ the corresponding conditional partition. Then all $p_{\mathcal{W}} \in \mathrm{Im}(K)$ are disconnected.*

*Proof.* Let $p_{\mathcal{W}} \in \mathrm{Im}(K)$. There exists $I \subseteq [s(\mathcal{U})]$ with $|I| \geq 2$ and $\emptyset \neq \mathcal{W}_i \subseteq \mathcal{V}_i(\mathcal{U})$ for all $i \in I$ such that $\mathcal{W} = \mathcal{W}_I = \bigcup_{i\in I} \mathcal{W}_i$. Now, let $i \neq j \in I$ and $w_i \in \mathcal{W}_i \subseteq \mathcal{V}_i(\mathcal{U})$ and $w_j \in \mathcal{W}_j \subseteq \mathcal{V}_j(\mathcal{U})$. Since $\mathcal{V}_i(\mathcal{U})$ and $\mathcal{V}_j(\mathcal{U})$ are components of $\mathcal{G}^{\setminus\mathcal{U}}$, there is no walk connecting $w_i$ and $w_j$ in $\mathcal{G}^{\setminus\mathcal{U}}$. Since $\mathcal{U} \subseteq \mathcal{V} \setminus \mathcal{W}_I$, there can also be no such walk in $\mathcal{G}^{\setminus(\mathcal{V}\setminus\mathcal{W}_I)}$. This shows that $\mathcal{V} \setminus \mathcal{W}_I$ is a cutset, and thus $p_{\mathcal{W}} = p_{\mathcal{W}_I}$ is disconnected. □

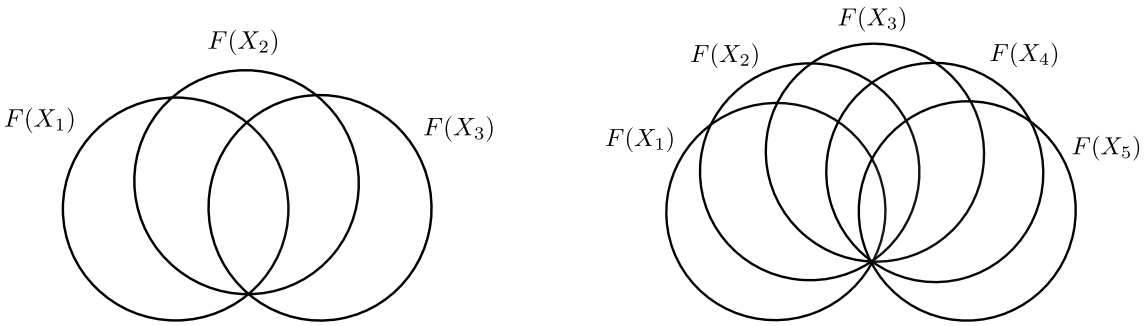



**Figure 8:** If elements $X_1, \dots, X_n \in M$ form an $F$-Markov chain, then many atoms in the $F$-diagram disappear by Corollary 4.24. The only atoms that remain are those corresponding to "intervals" in $[n]$. This leads to a fan-like structure of the $F$-diagram, as visualized here for $n = 3$ and $n = 5$.

The following proof of the characterization of $F$-Markov random fields is similar to the one given in Yeung et al. [2002], Theorem 8, for the special case of probabilistic Markov random fields.

*Proof of Theorem 2.21.* Assume that $X_1, \dots, X_n$ form an $F$-Markov random field with respect to $\mathcal{G}$. Let $p_{\mathcal{W}}$ be a disconnected atom. We need to show $\widetilde{F}(p_{\mathcal{W}}) = 0$.

Define $\mathcal{U} := \mathcal{V} \setminus \mathcal{W}$. By assumption of $\mathcal{W}$ being disconnected, $\mathcal{U}$ is a cutset. Thus, $s(\mathcal{U}) \geq 2$ and we have components $\mathcal{V}_1(\mathcal{U}), \dots, \mathcal{V}_{s(\mathcal{U})}(\mathcal{U})$ in $\mathcal{G}^{\setminus \mathcal{U}}$. Since $X_1, \dots, X_n$ form an $F$-Markov random field with respect to $\mathcal{G}$, we obtain by Proposition 3.5 the $F$-independence

$$F : \coprod_{i=1}^{s(\mathcal{U})} X_{\mathcal{V}_i(\mathcal{U})} \mid X_{\mathcal{U}}.$$

Now, notice that this is precisely the $F$-FCMI corresponding to the conditional partition $K = \big(\mathcal{U}, \mathcal{V}_i(\mathcal{U}), 1 \leq i \leq s(\mathcal{U})\big)$. By Theorem 4.21, $\widetilde{F}(p_{\mathcal{W}'}) = 0$ for all atoms $p_{\mathcal{W}'} \in \mathrm{Im}(K)$. Now, notice that $\bigcup_{i=1}^{s(\mathcal{U})} \mathcal{V}_i(\mathcal{U}) = \mathcal{V} \setminus \mathcal{U} = \mathcal{W}$, which, due to $s(\mathcal{U}) \geq 2$ and $\mathcal{V}_i(\mathcal{U}) \neq \emptyset$, shows that $p_{\mathcal{W}} \in \mathrm{Im}(K)$. This shows $\widetilde{F}(p_{\mathcal{W}}) = 0$, and we are done.

For the other direction, assume that $\widetilde{F}$ vanishes on all disconnected atoms. Now, let $\mathcal{U} \subseteq \mathcal{V}$ be a cutset, with components $\mathcal{V}_1(\mathcal{U}), \dots, \mathcal{V}_{s(\mathcal{U})}(\mathcal{U})$ of $\mathcal{G}^{\setminus \mathcal{U}}$. By Proposition 3.5, we need to show the $F$-independence

$$F : \coprod_{i=1}^{s(\mathcal{U})} X_{\mathcal{V}_i(\mathcal{U})} \mid X_{\mathcal{U}}.$$

This is the $F$-FCMI corresponding to the conditional partition $K = \big(\mathcal{U}, \mathcal{V}_i(\mathcal{U}), 1 \leq i \leq s(U)\big)$. Thus, by Theorem 4.21, we need to show that $\widetilde{F}(p_{\mathcal{W}}) = 0$ for all $p_{\mathcal{W}} \in \mathrm{Im}(K)$. Since all $p_{\mathcal{W}} \in \mathrm{Im}(K)$ are disconnected by Lemma 4.23, this follows from assuming that $\widetilde{F}$ vanishes on all disconnected atoms. □

We refer back to Figure 3 for a visualization of how $F$-diagrams of $F$-Markov random fields look like. That visualization focused on $I$-Markov random fields, but is entirely correct also in our general case, as the preceding theorem shows.

**Corollary 4.24.** *With all notation as above, the following two statements are equivalent:*

- $X_1, \dots, X_n$ *form an $F$-Markov chain.*
- $\widetilde{F}(p_{\mathcal{W}}) = 0$ *for all $\mathcal{W} \subseteq [n]$ that do* not *only contain consecutive numbers.*

*Proof.* For the proof, we specialize to the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = [n]$ and $\mathcal{E} = \big\{\{i, i+1\} \mid i = 1, \dots, n-1\big\}$, so $\mathcal{G}$ is intuitively a chain. Then from Proposition 3.6 we know that the first statement is equivalent to $X_1, \dots, X_n$ forming an $F$-Markov random field with respect to $\mathcal{G}$. Since the disconnected atoms $p_{\mathcal{W}}$ with respect to $\mathcal{G}$ are precisely those where $\mathcal{W}$ does not only consist of consecutive numbers, the result follows from Theorem 2.21. □

The effect of the preceding corollary on $F$-diagrams is visualized in Figure 8.

**Corollary 4.25.** *Assume $X_1, \dots, X_n$ form an $F$-Markov Chain. Then for all $I, J \subseteq [n]$ with $\emptyset \neq I = \{i_1 < i_2 < \dots < i_q\}$, the following equality holds:*

$$X_J.F(X_{i_1}; X_{i_2}; \dots; X_{i_q}) = X_J.F(X_{i_1}; X_{i_q}).$$

*Proof.* We can without loss of generality assume $X_J = \mathbf{1}$. Using Hu's Theorem 2.8, we obtain

$$\begin{aligned} F(X_{i_1}; \dots; X_{i_q}) = \widetilde{F}\left(\bigcap_{k=1}^{q} \widetilde{X}_{i_k}\right) &= \sum_{L:\ i_1,\dots,i_q \in L} \widetilde{F}(p_L) \overset{(\star)}{=} \sum_{L:\ i_1, i_q \in L} \widetilde{F}(p_L) \\ &= \widetilde{F}\left(\widetilde{X}_{i_1} \cap \widetilde{X}_{i_q}\right) = F(X_{i_1}; X_{i_q}). \end{aligned}$$

In $(\star)$, we used Corollary 4.24: both sums are equal since for all summands where $L$ does not contain the whole "interval" $[i_1 : i_q]$, we have $\widetilde{F}(p_L) = 0$. □

## 5 Probabilistic Independences and Markov Random Fields

In this section, we specialize the results from Section 4 to the probabilistic setting. Since we want to study probabilistic FCMIs and Markov random fields, we want to restrict our information measures — e.g., Shannon entropy, Kullback-Leibler divergence, and cross-entropy — to probability mass functions satisfying these properties. However, to be able to work in the setting of Section 2.2 of a general function $F : M \to G$, we need a monoid action, which in the probabilistic case involves conditional probability mass functions. We thus need to make sure that properties of distributions such as "forms a Markov random field" are preserved under conditioning. We prove this for several properties in Section 5.1.

In Section 5.2 we briefly study a general class of information functions that satisfy the chain rule, and motivate the case of higher-order cross-entropy and Kullback-Leibler divergence by a connection to the cluster cross-entropy from Cocco and Monasson [2012]. Building on Section 5.1, in Section 5.3, we restrict general information functions to stable properties like "forms a Markov random field" and show that this preserves the monoid action and the chain rule. In Theorem 5.12, we then recover the notion of a Markov random field with respect to restricted information functions, providing examples for the $F$-Markov random fields studied in Section 4. In Section 5.4 we then specialize to the case of Kullback-Leibler Markov chains with fixed transition probabilities between time-steps, reminiscent of the physical laws. This leads to Theorem 5.13 and Corollary 5.14, where we interpret a weak version of the second law of thermodynamics in terms of a degeneracy of Kullback-Leibler diagrams. Finally, in Section 5.5, we use the Kullback-Leibler decomposition over Markov chains that follows from Theorem 5.12 to derive the explicit decomposition of the evidence lower bound for diffusion models, demonstrating the applicability of our work to machine learning.

The results in Yeung et al. [2002] of $I$-diagram characterizations of FCMIs and Markov random fields focus on *fixed* probability mass functions, thus only looking at what we will call a *slice* of the $I$-diagram. This does not directly correspond to our supposed generalizations; after all, they always involve a monoid action, and thus let the probability mass functions unspecified in the probabilistic case. In Appendix D, we demonstrate that this is not a problem by showing that our Theorems 4.21 and 2.21 actually imply the main results in Yeung et al. [2002]. This validates the claim that our results are indeed generalizations.

In this whole section, let $\Omega$ be a countable, discrete sample space, with the probability mass function *not* yet fixed, and assume that all random variables $X : \Omega \to E_X$ have a finite discrete value space $E_X$. Also, recall the notions of the conditional independence $X \perp\!\!\!\perp_P Y \mid Z$ of random variables on $\Omega$ with respect to probability mass function $P$, and recall that equivalence classes of random variables form a monoid according to Proposition 2.1. Conditional independence is well-defined at the level of equivalence classes, giving rise to a separoid (Proposition 2.13), and we do not distinguish in notation between $X$ and its equivalence class.

## 5.1 Stability under Conditioning

In this subsection, we show that for random variables $X_1, \dots, X_n$, the property to be an FCMI, the property to be a Markov random field, and the property of two probability mass functions to have "equal transition probabilities" between time-steps in a Markov chain, are all *stable under conditioning*. That means that if they hold with respect to a probability mass function $P$, they also hold with respect to $P|_{Y=y}$ for arbitrary $Y = X_I$ with $I \subseteq [n]$ and $y \in E_Y$.

We will later also need the following definition, for example for defining the Kullback-Leibler divergence:

**Definition 5.1** (Absolutely Continuous)**.** *Let $P, Q \in \Delta(\Omega)$ be two probability mass functions. Then $P$ is said to be absolutely continuous with respect to $Q$, written $P \ll Q$, if the following implication holds for all $\omega \in \Omega$:*

$$Q(\omega) = 0 \quad \Longrightarrow \quad P(\omega) = 0.$$

We now specialize the notions of conditional mutual independences and FCMIs from general separoids to the probabilistic case to make wordings and notation more efficient. We did the similar specializations already for Markov random fields and Markov chains in Definitions 2.14 and 2.16.

**Terminology 5.2.** *Let $X_1, \dots, X_n, Y$ be random variables on $\Omega$, $P \in \Delta(\Omega)$ a probability mass function, and $\perp\!\!\!\perp_P$ the corresponding separation relation. If $X_1, \dots, X_n$ are mutually independent given $Y$ with respect to $\perp\!\!\!\perp_P$ (Definitions 3.1), then we write*

$$P : \coprod_{i=1}^{n} X_i \mid Y$$

*and call this a conditional mutual $P$-independence. Let $K = \big(J, L_i, 1 \leq i \leq q\big)$ be conditional partition of $[n]$ (Definition 4.16). Then the FCMI (Definition 3.3)*

$$P : \coprod_{i=1}^{q} X_{L_i} \mid X_J$$

*is called a full conditional mutual $P$-independence (with respect to $X_1, \dots, X_n$) — or $P$-FCMI for short. If it holds, then we say $K$ induces a $P$-FCMI.*

We first need two lemmas:

**Lemma 5.3.** *Let $X, Y, Z$ be three random variables on $\Omega$ and $P \in \Delta(\Omega)$ a probability mass function. If $P(y, z) \neq 0$, then*

$$\big(P|_{Y=y}\big)(x \mid z) = P(x \mid y, z).$$

*Proof.* We have:

$$\big(P|_{Y=y}\big)(x \mid z) = \Big(\big(P|_{Y=y}\big)|_{Z=z}\Big)(x) = \big(P|_{YZ=(y,z)}\big)(x) = P(x \mid y, z).$$

The second step can be showed by explicit computation. □

**Lemma 5.4.** *Let $U, V, W, Y$ be random variables on $\Omega$ and $P \in \Delta(\Omega)$ a probability mass function such that the following independence holds:*

$$U \underset{P}{\perp\!\!\!\perp} V \mid WY.$$

*Let $y \in E_Y$ be arbitrary with $P(y) \neq 0$. Then the following independence follows:*

$$U \underset{P|_{Y=y}}{\perp\!\!\!\perp} V \mid W.$$

*Proof.* Let $u \in E_U, v \in E_V$ and $w \in E_W$ be arbitrary. We want to show the following:

$$\big(P|_{Y=y}\big)(u, v, w) = \big(P|_{Y=y}\big)(u \mid w) \cdot \big(P|_{Y=y}\big)(v, w). \tag{23}$$

If $P(w \mid y) = 0$, then both sides of Equation (23) vanish and the result is clear. Thus, we can assume $P(w \mid y) \neq 0$, and therefore, together with $P(y) \neq 0$, we obtain $P(w, y) \neq 0$. By Lemma 5.3, this results in $\big(P|_{Y=y}\big)(u \mid w) = P(u \mid w, y)$. We obtain that the desired Equation (23) is equivalent to the following:

$$\frac{P(u, v, w, y)}{P(y)} = P(u \mid w, y) \cdot \frac{P(v, w, y)}{P(y)},$$

which is equivalent to

$$P(u, v, w, y) = P(u \mid w, y) \cdot P(v, w, y).$$

This follows from the assumption $U \underset{P}{\perp\!\!\!\perp} V \mid WY$. □

**Proposition 5.5.** *Let $X_1, \dots, X_n$ be random variables on $\Omega$, $P \in \Delta(\Omega)$ a probability mass function, and $K = \big(J, L_i, 1 \leq i \leq q\big)$ a conditional partition of $[n]$ that induces a $P$-FCMI with respect to $X_1, \dots, X_n$. Then for all $I \subseteq [n]$, for $Y := X_I$, and for all $y \in E_Y$ with $P(y) \neq 0$, $K$ also induces a $\big(P|_{Y=y}\big)$–FCMI with respect to $X_1, \dots, X_n$.*

*Proof.* By assumption, we have the $P$-FCMI $P : \perp\!\!\!\perp_{i=1}^{q} X_{L_i} \mid X_J$. This means that for all $i \in \{1, \dots, q\}$, we have

$$X_{L_i} \underset{P}{\perp\!\!\!\perp} X_{\bigcup_{l \neq i} L_l} \mid X_J. \tag{24}$$

Now, decompose $I$ into three sets $I_i = I \cap L_i$, $I_{\setminus i} = I \cap \bigcup_{l \neq i} L_l$ and $I_J = I \cap J$. We obtain the corresponding random variables $Y_i = X_{I_i}$, $Y_{\setminus i} = X_{I_{\setminus i}}$ and $Y_J = X_{I_J}$. Since the monoid of random variables is idempotent, the above pairwise $P$-independence is equivalent to the following:

$$Y_i X_{L_i} \underset{P}{\perp\!\!\!\perp} Y_{\setminus i} X_{\bigcup_{l \neq i} L_l} \mid X_J Y_J.$$

Applying weak union (S4) to both the left and right factor, we obtain

$$X_{L_i} \underset{P}{\perp\!\!\!\perp} X_{\bigcup_{l \neq i} L_l} \mid X_J \big(Y_i Y_{\setminus i} Y_J\big).$$

Since $K$ is *full*, meaning the contained sets cover all of $[n]$, we obtain $I = I_i \cup I_{\setminus i} \cup I_J$ and consequently $Y_i Y_{\setminus i} Y_J = Y$. Lemma 5.4 then shows the $\big(P|_{Y=y}\big)$–independence $X_{L_i} \underset{P|_{Y=y}}{\perp\!\!\!\perp} X_{\bigcup_{l \neq i} L_l} \mid X_J$. Since $i$ was arbitrary, we obtain the $\big(P|_{Y=y}\big)$–FCMI $P|_{Y=y} : \perp\!\!\!\perp_{i=1}^{q} X_{L_i} \mid X_J$. That was to show. □

**Corollary 5.6.** *Let $X_1, \dots, X_n$ be random variables on $\Omega$, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ a graph with $\mathcal{V} = [n]$, and $P \in \Delta(\Omega)$ a probability mass function. Assume $X_1, \dots, X_n$ form a $P$-Markov random field with respect to $\mathcal{G}$.*

*Then for all $I \subseteq [n]$, for $Y := X_I$, and for all $y \in E_Y$ with $P(y) \neq 0$, $X_1, \dots, X_n$ also form a $\big(P|_{Y=y}\big)$–Markov random field with respect to $\mathcal{G}$.*

*Proof.* We show this by proving the cutset property, which by Proposition 3.5 is equivalent to the global Markov property. Thus, let $\mathcal{U} \subseteq [n]$ be a cutset and $\mathcal{V}_1(\mathcal{U}), \dots, \mathcal{V}_{s(\mathcal{U})}(\mathcal{U})$ the corresponding components. By assumption we know that the conditional partition $K = \big(\mathcal{U}, \mathcal{V}_i(\mathcal{U}), 1 \leq i \leq s(\mathcal{U})\big)$ induces a $P$-FCMI. Then Proposition 5.5 shows that we also obtain a $\big(P|_{Y=y}\big)$–FCMI, meaning we have $P|_{Y=y} : \perp\!\!\!\perp_{i=1}^{s(\mathcal{U})} X_{\mathcal{V}_i(\mathcal{U})} \mid X_{\mathcal{U}}$. This was to show. □

Next, we show that the property to have "equal transition probabilities" between time-steps in a Markov chain is stable under conditioning. We will use this in Section 5.4 to illustrate a weak version of the second law of thermodynamics. First, we state a simple lemma about general separoids that we will use:

**Lemma 5.7** (Dawid [2001], Lemma 1.1)**.** *Let $(M, \perp\!\!\!\perp)$ be a separoid. For $X, Y, Z \in M$, we have the following equivalence:*

$$X \perp\!\!\!\perp Y \mid Z \quad \Longleftrightarrow \quad XZ \perp\!\!\!\perp YZ \mid Z.$$

**Proposition 5.8.** *Let $X_1, \dots, X_n$ be random variables on $\Omega$ and $(P, Q) \in \Delta(\Omega)^2$ be two probability mass functions such that $P$ is absolutely continuous with respect to $Q$, see Definition 5.1. Assume that $X_1, \dots, X_n$ form a $P$-Markov chain* and *$Q$-Markov chain. Additionally, assume that $P$ and $Q$ have the same transition probabilities, i.e.: $P(x_i \mid x_{i-1}) = Q(x_i \mid x_{i-1})$ for all $i$ and all $x_{i-1}, x_i$ with $P(x_{i-1}) \neq 0$.*

*Then for all $I \subseteq [n]$, for $Y := X_I$, and for all $y \in E_Y$ with $P(y) \neq 0$, $P|_{Y=y}$ is also absolutely continuous with respect to $Q|_{Y=y}$ and $X_1, \dots, X_n$ also forms a $\big(P|_{Y=y}\big)$–Markov chain and a $\big(Q|_{Y=y}\big)$–Markov chain. Additionally, $P|_{Y=y}$ and $Q|_{Y=y}$ also have the same transition probabilities, i.e.: $(P|_{Y=y})(x_i \mid x_{i-1}) = (Q|_{Y=y})(x_i \mid x_{i-1})$ for all $i$ and all $x_{i-1}, x_i$ with $(P|_{Y=y})(x_{i-1}) \neq 0$.*

*Proof.* That $P|_{Y=y}$ is absolutely continuous with respect to $Q|_{Y=y}$ is clear. That $X_1, \dots, X_n$ also forms a $\big(P|_{Y=y}\big)$–Markov chain and a $\big(Q|_{Y=y}\big)$–Markov chain follows from the fact that Markov chains are Markov random fields by Proposition 3.6, and from Corollary 5.6.

For the second statement, let $i$ be given. Write $I = I_- \cup I_+$ with $I_- \subseteq [i-1]$, $I_+ \subseteq [i:n]$. Correspondingly, we write $Y = Y_- Y_+$. Remember that being a Markov chain implies being a Markov random field according to Proposition 3.6. Together with Lemma 5.7 and the decomposition separoid axiom (S3), one can show the following two independences:

$$X_i \underset{P}{\perp\!\!\!\perp} Y_- \mid X_{i-1}Y_+, \qquad X_i \underset{Q}{\perp\!\!\!\perp} Y_- \mid X_{i-1}Y_+.^{10}$$

It follows

$$\begin{aligned}
\big(P|_{Y=y}\big)(x_i \mid x_{i-1}) &\overset{(\star)}{=} P(x_i \mid x_{i-1}, y) && \text{(Lemma 5.3)} \\
&= P(x_i \mid x_{i-1}, y_-, y_+) \\
&= P(x_i \mid x_{i-1}, y_+) \\
&= \frac{P(x_i, y_+ \mid x_{i-1})}{P(y_+ \mid x_{i-1})} \\
&= \frac{P(x_i \mid x_{i-1}) \cdot P(y_+ \mid x_{i-1}, x_i)}{\sum_{x_i'} P(x_i' \mid x_{i-1}) \cdot P(y_+ \mid x_{i-1}, x_i')},
\end{aligned}$$

In $(\star)$, to apply the lemma, we used $P(y) \neq 0$ and the additional assumption that $\big(P|_{Y=y}\big)(x_{i-1}) \neq 0$, which results in $P(y, x_{i-1}) \neq 0$. Due to $P$ being absolutely continuous with respect to $Q$, we also obtain $Q(y, x_{i-1}) \neq 0$. Thus, we similarly get:

$$\big(Q|_{Y=y}\big)(x_i \mid x_{i-1}) = \frac{Q(x_i \mid x_{i-1}) \cdot Q(y_+ \mid x_{i-1}, x_i)}{\sum_{x_i'} Q(x_i' \mid x_{i-1}) \cdot Q(y_+ \mid x_{i-1}, x_i')}.$$

By assumption, we already know $P(x_i \mid x_{i-1}) = Q(x_i \mid x_{i-1})$. The preceding computations therefore show that it is enough to prove that $P(y_+ \mid x_{i-1}, x_i) = Q(y_+ \mid x_{i-1}, x_i)$. We can assume that $I_+ \subseteq [i+1:n]$ since otherwise $x_i$ determines the first entry of $y_+$, which can then be removed from the expression if it is equal to $x_i$. Then, let $J \subseteq [i+1:n]$ be such that $J \dot\cup I_+ = [i+1:n]$. We obtain:

$$\begin{aligned}
P(y_+ \mid x_{i-1}, x_i) &= \sum_{x_J} P\big(x_{[i+1:n]} \mid x_{i-1}, x_i\big) \\
&= \sum_{x_J} \prod_{j=i}^{n-1} P(x_{j+1} \mid x_j) && \text{(Markov chain property)} \\
&= \sum_{x_J} \prod_{j=i}^{n-1} Q(x_{j+1} \mid x_j)
\end{aligned}$$

[10] More precisely, the global Markov property ensures $X_i \perp\!\!\!\perp_P X_{[i-2]} \mid X_{i-1}Y_+$ (Note that this is trivial if $X_i$ is part of $Y_+$, and otherwise we have the disjointness stated in the global Markov property). Then by Lemma 5.7, we can add the variables $X_{i-1}Y_+$ to the left-hand-sides. The decomposition separoid axiom (S3) then allows to remove the variables that we do not need for the final result.

$$= \sum_{x_J} Q\big(x_{[i+1:n]} \mid x_{i-1}, x_i\big) \qquad \text{(Markov chain property)}$$
$$= Q(y_+ \mid x_{i-1}, x_i).$$

This was to show. □

## 5.2 Information Functions Satisfying the Chain Rule

Before we show in the next subsection that information functions can always be restricted to stable probability sets, we first define a class of information functions to work with. This includes Shannon entropy, Kullback-Leibler divergence, and cross-entropy. We also connect higher-order cross-entropy and Kullback-Leibler divergence to the notion of cluster entropy defined in Cocco and Monasson [2012].

To capture many information functions, we work in the following general setting very similar to Baudot and Bennequin [2015], Section 5. Let $M$ be the monoid of equivalence classes of random variables on the fixed sample space $\Omega$, and let $r \geq 0$ be a natural number. We denote tuples of probability mass functions in $\Delta(\Omega)^{r+1}$ by $\big(P \| Q_1, \dots, Q_r\big) \coloneqq \big(P, Q_1, \dots, Q_r\big)$ to remind of the notation inside the Kullback-Leibler divergence. Then, we set

$$\widetilde{\Delta(\Omega)^{r+1}} \coloneqq \Big\{ \big(P \| Q_1, \dots, Q_r\big) \in \Delta(\Omega)^{r+1} \;\Big|\; P \ll Q_1, \dots, P \ll Q_r \Big\}. \tag{25}$$

Here, $P \ll Q$, means that $P$ is absolutely continuous with respect to $Q$, see Definition 5.1. In applications, $r = 0$ would be used for Shannon entropy and $\alpha$-entropy, and $r = 1$ for Kullback-Leibler divergence, $\alpha$–Kullback-Leibler divergence and cross-entropy, indicating the number of additional probability mass functions. We do not have applications for $r \geq 2$ in mind.

Let

$$G \coloneqq \mathrm{Meas}\Big(\widetilde{\Delta(\Omega)^{r+1}}, \mathbb{R}\Big) \coloneqq \Big\{ f : \widetilde{\Delta(\Omega)^{r+1}} \to \mathbb{R} \;\Big|\; f \text{ is measurable} \Big\}, \tag{26}$$

which clearly is an abelian group. We also assume an additive monoid action of $M$ on $G$ given by the formula

$$\big[X.f\big]\big(P \| Q_1, \dots, Q_r\big) = \sum_{x \in E_X} g\big(P(x), Q_1(x), \dots, Q_r(x)\big) \cdot f\big(P|_{X=x} \| Q_1|_{X=x}, \dots, Q_r|_{X=x}\big) \tag{27}$$

for some measurable function $g : [0,1]^{r+1} \to \mathbb{R}$ with $g(0, \star, \dots, \star) = 0$.[11] We assume $g$ to be the same function irrespective of $X, f$, and $P, Q_1, \dots, Q_r$. Note that $P \ll Q_i$ implies that also $P|_{X=x} \ll Q_i|_{X=x}$, which means that the preceding formula is actually defined.

Furthermore, we assume to have a function $F : M \to G$ that satisfies the chain rule, meaning that for all $X, Y \in M$, we have

$$F(XY) = F(X) + X.F(Y).$$

The reader can verify that all of Shannon entropy, Kullback-Leibler divergence, $\alpha$-entropy, $\alpha$–Kullback-Leibler divergence, and cross-entropy, fit precisely under this umbrella. Precise definitions that fit our exact framework can be found in Lang et al. [2025], Sections 2 and 5. For example, for Shannon entropy we have $r = 0$ and $G = \mathrm{Meas}\big(\Delta(\Omega), \mathbb{R}\big)$. The action is given by

$$\big[X.f\big](P) \coloneqq \sum_{x \in E_X} P(x) \cdot f\big(P|_{X=x}\big),$$

and the Shannon entropy $I : M \to G$ is given by

$$\big[I(X)\big](P) \coloneqq -\sum_{x \in E_X} P(x) \cdot \log P(x), \tag{28}$$

[11] The condition $g(0, \star, \dots, \star) = 0$ is satisfied in all examples we consider and ensures that we have a zero coefficient whenever the conditionals are not defined. We then define the product as simply zero. Additionally, note that due to absolute continuity, we have $P(x) = 0$ whenever $Q_i(x) = 0$ for some $i$. Therefore, it is enough to have the vanishing condition for $g$ in the first component only.

which indeed satisfies the chain rule $I(XY) = I(X) + X.I(Y)$; see Section 2.1. Note that $0 \cdot \log 0 \coloneqq 0$ in this formula. Similarly, for Kullback-Leibler divergence we have $r = 1$ and $G = \mathrm{Meas}\left(\widetilde{\Delta(\Omega)^2}, \mathbb{R}\right)$. The action is given by

$$\big[X.f\big](P\|Q) \coloneqq \sum_{x \in E_X} P(x) \cdot f\big(P|_{X=x} \| Q|_{X=x}\big). \tag{29}$$

The Kullback-Leibler divergence $D : M \to G$ is then defined by

$$\big[D(X)\big](P\|Q) \coloneqq \sum_{x \in E_X} P(x) \cdot \log \frac{P(x)}{Q(x)}. \tag{30}$$

This also satisfies the chain rule: $D(XY) = D(X) + X.D(Y)$. Note that $0 \cdot \log \frac{0}{Q(x)} \coloneqq 0$ in this formula. Similarly, the cross-entropy $C : M \to G$ is defined by

$$\big[C(X)\big](P\|Q) \coloneqq \sum_{x \in E_X} P(x) \cdot \log \frac{1}{Q(x)}. \tag{31}$$

Clearly, one then has $C(X) = I(X) + D(X)$.

Once these definitions are set in place, one can study higher-order terms $F : M^q \to G$ as in Equation (6), and visualize them in $F$-diagrams. The higher-order terms for Shannon entropy $I$, i.e., the mutual information and interaction information, are frequently studied, whereas the higher-order terms for $C$ and $D$ have gotten substantially less attention. So before we continue to study the $F$-diagrams of such functions for Markov random fields in the coming section, we briefly connect these higher-order terms to the cluster entropy from Cocco and Monasson [2012].

Namely, Cocco and Monasson [2012] define the cluster (cross-)entropy[12] $\Delta C$ for subsets of $n$ random variables $X_1, \dots, X_n$ by

$$\Delta C\big(\,;_{j \in J} X_j\big) \coloneqq \sum_{K \subseteq J} (-1)^{|K|-|J|} \cdot C(X_K). \tag{32}$$

In contrast, we define higher-order cross-entropies as in Equation (6) by the inductive formula

$$C\big(X_{j_1}; \dots; X_{j_q}\big) = C\big(X_{j_1}; \dots; X_{j_{q-1}}\big) - X_{j_q}.C\big(X_{j_1}; \dots; X_{j_{q-1}}\big).$$

In Lang et al. [2025], Corollary 3.5, part 6, it is deduced from the generalized Hu theorem that this obeys the following inclusion-exclusion type formula:

$$C\big(\,;_{j \in J} X_j\big) = \sum_{K \subseteq J} (-1)^{|K|+1} \cdot C(X_K). \tag{33}$$

Comparing Equations (32) and (33), we see that our higher-order cross-entropies and the cluster cross-entropies differ by a simple sign that depends on the order:

$$C\big(\,;_{j \in J} X_j\big) = (-1)^{|J|+1} \cdot \Delta C\big(\,;_{j \in J} X_j\big).$$

Higher-order Kullback-Leibler divergence is then fully determined by higher-order cross-entropy and information, as $D\big(\,;_{j \in J} X_j\big) = C\big(\,;_{j \in J} X_j\big) - I\big(\,;_{j \in J} X_j\big)$. We hope that these connections provide motivation for the reader to study information diagrams for functions such as cross-entropy and Kullback-Leibler divergence, and that the study of these diagrams, in turn, could help to better understand concepts like the cluster cross-entropies.

## 5.3 Restricting Information Measures to Stable Probability Sets

In this subsection, we show that information functions that "satisfy the chain rule" can, under mild conditions, always be restricted to subsets of probability mass functions as long as those are stable. We define this notion in Definition 5.9, which encompasses the examples we saw in the previous subsection. In Theorem 5.12, we will then see that restricting to probability mass functions that give rise to Markov random fields leads to the formalism of an $F$-Markov random field, as studied in Section 4.5. Let the notation throughout be as in the preceding subsection.

[12] The authors often write "entropy" when they actually mean cross-entropy.

**Definition 5.9** (Stable Property)**.** *Let $n \geq 0$, $X_1, \dots, X_n \in M$, $r \geq 0$, and $\mathcal{R} = \mathcal{R}(X_1, \dots, X_n)$ be a property of probability tuples with $r+1$ entries, which means that for all $(P\|Q_1, \dots, Q_r) \in \Delta(\Omega)^{r+1}$, $\mathcal{R}(P\|Q_1, \dots, Q_r)$ is either true or false. $\mathcal{R}$ is called* stable *if the following holds:*

- *$\mathcal{R}$ is* well-defined*, i.e., if $Y_1, \dots, Y_n$ are random variables on $\Omega$ with $Y_i \sim X_i$, then*
$$\mathcal{R}(X_1, \dots, X_n) = \mathcal{R}(Y_1, \dots, Y_n);$$
- *$\mathcal{R}$ is* measurable*, i.e., the set*
$$\Big\{(P\|Q_1, \dots, Q_r) \in \Delta(\Omega)^{r+1} \;\Big|\; \mathcal{R}(P\|Q_1, \dots, Q_r)\Big\}$$
*is a measurable subset of $\Delta(\Omega)^{r+1}$;*
- *$\mathcal{R}$ is* stable under conditioning with respect to $X_1, \dots, X_n$*, i.e.: let $M_{\mathcal{R}}$ be the submonoid of $M$ generated by $X_1, \dots, X_n$.*[13] *Then we assume that for all $(P\|Q_1, \dots, Q_r) \in \Delta(\Omega)^{r+1}$, all $Y \in M_{\mathcal{R}}$, and all $y \in E_Y$ with $P(y) \neq 0$, following implication holds:*
$$\mathcal{R}(P\|Q_1, \dots, Q_r) \quad \Longrightarrow \quad \mathcal{R}(P|_{Y=y}\|Q_1|_{Y=y}, \dots, Q_r|_{Y=y}).$$

Now, let $X_1, \dots, X_n \in M$, and let $\mathcal{R} = \mathcal{R}(X_1, \dots, X_n)$ be a stable property of probability tuples with $r+1$ elements. Define

$$\big(\widetilde{\Delta(\Omega)^{r+1}}\big)_{\mathcal{R}} \coloneqq \Big\{(P\|Q_1, \dots, Q_r) \in \widetilde{\Delta(\Omega)^{r+1}} \;\Big|\; \mathcal{R}(P\|Q_1, \dots, Q_r)\Big\}.$$

Then, define

$$G_{\mathcal{R}} \coloneqq \mathrm{Meas}\left(\big(\widetilde{\Delta(\Omega)^{r+1}}\big)_{\mathcal{R}}, \mathbb{R}\right) \tag{34}$$

and the action $\cdot_{\mathcal{R}} : M_{\mathcal{R}} \times G_{\mathcal{R}} \to G_{\mathcal{R}}$ by

$$\big[X.{}_{\mathcal{R}} f\big](P\|Q_1, \dots, Q_r) \coloneqq \big[X.\tilde{f}\big](P\|Q_1, \dots, Q_r), \tag{35}$$

where $\tilde{f}$ is any measurable extension of $f$ to all of $\widetilde{\Delta(\Omega)^{r+1}}$. Note that this means that $X.\tilde{f}$ is a measurable extension of $X.{}_{\mathcal{R}} f$, and that $X.{}_{\mathcal{R}} f$ is thus again measurable.

**Lemma 5.10.** *The function $\cdot_{\mathcal{R}} : M_{\mathcal{R}} \times G_{\mathcal{R}} \to G_{\mathcal{R}}$, as defined above, is a well-defined additive monoid action.*

*Proof.* For well-definedness, one first needs to check that for all measurable $f : \big(\widetilde{\Delta(\Omega)^{r+1}}\big)_{\mathcal{R}} \to \mathbb{R}$, a measurable extension $\tilde{f}$ to all of $\widetilde{\Delta(\Omega)^{r+1}}$ exists: note that from the fact that $\mathcal{R}$ is stable and thus measurable, we immediately obtain that $\big(\widetilde{\Delta(\Omega)^{r+1}}\big)_{\mathcal{R}}$ is a measurable subset of $\widetilde{\Delta(\Omega)^{r+1}}$. Then $\tilde{f}$ can, for example, be constructed by mapping all elements in $\widetilde{\Delta(\Omega)^{r+1}} \setminus \big(\widetilde{\Delta(\Omega)^{r+1}}\big)_{\mathcal{R}}$ to zero, which is clearly measurable.

Furthermore, one needs to check that the definition is independent of the extension $\tilde{f}$ and the representative of the *equivalence class* $X$. The first requirement follows from Equation (27): $g$ was assumed to be independent of $\tilde{f}$, and $\tilde{f}$ is only applied to elements on which already $f$ is uniquely defined, due to the assumption of $\mathcal{R}$ being stable under conditioning. That the formula is independent of the representative of $X$ then follows since the original action was well-defined.

That $\cdot_{\mathcal{R}}$ is additive and $\mathbf{1} \in M_{\mathcal{R}}$ acts neutrally is obvious. We now show that the action is associative:

$$\begin{aligned}
\Big[(XY).{}_{\mathcal{R}} f\Big](P\|Q_1, \dots, Q_r) &= \Big[(XY).\tilde{f}\Big](P\|Q_1, \dots, Q_r) \\
&= \Big[X.\big(Y.\tilde{f}\big)\Big](P\|Q_1, \dots, Q_r) \\
&\overset{(\star)}{=} \Big[X.{}_{\mathcal{R}}\big(Y.{}_{\mathcal{R}} f\big)\Big](P\|Q_1, \dots, Q_r).
\end{aligned}$$

In step $(\star)$, we used that $Y.\tilde{f}$ is by definition a measurable extension of $Y.{}_{\mathcal{R}} f$, which by definition of the "outer" action $\cdot_{\mathcal{R}}$ gives the result. This finishes the proof. □

[13] I.e., this submonoid consists of all $X_I$ for $I \subseteq [n]$.

Finally, let $F : M \to G$ be a function satisfying the chain rule as in Section 5.2 and define

$$F_{\mathcal{R}} : M_{\mathcal{R}} \to G_{\mathcal{R}}, \quad \big[F_{\mathcal{R}}(X)\big](P\|Q_1, \dots, Q_r) \coloneqq \big[F(X)\big](P\|Q_1, \dots, Q_r). \tag{36}$$

In other words, $F_{\mathcal{R}}(X)$ is just the restriction of $F(X)$ to the tuples of probability mass functions satisfying property $\mathcal{R}$.

**Proposition 5.11.** *Let $r \geq 0$ and $G$ be the abelian group defined in Equation (26). Assume $M$, the monoid of equivalence classes of random variables on $\Omega$, acts additively on $G$ by Equation (27), and that $F : M \to G$ is a function satisfying the chain rule Equation (5).*

*Now, let $X_1, \dots, X_n$ be random variables on $\Omega$ and $\mathcal{R} = \mathcal{R}(X_1, \dots, X_n)$ a stable property of tuples of $r+1$ probability mass functions. Let $M_{\mathcal{R}}$ be the submonoid generated by $X_1, \dots, X_n$ and $G_{\mathcal{R}}$ be defined as in Equation (34). Let $M_{\mathcal{R}}$ act on $G_{\mathcal{R}}$ as in Equation (35). Then the restricted information function $F_{\mathcal{R}} : M_{\mathcal{R}} \to G_{\mathcal{R}}$, defined in Equation (36), satisfies the chain rule:*

$$F_{\mathcal{R}}(XY) = F_{\mathcal{R}}(X) + X.{}_{\mathcal{R}}F_{\mathcal{R}}(Y).$$

*Proof.* We have

$$\begin{aligned}
\big[F_{\mathcal{R}}(XY)\big]\big(P\|Q_1, \dots, Q_r\big) &= \big[F(XY)\big]\big(P\|Q_1, \dots, Q_r\big) \\
&= \big[F(X)\big]\big(P\|Q_1, \dots, Q_r\big) + \big[X.F(Y)\big]\big(P\|Q_1, \dots, Q_r\big) \\
&\overset{(\star)}{=} \big[F_{\mathcal{R}}(X)\big]\big(P\|Q_1, \dots, Q_r\big) + \big[X.{}_{\mathcal{R}}F_{\mathcal{R}}(Y)\big]\big(P\|Q_1, \dots, Q_r\big) \\
&= \big[F_{\mathcal{R}}(X) + X.{}_{\mathcal{R}}F_{\mathcal{R}}(Y)\big]\big(P\|Q_1, \dots, Q_r\big).
\end{aligned}$$

In step $(\star)$, in the right-hand term, we used that $F(Y)$ extends $F_{\mathcal{R}}(Y)$, which by definition of the action $.{}_{\mathcal{R}}$ results in $X.F(Y)$ extending $X.{}_{\mathcal{R}}F_{\mathcal{R}}(Y)$. □

The preceding proposition ensures that we can apply Hu's Theorem 2.8 to $F_{\mathcal{R}}$, which we will make use of in the next section on Kullback-Leibler Markov chains.

In the next theorem, we restrict this setting somewhat further to obtain a result on $F_{\mathcal{R}}$-Markov random fields that will apply to Shannon entropy, Kullback-Leibler divergence and cross-entropy. The result will, however, as stated not apply to $\alpha$-entropy and $\alpha$–Kullback-Leibler divergence.

For the theorem, we make the stronger assumption that the action of $M$ on $G$ is given by

$$\big[X.f\big]\big(P\|Q_1, \dots, Q_r\big) = \sum_{x \in E_X} P(x) \cdot f\big(P|_{X=x}\|Q_1|_{X=x}, \dots, Q_r|_{X=x}\big), \tag{37}$$

meaning that, in Equation (27), we have $g(p, \star, \dots, \star) = p$. Additionally, we make the assumption that $F$ is given by

$$\big[F(X)\big]\big(P\|Q_1, \dots, Q_r\big) = \sum_{x \in E_X} P(x) \cdot h\big(P(x), Q_1(x), \dots, Q_r(x)\big) \tag{38}$$

for some function $h : [0,1]^{r+1} \to \mathbb{R}$. Clearly, all these stronger assumptions hold for Shannon entropy, Kullback-Leibler divergence, and cross-entropy, see, e.g., Equations (28) and (30).[14]

Now, fix random variables $X_1, \dots, X_n$ on $\Omega$ and a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = [n]$. We now define the property

$$\mathcal{MRF} \coloneqq \mathcal{MRF}(\mathcal{G}; X_1, \dots, X_n). \tag{39}$$

When applied to a probability tuple $\big(P\|Q_1, \dots, Q_r\big)$, this evaluates to "true" if $X_1, \dots, X_n$ form a $P$-Markov random field and, for all $i$, a $Q_i$-Markov random field, with respect to $\mathcal{G}$. We know from Corollary 5.6 that $\mathcal{MRF}$ is stable under conditioning. It is also clearly well-defined, since conditional independences do not depend on the equivalence classes of the involved random variables. Finally, it is measurable since it is a conjunction of conditional independence relations: the subset of probability mass functions that satisfy an independence relation can by Equation (8) easily seen to be measurable.

[14] To make those examples fit the framework precisely, one needs to define $\log(0)$ to be an arbitrary real number instead of $-\infty$. This does not change the result as $\log(0)$ is always multiplied with the coefficient 0.

Thus, $\mathcal{MRF}$ is a stable property and we obtain the function $F_{\mathcal{MRF}}$ that satisfies the chain rule by Proposition 5.11. Consequently, we obtain a separation rule $\perp\!\!\!\perp_{F_{\mathcal{MRF}}}$ by Proposition 4.8. In the following, we slightly generalize from this by considering properties that *imply* the Markov random field property, which is crucial in the next section when restricting to properties that are motivated by physics.

**Theorem 5.12.** *Let $r \geq 0$ and $G$ be the abelian group defined in Equation (26). Assume $M$, the monoid of equivalence classes of random variables on $\Omega$, acts additively on $G$ by Equation (37), and that $F : M \to G$ is a function of the form Equation (38) satisfying the chain rule Equation (5).*

*Now, let $X_1, \dots, X_n$ be random variables on $\Omega$, $\mathcal{G}$ a graph with vertex set $[n]$, and $\mathcal{R} = \mathcal{R}(\mathcal{G}; X_1, \dots, X_n)$ a stable property that implies the property $\mathcal{MRF}$ from Equation (39):*

$$\forall (P \| Q_1, \dots, Q_r) \in \widetilde{\Delta(\Omega)^{r+1}} : \quad \mathcal{R}(P \| Q_1, \dots, Q_r) \quad \Longrightarrow \quad \mathcal{MRF}(P \| Q_1, \dots, Q_r).$$

*Then $X_1, \dots, X_n$ form an $F_{\mathcal{R}}$-Markov random field with respect to $\mathcal{G}$, where $F_{\mathcal{R}}$ is defined as in Equation (36) as a restriction of $F$.*

*Proof.* Let $\mathcal{A}, \mathcal{B}, \mathcal{C} \subseteq \mathcal{V} = [n]$ be disjoint vertex sets such that $\mathcal{C}$ separates $\mathcal{A}$ from $\mathcal{B}$. We need to show the $F_{\mathcal{R}}$-independence $X_{\mathcal{A}} \perp\!\!\!\perp_{F_{\mathcal{R}}} X_{\mathcal{B}} \mid X_{\mathcal{C}}$, which, by Proposition 4.13, is equivalent to the equality

$$(X_{\mathcal{C}} X_{\mathcal{B}}).{}_{\mathcal{R}} F_{\mathcal{R}}(X_{\mathcal{A}}) = X_{\mathcal{C}}.{}_{\mathcal{R}} F_{\mathcal{R}}(X_{\mathcal{A}}). \tag{40}$$

For all $(P \| Q_1, \dots, Q_r) \in \left(\widetilde{\Delta(\Omega)^{r+1}}\right)_{\mathcal{R}}$, we obtain:

$$\begin{aligned}
&\Big[(X_{\mathcal{C}} X_{\mathcal{B}}).{}_{\mathcal{R}} F_{\mathcal{R}}(X_{\mathcal{A}})\Big](P \| Q_1, \dots, Q_r) \\
&= \sum_{x_{\mathcal{C}}, x_{\mathcal{B}}} P(x_{\mathcal{C}}, x_{\mathcal{B}}) \cdot \Big[F_{\mathcal{R}}(X_{\mathcal{A}})\Big]\Big(P|_{X_{\mathcal{C}} X_{\mathcal{B}} = (x_{\mathcal{C}}, x_{\mathcal{B}})} \parallel Q_1|_{X_{\mathcal{C}} X_{\mathcal{B}} = (x_{\mathcal{C}}, x_{\mathcal{B}})}, \dots, Q_r|_{X_{\mathcal{C}} X_{\mathcal{B}} = (x_{\mathcal{C}}, x_{\mathcal{B}})}\Big) \\
&= \sum_{x_{\mathcal{C}}, x_{\mathcal{B}}} P(x_{\mathcal{C}}, x_{\mathcal{B}}) \sum_{x_{\mathcal{A}}} P(x_{\mathcal{A}} \mid x_{\mathcal{C}}, x_{\mathcal{B}}) \cdot h\Big(P(x_{\mathcal{A}} \mid x_{\mathcal{C}}, x_{\mathcal{B}}),\ Q_1(x_{\mathcal{A}} \mid x_{\mathcal{C}}, x_{\mathcal{B}}), \dots, Q_r(x_{\mathcal{A}} \mid x_{\mathcal{C}}, x_{\mathcal{B}})\Big) \\
&\overset{(\star)}{=} \sum_{x_{\mathcal{C}}, x_{\mathcal{B}}} P(x_{\mathcal{C}}, x_{\mathcal{B}}) \sum_{x_{\mathcal{A}}} P(x_{\mathcal{A}} \mid x_{\mathcal{C}}, x_{\mathcal{B}}) \cdot h\Big(P(x_{\mathcal{A}} \mid x_{\mathcal{C}}),\ Q_1(x_{\mathcal{A}} \mid x_{\mathcal{C}}), \dots, Q_r(x_{\mathcal{A}} \mid x_{\mathcal{C}})\Big) \\
&= \sum_{x_{\mathcal{C}}, x_{\mathcal{A}}} \left(\sum_{x_{\mathcal{B}}} P(x_{\mathcal{A}}, x_{\mathcal{C}}, x_{\mathcal{B}})\right) \cdot h\Big(P(x_{\mathcal{A}} \mid x_{\mathcal{C}}),\ Q_1(x_{\mathcal{A}} \mid x_{\mathcal{C}}), \dots, Q_r(x_{\mathcal{A}} \mid x_{\mathcal{C}})\Big) \\
&= \sum_{x_{\mathcal{C}}} P(x_{\mathcal{C}}) \sum_{x_{\mathcal{A}}} P(x_{\mathcal{A}} \mid x_{\mathcal{C}}) \cdot h\Big(P(x_{\mathcal{A}} \mid x_{\mathcal{C}}),\ Q_1(x_{\mathcal{A}} \mid x_{\mathcal{C}}), \dots, Q_r(x_{\mathcal{A}} \mid x_{\mathcal{C}})\Big) \\
&= \sum_{x_{\mathcal{C}}} P(x_{\mathcal{C}}) \cdot \Big[F_{\mathcal{R}}(X_A)\Big]\Big(P|_{X_{\mathcal{C}} = x_{\mathcal{C}}} \parallel Q_1|_{X_{\mathcal{C}} = x_{\mathcal{C}}}, \dots, Q_r|_{X_{\mathcal{C}} = x_{\mathcal{C}}}\Big) \\
&= \Big[X_{\mathcal{C}}.{}_{\mathcal{R}} F_{\mathcal{R}}(X_A)\Big](P \| Q_1, \dots, Q_r).
\end{aligned}$$

In the calculation, in step $(\star)$ we used the assumption that

$$X_{\mathcal{A}} \perp\!\!\!\perp_{P} X_{\mathcal{B}} \mid X_{\mathcal{C}}, \quad X_{\mathcal{A}} \perp\!\!\!\perp_{Q_1} X_{\mathcal{B}} \mid X_{\mathcal{C}}, \quad \dots \quad X_{\mathcal{A}} \perp\!\!\!\perp_{Q_r} X_{\mathcal{B}} \mid X_{\mathcal{C}},$$

which is due to $(P \| Q_1, \dots, Q_r) \in \left(\widetilde{\Delta(\Omega)^{r+1}}\right)_{\mathcal{R}} \subseteq \left(\widetilde{\Delta(\Omega)^{r+1}}\right)_{\mathcal{MRF}}$. The other steps follow from the definitions. Overall, this shows Equation (40), and we are done. □

## 5.4 The Second Law of Thermodynamics

In this section, we study a weak version of the second law of thermodynamics and show that it can be represented visually by the degeneracy of a certain Kullback-Leibler diagram.

We fix random variables $X_1, \dots, X_n$ on $\Omega$. We now work with the rather special property $\mathcal{P} = \mathcal{P}(X_1, \dots, X_n)$ of probability tuples $(P \| Q) \in \Delta(\Omega)^2$ defined as follows: $\mathcal{P}(P \| Q)$ holds if and only if

- $P$ is absolutely continuous with respect to $Q$;
- $X_1, \dots, X_n$ forms a $P$-Markov chain and $Q$-Markov chain; and
- for all $i \geq 2$ and all $(x_{i-1}, x_i)$ with $P(x_{i-1}) \neq 0$, we have $P(x_i \mid x_{i-1}) = Q(x_i \mid x_{i-1})$.

The only difference of $P$ and $Q$ then stems from the initial distributions $P(X_1)$, $Q(X_1)$. Intuitively, the letter "P" in the property $\mathcal{P} = \mathcal{P}(X_1, \dots, X_n)$ is supposed to remind of "Physics". Namely, let $(P, Q) \in \widetilde{\Delta(\Omega)^2}$ with $\mathcal{P}\big(P \| Q\big)$, then:

- $i = 1, \dots, n$ indexes time points with equal separation;
- The value spaces $E_{X_i}$ model the sets of possible micro states at time point $i$ (in reality, all these spaces are equal to each other, but we need not make this assumption);
- $P(X_1)$ and $Q(X_1)$ are the initial distributions of $P$ and $Q$, which can be considered as macro states of the universe at some fixed point in time; later, in the context of the second law of thermodynamics, we will choose $P(X_1)$ to be arbitrary and $Q(X_1)$ to be the uniform distribution;
- $P(x_i \mid x_{i-1}) = Q(x_i \mid x_{i-1})$ models the likelihood that $x_{i-1}$ evolves to $x_i$ according to the "physical laws" described by $P$ and $Q$; In reality, the physical laws are fixed and time independent, which means there is one transition matrix $T$ such that $P(x_i \mid x_{i-1}) = Q(x_i \mid x_{i-1}) = T(x_i \mid x_{i-1})$ for all $i$ and all $x_{i-1}, x_i$, but we do not yet make this stronger assumption;
- $P(X_i)$ and $Q(X_i)$ are the macro states of the universe at time points $i$ when evolving $P(X_1)$ and $Q(X_1)$ according to the "physical laws" specified by $P$ and $Q$. These laws are equal to each other by assumption.

In the following, we want to investigate a weak version of the second law of thermodynamics, which states that under some conditions, entropy cannot decrease over time. We show this by first investigating the Kullback-Leibler diagram for the property $\mathcal{P}$ in Theorem 5.13, which will lead to the desired result in Corollary 5.14.

Formally, let $D : M \to \mathrm{Meas}\left(\widetilde{\Delta(\Omega)^2}, \mathbb{R}\right)$ be the Kullback-Leibler divergence, see Equation (30). The property $\mathcal{P}$ is stable under conditioning by Proposition 5.8, and is also easily seen to be measurable and well-defined. Therefore, by Proposition 5.11, the restricted Kullback-Leibler divergence

$$D_{\mathcal{P}} : M_{\mathcal{P}} \to \mathrm{Meas}\left(\left(\widetilde{\Delta(\Omega)^2}\right)_{\mathcal{P}}, \mathbb{R}\right) =: G_{\mathcal{P}}$$

satisfies the chain rule, where $M_{\mathcal{P}}$ is the monoid generated by $X_1, \dots, X_n$. Accordingly, by Hu's Theorem 2.8, we obtain a corresponding $G_{\mathcal{P}}$-valued measure $\widetilde{D_{\mathcal{P}}} : 2^{\widetilde{X}} \to G_{\mathcal{P}}$. The following theorem shows that this measure degenerates, as visualized in Figure 9:

**Theorem 5.13** (Structure of the $D_{\mathcal{P}}$-diagram)**.** *Let $D : M \to G$ be the Kullback-Leibler divergence from Equation (30). Let $X_1, \dots, X_n$ be random variables on $\Omega$ and consider the property $\mathcal{P} = \mathcal{P}(X_1, \dots, X_n)$ defined above, leading to a restriction $D_{\mathcal{P}} : M_{\mathcal{P}} \to G_{\mathcal{P}}$ and a $G_{\mathcal{P}}$-valued measure $\widetilde{D_{\mathcal{P}}} : 2^{\widetilde{X}} \to G_{\mathcal{P}}$. One has the following:*

- *For all atoms $p_I$ where $I$ does not only consist of consecutive numbers, one has $\widetilde{D_{\mathcal{P}}}(p_I) = 0$;*
- *Let $p_{[i:k]}$ with $1 \leq i \leq k \leq n$ be an atom that consists of only consecutive numbers. If $i \geq 2$, then $\widetilde{D_{\mathcal{P}}}\big(p_{[i:k]}\big) = 0$.*

*Proof.* By Proposition 3.6 and Theorem 5.12, $X_1, \dots, X_n$ forms a $D_{\mathcal{P}}$-Markov random field with respect to the graph $\mathcal{G}$ with edges $i - (i+1)$ for $i = 1, \dots, n-1$. By Proposition 3.6 again, we know that this is equivalent to forming a $D_{\mathcal{P}}$–Markov chain. Then by Corollary 4.24, one has $\widetilde{D_{\mathcal{P}}}(p_I) = 0$ if $I \subseteq [n]$ does *not* only consist of consecutive numbers. Thus, only the atoms $p_I$ of the form $I = [i:k]$ are of relevance for the $D_{\mathcal{P}}$-diagram.

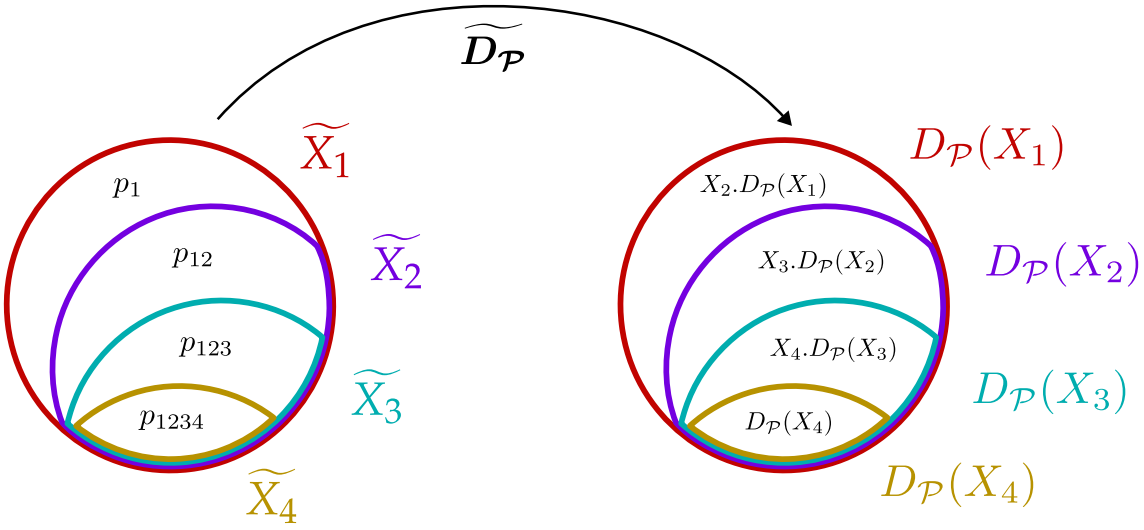



**Figure 9:** If one restricts the set of tuples $(P, Q)$ of probability mass functions to those that have equal transition probabilities and individually give rise to a Markov chain $X_1, \dots, X_n$, then this leads to a restricted Kullback-Leibler divergence $D_{\mathcal{P}}$ and the corresponding measure $\widetilde{D_{\mathcal{P}}}$. Theorem 5.13 shows that in the corresponding $D_{\mathcal{P}}$-diagram, only atoms of the form $p_{12\dots k}$ remain. This is visualized here for the case $n = 4$ by omitting all atoms that are mapped to zero by $\widetilde{D_{\mathcal{P}}}$. *Intuitively, the Kullback-Leibler divergence shrinks over time.* In particular, this leads by Hu's Theorem 2.8 to the chain rule $D_{\mathcal{P}}(X_{i-1}) = D_{\mathcal{P}}(X_i) + X_i.D_{\mathcal{P}}(X_{i-1})$, which we use in the proof of Corollary 5.14 to deduce a weak version of the second law of thermodynamics.

Let $i \geq 2$. Using Lemma 4.4, Corollary 4.25, and Equation (6), we have

$$\begin{aligned}
\widetilde{D_{\mathcal{P}}}\big(p_{[i:k]}\big) &= \big(X_{[i-1]}X_{[k+1:n]}\big).D_{\mathcal{P}}\big(X_i; \dots; X_k\big) \\
&= \big(X_{[i-1]}X_{[k+1:n]}\big).D_{\mathcal{P}}\big(X_i; X_k\big) \\
&= \big(X_{[i-1]}X_{[k+1:n]}\big).D_{\mathcal{P}}\big(X_i\big) - \big(X_{[i-1]}X_{[k:n]}\big).D_{\mathcal{P}}\big(X_i\big) \\
&= \big(X_{[i-2]}X_{[k+1:n]}\big).\Big(X_{i-1}.D_{\mathcal{P}}\big(X_i\big)\Big) - \big(X_{[i-2]}X_{[k:n]}\big).\Big(X_{i-1}.D_{\mathcal{P}}\big(X_i\big)\Big)
\end{aligned}$$

Since the monoid action sends zero elements to zero, it is enough to show $X_{i-1}.D_{\mathcal{P}}(X_i) = 0$. For arbitrary $\big(P \| Q\big) \in \big(\widetilde{\Delta(\Omega)^2}\big)_{\mathcal{P}}$, we have

$$\begin{aligned}
\big[X_{i-1}.D_{\mathcal{P}}(X_i)\big]\big(P\|Q\big) &= \sum_{x_{i-1}} P(x_{i-1}) \cdot \big[D_{\mathcal{P}}(X_i)\big]\big(P|_{X_{i-1}=x_{i-1}} \parallel Q|_{X_{i-1}=x_{i-1}}\big) \\
&= \sum_{x_{i-1}} P(x_{i-1}) \sum_{x_i} P(x_i \mid x_{i-1}) \cdot \log \frac{P(x_i \mid x_{i-1})}{Q(x_i \mid x_{i-1})} \\
&= \sum_{x_{i-1}} P(x_{i-1}) \sum_{x_i} P(x_i \mid x_{i-1}) \cdot \log 1 \\
&= 0,
\end{aligned}$$

where in the third step we used the assumptions of equal transition probabilities, which holds since $\mathcal{P}(P\|Q)$ is true. That was to show. □

We deduce the following weak version of the second law of thermodynamics:

**Corollary 5.14** (Second Law of Thermodynamics)**.** *Let all notation be as above. Assume that the joint $X_1 \cdots X_n : \Omega \to E_{X_1} \times \cdots \times E_{X_n}$ is surjective. Let $P$ be a probability mass function on $\Omega$ such that $X_1, \dots, X_n$ forms a $P$-Markov chain. Assume that $E_{X_i} = E_{X_j}$ for all $i, j = 1, \dots, n$ and that there is a time-independent doubly stochastic transition matrix $T$ such that $P(x_i \mid x_{i-1}) =$*

*$T(x_i \mid x_{i-1})$ for all $i = 2, \dots, n$ and $(x_{i-1}, x_i)$ with $P(x_{i-1}) \neq 0$. Let $I : M \to G$ be the Shannon entropy, as defined in Equation (28). Then for all $i = 2, \dots, n$, one has*

$$\big[I(X_{i-1})\big](P) \leq \big[I(X_i)\big](P),$$

*i.e., entropy cannot decrease.*

*Proof.* The proof is a diagrammatic representation of the reasoning in Cover and Thomas [2006], Chapter 4.4. Let $Q$ be any probability mass function on $\Omega$ such that

- $Q_{X_1}$ is the uniform distribution on $X_1$;
- $Q(x_i \mid x_{i-1}) = T(x_i \mid x_{i-1})$ for all $i = 2, \dots, n$ and $(x_{i-1}, x_i)$.[15]

Then, as is well-known, the fact that $T$ is doubly stochastic implies that $Q_{X_i}$ is uniform for all $i = 1, \dots, n$. Let $D$ be the Kullback-Leibler divergence. We obtain

$$\begin{aligned}
\big[D(X_{i-1})\big](P\|Q) &= -\sum_{x_{i-1}} P(x_{i-1}) \log \frac{Q(x_{i-1})}{P(x_{i-1})} \\
&= -\sum_{x_{i-1}} P(x_{i-1}) \log \frac{1}{\big|E_{X_{i-1}}\big|} - \big[I(X_{i-1})\big](P) \\
&= C - \big[I(X_{i-1})\big](P),
\end{aligned}$$

where $C$ is a constant, i.e., independent of $i, P$ and $Q$, and where the uniformity of $Q_{X_{i-1}}$ was used in the second step. Thus, it is enough to show that $\big[D(X_{i-1})\big](P\|Q)$ satisfies the following inequality:

$$\big[D(X_{i-1})\big](P\|Q) \geq \big[D(X_i)\big](P\|Q).$$

Now, note that $P \ll Q$ since $Q$ is nowhere zero. Also, by assumption, $P$ and $Q$ both have equal transition probabilities given by $T$. We obtain $\big(P\|Q\big) \in \Big(\widetilde{\Delta(\Omega)^2}\Big)_{\mathcal{P}}$. Thus, Theorem 5.13, as visualized by Figure 9, shows

$$\big[D(X_{i-1})\big](P\|Q) = \big[D(X_i)\big](P\|Q) + \big[X_i.D(X_{i-1})\big](P\|Q) \geq \big[D(X_i)\big](P\|Q), \tag{41}$$

where in the last step we used that conditional Kullback-Leibler divergence is non-negative. That was to show. □

## 5.5 Diffusion Models

We briefly elaborate another simple illustration of our theory, namely its help in deriving the explicit decomposition of the evidence lower bound (ELBO) in diffusion models [Sohl-Dickstein et al., 2015]. Diffusion models are used in text-to-image generation models like Dalle [Ramesh et al., 2021] and Imagen [Saharia et al., 2022] that recently received widespread attention.

In the classical formulation of diffusion models, the data is assumed to come from a distribution $Q(X) \in \Delta(E_X)$ that is progressively transformed by fixed conditional noise distributions $Q(Z_1 \mid X)$ and $Q(Z_t \mid Z_{t-1})$, $t = 2, \dots, T$ over latent value spaces $E_{Z_1}, \dots, E_{Z_T}$. One can then form the joint distribution $Q(X, \mathbf{Z}) = Q(X) \cdot Q(Z_1 \mid X) \cdot \prod_{t=2}^{T} Q(Z_t \mid Z_{t-1})$. $X, Z_1, \dots, Z_T$ then form a Markov chain with respect to $Q$.

The goal for diffusion models is to learn a model distribution $P(X)$ that is close to the data distribution $Q(X)$. This is done by fixing a latent distribution $P(Z_T)$ and initializing denoising distributions $P(Z_{t-1} \mid Z_t)$, $t = 2, \dots, T$ and $P(X \mid Z_1)$ that are parameterized by deep neural networks. As before, one can then also build the joint distribution $P(X, \mathbf{Z}) = P(Z_T) \cdot \prod_{t=2}^{T} P(Z_{t-1} \mid Z_t) \cdot P(X \mid Z_1)$. Clearly, $Z_T, \dots, Z_1, X$ then form a Markov chain with respect to $P$, and due to symmetry (compare Proposition 3.6), also $X, Z_1, \dots, Z_T$ form a Markov chain with respect to $P$.

[15] To see that this exists, first construct a joint distribution on $E_{X_1} \times \cdots \times E_{X_n}$ with these properties by multiplying the initial distribution with all the transitional distributions. Then, "pull it back" to $\Omega$, for which there is at least one possible construction since the joint $X_1 \cdots X_n$ is surjective.

Now, fix a datapoint $x \in E_X$ sampled from $Q(X)$. Since minimizing the Kullback-Leibler divergence between $Q(X)$ and $P(X)$ is equivalent to maximizing the log-likelihood of $P(X)$ for datapoints sampled from $Q(X)$, the goal is to make a gradient step that increases $\log P(x)$. In diffusion models, this is accomplished by constructing a so-called *evidence-lower bound* (ELBO) of this objective that can easily be optimized with respect to the neural network parameters and does not require intractable summations. This is in contrast to the marginal $P(x)$, which is a sum over all $\mathbf{z} \in E_{Z_1} \times \cdots \times E_{Z_T}$. The main result is as follows:

**Proposition 5.15** (Computing the ELBO for diffusion models)**.** *Assume $P(Z_T)$ is fixed and that $Q(Z_T \mid x) = P(Z_T)$.*[16] *Define the ELBO by*

$$\mathcal{L}(x) \coloneqq \sum_{\mathbf{z}} Q(\mathbf{z} \mid x) \log \frac{P(x, \mathbf{z})}{Q(\mathbf{z} \mid x)}.$$

*Then $\mathcal{L}(x) \leq \log P(x)$ and*

$$\mathcal{L}(x) = \sum_{z_1} Q(z_1 \mid x) \log P(x \mid z_1) - \sum_{t=2}^{T} \sum_{z_t} Q(z_t \mid x) \cdot D\Big(Q(Z_{t-1} \mid z_t, x) \;\Big\|\; P(Z_{t-1} \mid z_t)\Big).^{17} \tag{42}$$

*Proof.* The two claims involve two different factorizations of $P(x, \mathbf{z})$. For the first claim, we note

$$\begin{aligned} \mathcal{L}(x) &= \sum_{\mathbf{z}} Q(\mathbf{z} \mid x) \log \frac{P(x) \cdot P(\mathbf{z} \mid x)}{Q(\mathbf{z} \mid x)} \\ &= \log P(x) - D\big(Q(\mathbf{Z} \mid x) \parallel P(\mathbf{Z} \mid x)\big) \\ &\leq \log P(x), \end{aligned}$$

which follows from the non-negativity of Kullback-Leibler divergence.

For the second claim, we note

$$\begin{aligned} \mathcal{L}(x) &= \sum_{\mathbf{z}} Q(\mathbf{z} \mid x) \log \frac{P(x \mid \mathbf{z}) \cdot P(\mathbf{z})}{Q(\mathbf{z} \mid x)} \\ &= \sum_{\mathbf{z}} Q(\mathbf{z} \mid x) \log P(x \mid \mathbf{z}) - D\big(Q(\mathbf{Z} \mid x) \parallel P(\mathbf{Z})\big) \end{aligned}$$

Now, since $Z_T, \dots, Z_1, X$ form a Markov chain with respect to $P$, we have $P(x \mid \mathbf{z}) = P(x \mid z_1)$, and so the left part of the formula simplifies:

$$\sum_{\mathbf{z}} Q(\mathbf{z} \mid x) \log P(x \mid \mathbf{z}) = \sum_{z_1} Q(z_1 \mid x) \log P(x \mid z_1).$$

Thus, it just remains to evaluate the Kullback-Leibler divergence. We first perform the computation and then justify our steps:

$$\begin{aligned} D\big(Q(\mathbf{Z} \mid x) \parallel P(\mathbf{Z})\big) &\overset{(1)}{=} \big[D(\mathbf{Z})\big]\big(Q|_{X=x} \parallel P\big) \\ &\overset{(2)}{=} \left[\sum_{t=2}^{T} Z_t.D(Z_{t-1}) + D(Z_T)\right]\big(Q|_{X=x} \parallel P\big) \\ &\overset{(3)}{=} \sum_{t=2}^{T} \sum_{z_t} Q(z_t \mid x) \cdot D\Big(Q(Z_{t-1} \mid z_t, x) \;\Big\|\; P(Z_{t-1} \mid z_t)\Big) \\ &\quad + D\big(Q(Z_T \mid x) \parallel P(Z_T)\big). \end{aligned}$$

[16] In practice, this is achieved by adding so much noise to $x$ over the Markov chain $Q$ that eventually, all information in $x$ is destroyed.

[17] For simplicity, we write $D\big(Q(Y) \| P(Y)\big)$ for $[D(Y)](Q \| P)$.

Since $Q(Z_T \mid x)$ and $P(Z_T)$ are assumed identical, the last Kullback-Leibler divergence disappears, proving the original claim.

We now justify the steps. Step (1) just makes explicit that Kullback-Leibler divergence is a function whose arguments are a random variable *followed by* a pair of distributions, see Equation (30).

For step (2), remember that $X, Z_1, \dots, Z_T$ form a Markov chain with respect to $P$ and $Q$. The Markov chain property is conditionally stable by Corollary 5.6 and the fact that Markov chains are special Markov random fields (Proposition 3.6). Thus, $X, Z_1, \dots, Z_T$ also form a Markov chain with respect to $Q|_{X=x}$. Consequently, by Theorem 5.12, the *restriction* of $D$ to the Markov chain property gives rise to a $D$-Markov chain. Thus, using Corollary 4.24, as visualized in Figure 8, together with Hu's Theorem (Theorem 2.8), gives rise to the decomposition in step (2).

Step (3) simply uses the definition of the monoid action, Equation (29). This proves the claim. □

Comparing with the exposition in Bishop and Bishop [2023], we see that our derivation avoids a few complications: We do not need to compute explicit Bayesian posteriors, reason about marginalizations over many variables, or reason about terms that cancel each other. Since the Kullback-Leibler divergence is ubiquitous in machine learning, we think there could be many similar applications of our work that aim to find decompositions of loss functions over Markov random fields. We think that for Markov random fields that are more complex than the Markov chains encountered in diffusion models, our formalism could make finding useful decompositions and approximations of loss functions simpler than the alternative of more low-level reasoning.

# 6 Discussion

## 6.1 Major Findings: Characterizations of $F$–FCMIs, $F$–Markov Random Fields, and Probabilistic Applications

In this work, we have generalized the main results in Yeung et al. [2002], which characterize probabilistic full conditional mutual independences and Markov random fields in terms of $I$-diagrams [Hu, 1962, Yeung, 1991]. In doing so, we replaced the Shannon entropy $I$ with any function $F : M \to G$ from a commutative, idempotent monoid $M$ to an abelian group $G$ that satisfies the chain rule:

$$F(XY) = F(X) + X.F(Y).$$

The dot denotes an additive monoid action of $M$ on $G$ that generalizes the conditioning of information functions on random variables.

Consequently, we also replaced the familiar probabilistic conditional $P$-independence — which is characterized by a vanishing of conditional mutual information — by $F$-independence:

$$X \underset{F}{\perp\!\!\!\perp} Y \mid Z \quad :\Longleftrightarrow \quad Z.F(X;Y) = 0.$$

This independence relation gives rise to a separoid [Dawid, 2001], as we have shown in Proposition 4.8, a powerful framework in which one can study conditional independence relations. Separoids generally also allow us to study conditional *mutual* independences and full conditional mutual independences (FCMIs). These can be used to characterize Markov random fields by the *cutset property*, as was first observed in Yeung et al. [2002] for the classical case — see Section 3.2. When specializing to the separoid with the independence relation given by $\perp\!\!\!\perp_F$ we obtained the notions of $F$-FCMIs and $F$-Markov random fields that generalize the probabilistic counterparts.

By the generalized Hu Theorem 2.8 from Lang et al. [2025], for fixed elements $X_1, \dots, X_n \in M$, one obtains a $G$-valued measure $\widetilde{F}$ that describes relations of information terms $X_J.F\big(X_{L_1}; \dots; X_{L_q}\big)$ of arbitrary degree $q$, where $J, L_1, \dots, L_q \subseteq [n]$ are any subsets. These can then be visualized in $F$-diagrams, as we show for $n = 3$ in Figure 2. The core question asked in this work — generalizing the investigations in Yeung et al. [2002] — was how $F$-FCMIs and $F$-Markov random fields can be characterized in terms of the $F$-diagram. In Theorem 4.21, we found that $F$-FCMIs are characterized by a vanishing of the $F$-diagram on the atoms in the *image* of the corresponding conditional partition, see also Figures 6 and 7.

This results in a characterization of $F$-Markov random fields for a graph $\mathcal{G}$, Theorem 2.21, that is easy to interpret: an atom $p_{\mathcal{W}}$ disappears in the $F$-diagram if and only if it is *disconnected*, meaning that the vertex set $\mathcal{W}$ is disconnected when removing the vertices outside of $\mathcal{W}$ from $\mathcal{G}$. We visualized this result in Figure 3 for Shannon entropy, though the basic picture applies for arbitrary $F$. A visualization for the special case of a path-shaped graph — recovering $F$-Markov chains — can be found in Figure 8.

When looking closely at the results, which are always given as an equivalence of two or more statements, it becomes clear that one direction is always easy to prove — namely, if one starts assuming that a suitable set of atoms disappears in the $F$-diagram, then one can easily deduce an $F$-FCMI from it (possibly in the context of an $F$-Markov random field). After all, $\widetilde{F}$ is a measure, and so the vanishing of a set of atoms automatically leads to the vanishing of all sets *composed* of these atoms, and this includes by Hu's Theorem 2.8 precisely the sets encoding the $F$-FCMIs. However, going in the other direction is harder: why should an $F$-FCMI induce the vanishing of specific atoms? This problem is addressed by our main technique "subset determination", Theorem 4.1. It implies that whenever $\widetilde{F}(A) = 0$ for some set of atoms $A$, then we also have $\widetilde{F}(B) = 0$ for all subsets $B \subseteq A$, and in particular, $\widetilde{F}(p_I) = 0$ for all atoms $p_I \in A$. This is essentially based on inclusion-exclusion type arguments for $F$-diagrams, and uses the monoid action from $M$ on $G$. We illustrate how to use subset determination to our advantage in Figure 5. Essentially, subset determination replaces the use of inequalities in the original work [Yeung et al., 2002].

We then applied our results to the case where the information functions apply to probability mass functions. In Lemma 5.10 and Proposition 5.11 we showed that it is possible to restrict a general notion of information functions to conditionally stable properties while preserving the monoid action and the chain rule, which is similar to the use of adapted probability functors in information cohomology [Vigneaux, 2019]. Then, in Theorem 5.12, we restricted a narrower set of information functions $F$ — including Shannon entropy, Kullback-Leibler divergence, and cross-entropy — to stable properties $\mathcal{R}$ that *imply* the Markov random field property. The restriction $F_{\mathcal{R}}$ then gives rise to an $F_{\mathcal{R}}$–Markov random field; consequently, disconnected atoms disappear from the $F_{\mathcal{R}}$-diagram.

In Theorem 5.13, we applied this to the case of tuples of probability mass functions that give rise to a Markov chain and have the same conditional distributions — reminiscent of physics, where the conditional distributions are governed by the physical laws. It turns out that the corresponding $D$-diagram degenerates even further than the Markov chain property alone would predict; the Kullback-Leibler divergence of all regions outside the initial random variable disappears. The reason is that in those regions, one "conditions" on the initial variable according to Hu's Theorem 2.8, and since the conditional distributions are the same, one obtains zero Kullback-Leibler divergence.

The result is an ever-shrinking[18] sequence of Kullback-Leibler divergences across the Markov chain, as we visualize in Figure 9. This has the following consequence: if one of the two distributions is uniform, then the other one cannot move further away from it as measured by the Kullback-Leibler divergence, which means that its entropy cannot decrease over time. This is a weak version of the second law of thermodynamics, see Corollary 5.14.

Finally, we also used the Kullback-Leibler decomposition over Markov chains that results from Theorem 5.12 to obtain a conceptually simple derivation of the evidence lower bound in diffusion models, Proposition 5.15. This avoids some of the explicit computations used in prior expositions.

It is important to note that in most of these probabilistic results, we take a view in which information functions are still *functions* of yet unspecified probability mass functions. We simply restrict the sets of probability mass functions to those that satisfy stable properties, and can thereby preserve our general theory, the monoid action, and the "subset determination" property, Theorem 4.1, that does not generally hold for fixed probability mass functions. This raises the question of whether the $P$-independence results in Yeung et al. [2002] are actually covered by our work. We answer this affirmatively in Appendix D. In the corresponding proofs, we use the fact that a $P$-independence $X \underset{P}{\perp\!\!\!\perp} Y \mid Z$ is equivalent to the vanishing of the conditional mutual $P$-information $\big[Z.I(X;Y)\big](P)$, which is usually proved using Jensen's inequality. This reduction is then the only place where inequalities enter the theory, whereas they take a more central role in the proofs in Yeung et al. [2002].

[18] More precisely: non-increasing.

#### Further Findings: $F$-(Dual )Total Correlation, Cohomological Characterizations of Functions $F$, and Further Consequences

Our proof of the $F$-diagram characterization of $F$-FCMIs uses Theorem 4.15, in which we characterize conditional mutual $F$-independences by the vanishing of $F$-dual total correlation. This relates to $F$ in the same way that the classical dual total correlation from Han [1978] relates to Shannon entropy:

$$DTC_F(X_1;\dots;X_n) \coloneqq F\big(X_{[n]}\big) - \sum_{i=1}^{n} X_{[n]\setminus i}.F(X_i).$$

Conditional $F$-dual total correlation $Y.DTC_F\big(X_1,\dots,X_n\big)$ corresponds by Hu's Theorem 2.8 to a set of atoms, which coincide with those that vanish based on a conditional mutual $F$-independence $F : \perp\!\!\!\perp_{i=1}^{n} X_i \mid Y$. Again, we were able to use subset determination, Theorem 4.1, to prove the characterization.

In Appendix B, we also investigated $F$-total correlation, which generalizes the classical total correlation from Watanabe [1960]. We showed that it can also usually be used for a characterization of mutual $F$-independences. However, since it double-counts some atoms, the characterization only works for torsion-free groups $G$, and we provide a counter example in the case of torsion, see Example B.5. This is not a strong restriction — all groups that are used in practice seem to be derived from the real numbers $\mathbb{R}$, and therefore inherit the property to be torsion-free.

Subset determination implies that $F\big(X_{[n]}\big)$ determines the whole $F$-diagram for the variables $X_1,\dots,X_n$. In Appendix A, this led to a classification of functions $F : M \to G$ satisfying the chain rule, in the case that $M$ has a "top element" — similar to how $X_{[n]}$ is "on top" of all elements $X_I$ for $I \subseteq [n]$. We could show that such functions $F$ correspond precisely to elements in $G$ that are annihilated by the top element, which we also used in our constructions for Example B.5. In Remark A.6, we explained a cohomological interpretation of these findings: the first Hochschild cohomology group of $M$ with coefficients in $G$ disappears. This is in contrast to information cohomology [Baudot and Bennequin, 2015, Vigneaux, 2019], where the joint locality property ensures that the first cohomology group is nontrivial; in fact, it is generated by Shannon entropy!

Finally, in Appendix C, we explain that all results from Yeung et al. [2019] except those that depend on an order relation also generalize to our setting. Overall, this means that we have generalized most of the results in Kawabata and Yeung [1992], Yeung [1991], Yeung et al. [2002, 2019] from Shannon entropy to general functions $F$ satisfying the chain rule.

### 6.2 Conceivable Extensions of the Theory and Open Questions

#### $K$-Independence and Kolmogorov Complexity

We have set up the theory in such a way that it could in principle be extended beyond $F$-independence, since Section 3 introduces conditional (mutual) independences, Markov random fields, and Markov chains in the general context of separoids. In this context, we also characterized Markov random fields — defined in terms of the global Markov property — by the *cutset property* (Proposition 3.5), and showed that Markov chains can be equivalently described as Markov random fields corresponding to a path-shaped graph (Proposition 3.6).

We can think of one concrete way to potentially go beyond $F$-independence. Namely, Lang et al. [2025] also studied the case of functions $K : M \times M \to G$ satisfying the chain rule absent of any monoid action of $M$ on $G$: $K(XY) = K(X) + K(Y \mid X)$, where $K(Z) \coloneqq K(Z \mid \mathbf{1})$. One could then define the $K$-independence by $X \underset{K}{\perp\!\!\!\perp} Y \mid Z$ if $K(X;Y \mid Z) = 0$.

One can then ask questions similar to the ones answered in this work: In what generality is $(M, \perp\!\!\!\perp_K)$ a separoid? And when it is, does it allow to characterize conditional mutual $K$-independences and $K$-Markov random fields in terms of $K$-diagrams? Note that $K$-diagrams do indeed exist, as proven in Corollary 3.3 in Lang et al. [2025]. However, we expect $\perp\!\!\!\perp_K$ to not always satisfy the separoid axioms since our proof of the separoid axioms in Proposition 4.8 made use of subset determination, Theorem 4.1, which in turn builds on the monoid action which is not available in the setting using $K$.

In a specific case of interest, $K$ would be a version of Kolmogorov complexity [Li and Vitányi, 2019], especially Chaitin's prefix-free Kolmogorov complexity [Chaitin, 1987], restricted to sets

of strings that satisfy approximate independence relations. The fact that conditional mutual Kolmogorov complexity $K(x; y \mid z)$ is approximately non-negative implies that the separoid rules can be proved in this case, see Steudel et al. [2010].

### $F$-Bayesian Networks

Additionally, it would be interesting to attempt to go beyond Markov random fields by studying Bayesian networks. A priori, they can be defined in general separoids using the d-separation criterion for a directed acyclic graph [Pearl, 1985]. This can then be applied to the case of a separation $\perp\!\!\!\perp_F$ coming from a function $F$ satisfying the chain rule.

When applying such a theory to the probabilistic case, however, one will likely encounter problems: $F$-independences can only model conditionally stable cases, see Proposition 4.9. However, the independences in Bayesian networks are, due to colliders, not preserved under conditioning, see Remark 4.10. Therefore, this property is not stable, Definition 5.9. When restricting information functions to such sets of probability mass functions, one thus loses the monoid action that makes use of conditioning. Since the monoid action is crucially used in our main technique — subset determination, Theorem 4.1 — we expect such a theory to take a different route from the one for $F$–Markov random fields. Nevertheless, Steudel et al. [2010], Theorem 1, provides an information characterization of causal Bayesian networks for submodular information functions, which can be used as an inspiration.

### $O$-Information and $S$-Information

In Rosas et al. [2019] and Medina-Mardones et al. [2021], the $O$-information and $S$-information were defined as the difference and sum of the classical total correlation and dual total correlation, respectively. We have showed in this work that $F$-(dual )total correlation preserves its value in studying conditional mutual $F$-independences, and we therefore expect that $F$–$O$-information and $F$–$S$-information might provide useful insights into general characterizations of high-order interdependencies.

### (Cluster) Cross-Entropies and Kullback-Leibler Divergence

We think it is worthwhile to study the case where $F$ is cross-entropy or Kullback-Leibler divergence in greater detail. After all, much of machine learning and deep learning involves the minimization of a cross-entropy, or equivalently Kullback-Leibler divergence [Bishop, 2007, Bishop and Bishop, 2023]. This becomes especially interesting for graphical methods, including diffusion models [Sohl-Dickstein et al., 2015] that form the basis for widespread text-to-image generation methods like Dalle [Ramesh et al., 2021], Imagen [Saharia et al., 2022], and stable diffusion [Rombach et al., 2022]. Diffusion models involve a decomposition of a joint Kullback-Leibler divergence over a Markov chain as in Figure 8. It could be valuable to analyze the loss functions of more such graphical methods using Kullback-Leibler and cross-entropy-diagrams. Similarly, the (cluster) cross-entropies (see Section 5.2) used in adaptive cluster expansion [Cocco and Monasson, 2012] of Ising models, which form a Markov random field, deserve a further study.

## Conclusion

In this work we showed that it is possible to use the framework of general $F$-diagrams to study generalized notions of independences and Markov random fields. In the process, we generalized well-known notions like the (dual) total correlation and developed new methods such as subset determination. We were able to apply the general theory to the probabilistic case, and in particular to Kullback-Leibler diagrams on Markov chains. We think it is worthwhile to look into research areas such as machine learning, where graphical models and information functions such as cross-entropy are widespread, and to apply our generalized information-theoretic insights to such settings.

# Appendix

## A Cohomological Characterization of Functions Satisfying the Chain Rule

In this appendix we expand on Remark 4.7 and show how functions $F$ satisfying the chain rule can be classified. We will also interpret that result in terms of the vanishing of a cohomology group of degree 1. In this whole section, let $M$ be a commutative, idempotent monoid acting additively on an abelian group $G$.

**Definition A.1** (Bounded Monoid, Top Element)**.** *$M$ is called* bounded *if it contains a* top element*, i.e., an element $\top \in M$ such that $X \cdot \top = \top$ for all $X \in M$.*

*With the relation $\precsim$ on $M$ defined by $X \precsim Y$ if and only if $X \cdot Y = Y$, we see that a top element is equivalently described as the greatest element in $M$.*

Clearly, top elements are unique.

**Notation A.2.** *We denote by*

$$\mathrm{CR}(M,G) \coloneqq \Big\{ F : M \to G \;\Big|\; \forall X, Y \in M : \;\; F(XY) = F(X) + X.F(Y) \Big\}$$

*the set of functions satisfying the chain rule Equation (5).*

Note that $\mathrm{CR}(M,G)$ is itself an abelian group with addition

$$(F + F')(X) \coloneqq F(X) + F'(X).$$

Additionally, it carries an induced and well-defined additive monoid action $. : M \times \mathrm{CR}(M,G) \to \mathrm{CR}(M,G)$ given by

$$(X.F)(Y) \coloneqq X.F(Y).$$

**Notation A.3.** *Let $X \in M$ be any element. We denote by*

$$G^X \coloneqq \Big\{ g \in G \;\Big|\; X.g = 0 \Big\}$$

*the elements in $G$ that are annihilated by $X$. This is a subgroup of $G$. Furthermore, the action of $M$ on $G$ restricts to a well-defined additive monoid action $M \times G^X \to G^X$.*

**Definition A.4** (Module Homomorphism, Module Isomorphism)**.** *Let $G, H$ both be groups carrying an additive monoid action from $M$, which by abuse of notation we denote with the same symbol:*

$$. : M \times G \to G, \quad . : M \times H \to H.$$

*A function $\Phi : G \to H$ is called a* module homomorphism *if*

- *$\Phi$ is a group homomorphism: $\Phi(g + g') = \Phi(g) + \Phi(g')$ for all $g, g' \in G$;*
- *$\Phi$ commutes with the monoid actions: $\Phi(X.g) = X.\Phi(g)$ for all $X \in M$, $g \in G$.*

*A* module isomorphism *is a bijective module homomorphism.*

The following proposition shows that the functions $F : M \to G$ that satisfy the chain rule are, for bounded $M$ with top element $\top$, essentially the same as the elements in $G$ that are annihilated by $\top$.

**Proposition A.5.** *Let $M$ be a bounded, commutative, idempotent monoid with top element $\top$, $G$ an abelian group, and $. : M \times G \to G$ an additive monoid action. Define the pair of functions*

$$\mathrm{CR}(M,G) \underset{\Psi}{\overset{\Phi}{\rightleftarrows}} G^\top$$

*as follows:*

$$\forall F \in \mathrm{CR}(M,G): \quad \Phi(F) \coloneqq F(\top);$$
$$\forall g \in G^\top: \quad \Psi(g): M \to G, \quad \big[\Psi(g)\big](X) \coloneqq g - X.g.$$

*Then $\Phi$ and $\Psi$ are mutually inverse module isomorphisms.*

*Proof.* For $\Phi$, we need to check well-definedness, i.e., that $F(\top) \in G^\top$ for all $F \in \mathrm{CR}(M,G)$. Indeed, we have

$$F(\top) = F(\top\top) = F(\top) + \top.F(\top)$$

by the chain rule, from which $\top.F(\top) = 0$ follows. It is clear that $\Phi$ is a module homomorphism.

For $\Psi$, we also need to check that it is well-defined, i.e., that $\Psi(g)$ satisfies the chain rule for all $g \in G^\top$. Indeed, we have

$$\begin{aligned}\big[\Psi(g)\big](XY) &= g - (XY).g \\ &= g - X.g + X.g - X.(Y.g) \\ &= \big[\Psi(g)\big](X) + X.\big[g - Y.g\big] \\ &= \big[\Psi(g)\big](X) + X.\big[\Psi(g)\big](Y).\end{aligned}$$

[19]

That $\Psi$ is a module homomorphism is clear.

It remains to show that $\Psi$ and $\Phi$ are inverse to each other. By definition of $G^\top$, we have for all $g \in G^\top$:

$$\big(\Phi \circ \Psi\big)(g) = \Phi\big(\Psi(g)\big) = \big[\Psi(g)\big](\top) = g - \top.g = g.$$

In the other direction, let $F \in \mathrm{CR}(M,G)$ arbitrary. For all $X \in M$, we have $X \cdot \top = \top$ by definition of $\top$ and therefore

$$F(\top) = F(X \cdot \top) = F(X) + X.F(\top)$$

by the chain rule. It follows

$$\Big[\big(\Psi \circ \Phi\big)(F)\Big](X) = \big[\Psi\big(\Phi(F)\big)\big](X) = \big[\Psi\big(F(\top)\big)\big](X) = F(\top) - X.F(\top) = F(X).$$

This means $\big(\Psi \circ \Phi\big)(F) = F$, and we are done. □

**Remark A.6** (Cohomological Interpretation)**.** *For the interested reader, we interpret the preceding result in cohomological terms. Namely, consider the following cochain complex:*

$$G = \mathrm{Maps}(M^0, G) \xrightarrow{\Psi} \mathrm{Maps}\,(M, G) \xrightarrow{\delta} \mathrm{Maps}\,(M^2, G) \longrightarrow \dots$$

*Here, define $\Psi$ as in Proposition A.5 by $\big[\Psi(g)\big](X) \coloneqq g - X.g$ and $\delta$ by*

$$\big[\delta(F)\big](X;Y) \coloneqq X.F(Y) - F(XY) + F(X).$$

*These are Hochschild coboundary maps, as originally defined in Hochschild [1945], and also used for information cohomology in Baudot and Bennequin [2015], Vigneaux [2019]. Now, the fact that $\Psi$ is well-defined in Proposition A.5, i.e. that it only maps to functions satisfying the chain rule,*[20] *can now be expressed equivalently by saying that $\delta \circ \Psi = 0$, a crucial property for cochain complexes. The reason for this is that $\mathrm{Ker}(\delta) = \mathrm{CR}(M,G)$ is precisely the set of functions satisfying the chain rule. This allows to define the first cohomology group of the complex, given by $\mathrm{Ker}(\delta)/\mathrm{Im}(\Psi)$.*

*Now, the fact that $\Psi$ is surjective in the preceding proposition — meaning that it hits every function satisfying the chain rule — can now be expressed with $\mathrm{Im}(\Psi) = \mathrm{Ker}(\delta)$, meaning the*

[19] Actually, the assumption $g \in G^\top$ was not used in this computation. It works for all $g \in G$, which we use in Remark A.6

[20] Which, as we remarked, holds for $\Psi$ defined on all of $G$.

*first cohomology group vanishes:* $\mathrm{Ker}(\delta)/\mathrm{Im}(\Psi) = 0$. *This is shown by explicitly constructing the inverse* $\Phi$.

*Finally, the preceding proposition also shows that* $\Psi$ *is injective when restricting to* $G^\top$.

*Now, consider the case that* $M$ *is the monoid of equivalence classes of random variables on* $\Omega$ *and* $G = \mathrm{Meas}\big(\Delta(\Omega), \mathbb{R}\big)$. *Then the preceding discussion would suggest that Shannon entropy* $I$ *disappears in the first cohomology group. At first sight, one might think this contradicts the fundamental result of information cohomology, which shows that the first cohomology group is non-trivial and generated by Shannon entropy [Baudot and Bennequin, 2015]. The difference is explained by noting that Baudot and Bennequin [2015] also require their cochains to satisfy the* joint locality *property, which means that their cochains of degree zero are necessarily constant.*

## B Conditional Mutual $F$-Independences and $F$-Total Correlation

In this appendix, we provide a characterization of conditional mutual $F$-independences using $F$–total correlation. Different from the characterization using $F$–dual total correlation in Theorem 4.15, this characterization will only work when imposing an additional assumption on $G$, namely that it is torsion-free. The reason is that $F$–total correlation double counts atoms in the $F$-diagram; from a vanishing of the $F$–total correlation, one can then a priori only conclude that a *multiple* of each of the contained atoms disappears, and one needs $G$ to be torsion-free to be able to reduce the coefficient to 1.

As in the main text, we fix a commutative, idempotent monoid $M$ acting additively on an abelian group $G$ and a function $F : M \to G$ satisfying the chain rule Equation (5).

**Proposition B.1.** *Let* $X_1, \dots, X_n \in M$. *Then for all* $\emptyset \neq I \subseteq [n]$, *we have the following identities relating* $F$*–total correlation to higher* $F$*-interactions:*

1. $TC_F\big(\,;_{i\in I} X_i\big) = \sum_{\emptyset\neq L\subseteq I} \big(|L|-1\big)\cdot X_{I\setminus L}.F\big(\,;_{l\in L} X_l\big)$;
2. $\big(|I|-1\big)\cdot F\big(\,;_{i\in I} X_i\big) = \sum_{\emptyset\neq L\subseteq I}(-1)^{|I|-|L|}\cdot X_{I\setminus L}.TC_F\big(\,;_{l\in L} X_l\big)$.

*Proof.* To prove part 1, let the $X_i$, $i \in I$ be the full set of elements in Theorem 2.8. Then we obtain

$$\begin{aligned}
TC_F\big(\,;_{i\in I} X_i\big) &= \sum_{i\in I} F(X_i) - F(X_I) \\
&= \sum_{i\in I} \widetilde{F}\big(\widetilde{X}_i\big) - \widetilde{F}\big(\widetilde{X}_I\big) \\
&= \sum_{i\in I} \sum_{L\subseteq I, i\in L} \widetilde{F}(p_L) - \sum_{\emptyset\neq L\subseteq I} \widetilde{F}(p_L) \\
&= \sum_{\emptyset\neq L\subseteq I} \sum_{i\in L} \widetilde{F}(p_L) - \sum_{\emptyset\neq L\subseteq I} \widetilde{F}(p_L) \\
&= \sum_{\emptyset\neq L\subseteq I} \big(|L|-1\big)\cdot \widetilde{F}(p_L).
\end{aligned}$$

Now, note that $\widetilde{F}(p_L) = X_{I\setminus L}.F\big(\,;_{l\in L} X_l\big)$ by Lemma 4.4. The result follows.

To prove part 2, we evaluate the right-hand-side using part 1 on each summand:

$$\begin{aligned}
&\sum_{\emptyset\neq L\subseteq I} (-1)^{|I|-|L|}\cdot X_{I\setminus L}.TC_F(\,;_{l\in L}X_l)\\
&\qquad= \sum_{\emptyset\neq L\subseteq I} (-1)^{|I|-|L|}\cdot X_{I\setminus L}.\left(\sum_{\emptyset\neq K\subseteq L}\big(|K|-1\big)\cdot X_{L\setminus K}.F(\,;_{k\in K}X_k)\right)\\
&\qquad= \sum_{\emptyset\neq L\subseteq I}\sum_{\emptyset\neq K\subseteq L} (-1)^{|I|-|L|}\cdot\big(|K|-1\big)\cdot X_{I\setminus K}.F(\,;_{k\in K}X_k)\\
&\qquad= \sum_{\emptyset\neq K\subseteq I} (-1)^{|I|}\cdot\big(|K|-1\big)\cdot\left(\sum_{L:\ K\subseteq L\subseteq I}(-1)^{|L|}\right)\cdot X_{I\setminus K}.F(\,;_{k\in K}X_k).\\
&\qquad= \sum_{\emptyset\neq K\subseteq I} (-1)^{|I|}\cdot\big(|K|-1\big)\cdot(-1)^{|K|}\cdot\mathbb{1}_{I=K}\cdot X_{I\setminus K}.F(\,;_{k\in K}X_k)\\
&\qquad= (-1)^{2\cdot|I|}\cdot\big(|I|-1\big)\cdot X_{I\setminus I}.F(\,;_{i\in I}X_i)\\
&\qquad= \big(|I|-1\big)\cdot F(\,;_{i\in I}X_i).
\end{aligned}$$

In the fourth step, we used Lemma 4.6. □

**Definition B.2** (Torsion-Free Abelian Group)**.** *Let $G$ be an abelian group. Then $G$ is called* torsion-free *if $0\in G$ is the only element of finite order. In other words, for all $0\neq g\in G$ and all $0<k\in\mathbb{N}$, we have $k\cdot g\neq 0$. Equivalently, for all $k\in\mathbb{N}$ and $g\in G$, if $k\cdot g=0$ then $k=0$ or $g=0$.*

**Example B.3.** *The groups $\mathbb{Z}$, $\mathbb{R}$, $\mathrm{Meas}\big(\Delta(\Omega),\mathbb{R}\big)$ and all other groups encountered in Section 5 are torsion-free. $\mathbb{Z}/k\mathbb{Z}$ for $k\geq 1$ is a prototypical example of an abelian group with torsion: $k\cdot 1=0$.*

**Theorem B.4.** *Let $M$ be a commutative, idempotent monoid acting additively on a* torsion-free *abelian group $G$. Let $F: M\to G$ be a function satisfying the chain rule Equation (5). Let $X_1,\dots,X_n,Y\in M$. Then the following properties are equivalent:*

1. $F: \coprod_{i=1}^n X_i \mid Y$;
2. $Y.TC_F(\,;_{i\in[n]}X_i)=0$;
3. $(YX_{[n]\setminus I}).TC_F(\,;_{i\in I}X_i)=0$ *for all* $\emptyset\neq I\subseteq[n]$.

*Proof.* Assume 1. To prove 2, we use that by Proposition 3.2, we have $X_i \underset{F}{\perp\!\!\!\perp} X_{[i-1]} \mid Y$ for all $i=1,\dots,n$. By Proposition 4.13, this implies $Y.F\big(X_{[i]}\big)=Y.F\big(X_{[i-1]}\big)+Y.F(X_i)$. Inductively, we obtain $Y.F\big(X_{[n]}\big)=\sum_{i=1}^n Y.F(X_i)$ and therefore $Y.TC_F(\,;_{i\in[n]}X_i)=0$.

Now assume 2. To prove 3, we note

$$\begin{aligned}
0 &= X_{[n]\setminus I}.\Big(Y.TC_F(\,;_{i\in[n]}X_i)\Big)\\
&= Y.\left(X_{[n]\setminus I}.\left(\sum_{i=1}^n F(X_i)-F\big(X_{[n]}\big)\right)\right)\\
&= Y.\left(\sum_{i=1}^n X_{[n]\setminus I}.F(X_i)-X_{[n]\setminus I}.F\big(X_{[n]}\big)\right)\\
&\overset{(\star)}{=} Y.\left(\sum_{i\in I} X_{[n]\setminus I}.F(X_i)-X_{[n]\setminus I}.F(X_I)\right)\\
&= \big(YX_{[n]\setminus I}\big).\left(\sum_{i\in I}F(X_i)-F(X_I)\right)\\
&= \big(YX_{[n]\setminus I}\big).TC_F(\,;_{i\in I}X_i).
\end{aligned}$$

Step $(\star)$ can easily be seen using Hu's Theorem 2.8. That was to show.

Finally, assume 3. For proving 1, we can alternatively prove the equivalent property 3 of Theorem 4.15. We note that for all $\emptyset \neq I \subseteq [n]$, Proposition B.1 implies:

$$\begin{aligned}
\big(|I|-1\big) \cdot \big(YX_{[n]\setminus I}\big).F\big(\,;_{i\in I}X_i\big) &= \big(YX_{[n]\setminus I}\big).\Big(\big(|I|-1\big) \cdot F\big(\,;_{i\in I}X_i\big)\Big) \\
&= \big(YX_{[n]\setminus I}\big).\left(\sum_{\emptyset\neq L\subseteq I} (-1)^{|I|-|L|} \cdot X_{I\setminus L}.TC_F\big(\,;_{l\in L}X_l\big)\right) \\
&= \sum_{\emptyset\neq L\subseteq I} (-1)^{|I|-|L|} \cdot \big(YX_{[n]\setminus L}\big).TC_F\big(\,;_{l\in L}X_l\big) \\
&= 0.
\end{aligned}$$

If $|I| = 1$ then the factor $|I| - 1$ vanishes and what we showed is vacuous. If $|I| \geq 2$, then what we showed implies

$$\big(YX_{[n]\setminus I}\big).F\big(\,;_{i\in I}X_i\big) = 0$$

since $G$ is torsion-free, and we are done. □

We now show in an example that we cannot omit the assumption that $G$ is torsion-free in the preceding theorem. More precisely, if $G$ is not torsion-free, then property 2 does not necessarily imply property 1 anymore.

**Example B.5.** *Let $M = \big(\mathbb{Z}/2\mathbb{Z}, \cdot\big)$, where $\mathbb{Z}/2\mathbb{Z} = \{0, 1\}$. This is a commutative, idempotent monoid with the following multiplication rules:*

$$1 \cdot 1 = 1; \quad 1 \cdot 0 = 0; \quad 0 \cdot 1 = 0; \quad 0 \cdot 0 = 0.$$

*Furthermore, define $G \coloneqq \big(\mathbb{Z}/2\mathbb{Z}, +\big)$, which is an abelian group with the following addition rules:*

$$0 + 0 = 0; \quad 0 + 1 = 1; \quad 1 + 0 = 1; \quad 1 + 1 = 0.$$

*The last rule implies that $G$ has torsion. We define the action of $M$ on $G$ by $X.g \coloneqq X \cdot g$ for all $X \in M$ and $g \in G$. That is, the action is simply the usual multiplication in the ring $\big(\mathbb{Z}/2\mathbb{Z}, +, \cdot\big)$ and therefore clearly an additive monoid action. Finally, define*

$$F : M \to G, \quad F(X) \coloneqq 1 - X.$$

*We have*

$$F(XY) = 1 - XY = 1 - X + X - XY = 1 - X + X \cdot (1 - Y) = F(X) + X.F(Y),$$

*implying that $F$ satisfies the chain rule. Thus, $F$ satisfies all the properties of Hu's Theorem 2.8.*

*Now, set $X_1 \coloneqq X_2 \coloneqq X_3 \coloneqq 0 \in M$. Then we have*

$$TC_F(X_1; X_2; X_3) = F(0) + F(0) + F(0) - F(0 \cdot 0 \cdot 0) = 1 + 1 = 0.$$

*However, we have*

$$F(X_3; X_1X_2) = F(0; 0) = F(0) = 1.$$

*This means that the $F$-independence $\big(X_1X_2\big) \perp\!\!\!\perp_F X_3$ does* not *hold, implying that $X_1, X_2, X_3$ are* not *mutually independent. Thus we cannot omit the assumption of torsion-freeness in the Theorem.*

*Note that in light of Proposition A.5, the construction of $F$ can be explained as follows: $M$ is a bounded monoid with top element $0 \in M$. Then, $G^\top = G$, and so $1 \in G$ is one of the annihilated elements. We then see:*

$$\big[\Psi(1)\big](X) = 1 - X.1 = 1 - X = F(X),$$

*showing that $F = \Psi(1)$. The only other function satisfying the chain rule, $\Psi(0)$, is trivial, and does therefore not give rise to a counter example.*

# C General Consequences of Section 4

In this appendix, we take a look at Yeung et al. [2019] and explain which results generalize to our setting. As it turns out, all these results generalize without problem with the same proofs, or otherwise cannot even be formulated in our case since the formulations are based on inequalities, which does not make sense in the context of general abelian groups.

Fix a commutative, idempotent monoid $M$ acting on an abelian group $G$ and a function $F : M \to G$ satisfying the chain rule. We fix elements $X_1, \dots, X_n \in M$, giving rise to $\widetilde{F} : 2^{\widetilde{X}} \to G$ by Hu's Theorem 2.8, and a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = [n]$.

### Smallest Graph Representations

**Terminology C.1.** *If $X_1, \dots, X_n$ form an $F$-Markov random field with respect to $\mathcal{G}$, then we also say that $\mathcal{G}$ is a (graph)* representation *for $X_1, \dots, X_n$ with respect to $F$.*

**Definition C.2** (Subgraph)**.** *If $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ is a second graph with $\mathcal{V}' \subseteq \mathcal{V}$ and $\mathcal{E}' \subseteq \mathcal{E}$, then we say that $\mathcal{G}'$ is a* subgraph *of $\mathcal{G}$.*

**Definition C.3** (Smallest Graph Representation)**.** *$\mathcal{G}'$ is said to be the* smallest graph representation *for $X_1, \dots, X_n$ (with respect to $\perp\!\!\!\perp_F$) if $\mathcal{G}'$ is a representation for $X_1, \dots, X_n$ and a subgraph of all other representations for $X_1, \dots, X_n$.*

**Theorem C.4** (Smallest Graph Representation)**.** *Define $\widehat{\mathcal{G}} := \big([n], \widehat{\mathcal{E}}\big)$ with $\{i, j\} \in \widehat{\mathcal{E}}$ if and only if*

$$X_{[n]\setminus\{i,j\}}.F(X_i; X_j) = \widetilde{F}\big(p_{\{i,j\}}\big) \neq 0. \tag{43}$$

*If $X_1, \dots, X_n$ has a smallest graph representation, then it equals $\widehat{\mathcal{G}}$.*

*Proof.* This is simply Yeung et al. [2019], Theorem 3, generalized to our setting. Note that the first equality in Equation (43) is simply Lemma 4.4 and always holds. The original proof can be copied word for word. Only once, an "inequality sign" needs to be replaced by an "unequal sign". □

### Smallest Graphs for Subfields of Markov Random Fields

**Definition C.5** (Marginalization of Graph)**.** *Let $\mathcal{V}' \subseteq \mathcal{V}$. Then the Marginalization of $\mathcal{G}$ for $\mathcal{V}'$ is defined as $\mathcal{G}^*(\mathcal{V}') := (\mathcal{V}', \mathcal{E}')$ with $\{i, j\} \in \mathcal{E}'$ if and only if there is a walk from $i$ to $j$ in $\mathcal{G}$ with all intermediate vertices in $\mathcal{V} \setminus \mathcal{V}'$.*

**Notation C.6.** *Let $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ be a graph with $\mathcal{V}' \subseteq \mathcal{V} = [n]$. Then we write $\mathcal{G} \Longrightarrow \mathcal{G}'$ if for all $Y_1, \dots, Y_n$ that form an $F$-Markov random field with respect to $\mathcal{G}$, $Y_i, i \in \mathcal{V}'$ form an $F$-Markov random field with respect to $\mathcal{G}'$. Note that "$\mathcal{G} \Longrightarrow \mathcal{G}'$" is not a statement about graphs alone, as it depends on $F : M \to G$.*

**Theorem C.7** (Graphs for Subfields)**.** *Assume that there is at least one element $Z \in M$ with $F(Z) \neq 0$. Let $\mathcal{V}' \subseteq \mathcal{V}$ be a subset. Then $\mathcal{G}^*(\mathcal{V}')$ is the smallest graph $\mathcal{G}'$ such that $\mathcal{G} \Longrightarrow \mathcal{G}'$.*

*Proof.* This is precisely Yeung et al. [2019], Theorem 8. The proof is exactly the same, except that the condition $I(Z) > 0$ from the original paper is replaced by $F(Z) \neq 0$, and the constant random variable is replaced by $\mathbf{1} \in M$. □

**Remark C.8.** *If we drop the condition that there is $Z \in M$ with $F(Z) \neq 0$, then the conclusion is wrong. Indeed, if $F$ is trivial, then all $F$-independences always hold, meaning that the graph $\mathcal{G}' := (\mathcal{V}', \emptyset)$ with empty edge set would be the smallest graph with $\mathcal{G} \Longrightarrow \mathcal{G}'$. The existence of $Z$ with $F(Z) \neq 0$ ensures that we can for all edges in $\mathcal{G}^*\big(\mathcal{V}'\big)$ construct an $F$-Markov random field for $\mathcal{G}$ that is* faithful *to that edge by having a corresponding dependence.*

### Rewriting the Values of Atoms in $F$-Markov Random Fields

Section 5 in Yeung et al. [2019] discusses a characterization for forests of paths. Namely, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a forest of paths if and only if for all probability mass functions $P$ and all random variables $X_1, \dots, X_n$ that form a Markov random field with respect to $\mathcal{G}$ and $P$, the corresponding $I$-diagram is nonnegative. These results cannot be stated in our generalized setting since $F$ takes values in an unspecified abelian group $G$ that may not have an associated order relation to begin with. However, that section still contains an interesting result worth stating in our generalized setting:

**Theorem C.9** (Atoms in Markov Random Fields)**.** *Assume that $X_1, \dots, X_n$ form an $F$-Markov random field with respect to $\mathcal{G}$. Let $p_{\mathcal{I}}$ be a connected atom, where $\mathcal{I} \subseteq [n]$ has at least two elements. Define*

$$\mathcal{B} \coloneqq \Big\{ i \in \mathcal{I} \mid p_{\mathcal{I} \setminus i} \text{ is connected} \Big\}.$$

*Then one has*

$$\widetilde{F}(p_{\mathcal{I}}) = X_{\mathcal{V} \setminus \mathcal{I}}.F\big(\,;_{i \in \mathcal{I}} X_i\big) = X_{\mathcal{V} \setminus \mathcal{I}}.F\big(\,;_{i \in \mathcal{B}} X_i\big).$$

*Proof.* Note that the first equality is simply Lemma 4.4. The second equality is Yeung et al. [2019], Theorem 11. The original proof generalizes word for word to our setting. □

**Remark C.10.** *The preceding theorem deals with connected atoms. For disconnected atoms $p_{\mathcal{I}}$, Theorem 2.21 shows $\widetilde{F}(p_{\mathcal{I}}) = 0$.*

*Furthermore, note that Theorem C.9 generalizes a special case of Corollary 4.25: that corollary implies that for an $F$-Markov chain $X_1, \dots, X_n$ and an "interval" $I = [i : j] \subseteq [n]$ for $i < j$, one has*

$$X_{[n] \setminus I}.F\big(\,;_{i \in I} X_i\big) = X_{[n] \setminus I}.F(X_i; X_j).$$

### Trees and Drawing $F$-Diagrams for $F$-Markov Random Fields

The remaining results in Yeung et al. [2019] are essentially graph theoretic in nature and not dependent on $F$. In Theorem 9, they discuss when the marginalization of a graph is a tree. Finally, in Section 6, they discuss an algorithm for drawing an $I$-diagram, with two requirements relating to the graph $\mathcal{G}$: connected atoms are depicted as nonempty and disconnected atoms as empty. These diagrams are then suitable to express any $F$-diagram of any $F$-Markov random field corresponding to $\mathcal{G}$, as Theorem 2.21 shows. For the case of path-shaped graphs corresponding to Markov chains, we saw examples of such depictions in Figure 8. The algorithm in Section 6 of Yeung et al. [2019] works for any graph $\mathcal{G}$.

## D Slices of $I$-Diagrams

Our generalizations of the main results from Yeung et al. [2002], Theorems 4.21 and 2.21, work in the framework of $F$-independence, where $F$ satisfies the full set of assumptions in Hu's Theorem 2.8. In contrast, the original results were formulated for the $P$-independence with respect to a fixed probability mass function $P$, thus engaging only with one "slice" of the $I$-diagrams, as we explained in Equation (10). A priori, it may seem unclear whether these original results can actually be deduced from our supposed generalizations. In this section, we prove that this is indeed possible.

**Theorem D.1** (Characterization of $P$-FCMIs)**.** *Let $I : M \to \mathrm{Meas}\big(\Delta(\Omega), \mathbb{R}\big)$ be the Shannon entropy function. Let $X_1, \dots, X_n$ be random variables on $\Omega$ and $P \in \Delta(\Omega)$ a probability mass function, giving rise to $\widetilde{I}^P : 2^{\widetilde{X}} \to \mathbb{R}$ by Equation (10). Let $K = \big(J, L_i, 1 \leq i \leq q\big)$ be a conditional partition of $[n]$ with $q \geq 2$. Then the following two statements are equivalent:*

- *$K$ induces a $P$-FCMI: $P : \coprod_{i=1}^{q} X_{L_i} \mid X_J$;*
- *For all $p_W \in \mathrm{Im}(K)$: $\widetilde{I}^P(p_W) = 0$.*

*Proof.* Let $\mathcal{R}$ be the following property on elements $P' \in \Delta(\Omega)$:

$$\mathcal{R}(P') \quad :\iff \quad P' : \coprod_{i=1}^{q} X_{L_i} \mid X_J.$$

Then by Proposition 5.5, and using that it is also well-defined and measurable, this is stable property. Consequently, we can define the restricted entropy function $I_{\mathcal{R}} : M_{\mathcal{R}} \to \mathrm{Meas}\big(\Delta(\Omega)_{\mathcal{R}}, \mathbb{R}\big)$, which satisfies the chain rule according to Proposition 5.11. Now, we claim that the following conditional mutual independence holds:

$$I_{\mathcal{R}} : \coprod_{i=1}^{q} X_{L_i} \mid X_J. \tag{44}$$

By definition, conditional mutual independences are given by a set of pairwise independences, and these in turn by the vanishing of conditional mutual information. Thus, the claim reduces to the following: for all $i \in [q]$, we have

$$X_J.I_{\mathcal{R}}\big(X_{L_i}; X_{L_{\setminus i}}\big) = 0,$$

where $L_{\setminus i} \coloneqq \bigcup_{k \neq i} L_k$. But this is clear since for all $P'$ with $\mathcal{R}(P')$, we have $X_{L_i} \perp\!\!\!\perp_{P'} X_{L_{\setminus i}} \mid X_J$ and thus, as is well known, the vanishing of the corresponding conditional mutual information follows:

$$\Big[X_J.I_{\mathcal{R}}\big(X_{L_i}; X_{L_{\setminus i}}\big)\Big](P') = \Big[X_J.I\big(X_{L_i}; X_{L_{\setminus i}}\big)\Big](P') = 0.$$

This proves the claim, Equation (44).

Now, assume that the first statement holds, i.e., $P : \coprod_{i=1}^{q} X_{L_i} \mid X_J$. Then $\mathcal{R}(P)$ holds. From Equation (44), we obtain by Theorem 4.21 that for all $p_W \in \mathrm{Im}(K)$: $\widetilde{I_{\mathcal{R}}}(p_W) = 0$. It follows

$$\widetilde{I}^P(p_W) = \big[\widetilde{I}(p_W)\big](P) = \big[\widetilde{I_{\mathcal{R}}}(p_W)\big](P) = 0,$$

which proves one direction.

For the other direction, assume that $\widetilde{I}^P(p_W) = 0$ for all $p_W \in \mathrm{Im}(K)$. We want to show $P : \coprod_{i=1}^{q} X_{L_i} \mid X_J$. As is well-known, this amounts to showing the following for all $i \in [q]$:

$$\Big[X_J.I\big(X_{L_i}; X_{L_{\setminus i}}\big)\Big](P) = 0.$$

By Hu's Theorem 2.8, we thus need to show $\widetilde{I}^P\big(\widetilde{X}_{L_i} \cap \widetilde{X}_{L_{\setminus i}} \setminus \widetilde{X}_J\big) = 0$. But note that, clearly, we have $\widetilde{X}_{L_i} \cap \widetilde{X}_{L_{\setminus i}} \setminus \widetilde{X}_J \subseteq \mathrm{Im}(K)$. Thus, the claim follows from the assumption that $\widetilde{I}^P(p_W) = 0$ for all $p_W \in \mathrm{Im}(K)$, together with the fact that $\widetilde{I}^P$ is a signed measure and therefore additive over disjoint unions. □

Now we prove Yeung's main result, Theorem 2.15, from Yeung et al. [2002]:

*Proof of Theorem 2.15.* Let the property $\mathcal{R}$ be given by $\mathcal{R} \coloneqq \mathcal{MRF}\big(\mathcal{G}; X_1, \dots, X_n\big)$, see the discussion after Equation (39). Note that in our application of that definition, we have $r = 0$. Now, assume the first statement, i.e., $\mathcal{R}(P)$ holds. We also know from Theorem 5.12 that $X_1, \dots, X_n$ form an $I_{\mathcal{R}}$-Markov random field with respect to $\mathcal{G}$. Then Theorem 2.21 implies that $\widetilde{I_{\mathcal{R}}}(p_{\mathcal{W}}) = 0$ for all disconnected atoms $p_{\mathcal{W}}$. For these atoms, we then obtain

$$\widetilde{I}^P(p_{\mathcal{W}}) = \Big[\widetilde{I}(p_{\mathcal{W}})\Big](P) = \Big[\widetilde{I_{\mathcal{R}}}(p_{\mathcal{W}})\Big](P) = 0,$$

which proves one direction.

For the other direction, assume that $\widetilde{I}^P(p_{\mathcal{W}}) = 0$ for all disconnected atoms $p_{\mathcal{W}}$. By Proposition 3.5, we need to show that $X_1, \dots, X_n$ satisfies the cutset property with respect to $\mathcal{G}$ and $\perp\!\!\!\perp_P$. Thus, let $\mathcal{U}$ be a cutset of $\mathcal{G}$ and $K \coloneqq \big(\mathcal{U}, \mathcal{V}_i(\mathcal{U}), 1 \leq i \leq s(\mathcal{U})\big)$ the corresponding conditional partition. We need to show $P : \coprod_{i=1}^{s(\mathcal{U})} X_{\mathcal{V}_i(\mathcal{U})} \mid X_{\mathcal{U}}$. By Theorem D.1, we need to show that $\widetilde{I}^P(p_{\mathcal{W}}) = 0$ for all atoms $p_{\mathcal{W}} \in \mathrm{Im}(K)$. All such $p_{\mathcal{W}}$ are disconnected by Lemma 4.23, and so the result follows from the assumption. □

## References


Shun'ichi Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711, 2001. https://doi.org/10.1109/18.930911.

Pierre Baudot and Daniel Bennequin. The homological nature of entropy. *Entropy*, 17(5):3253–3318, 2015. ISSN 1099-4300. https://doi.org/10.3390/e17053253. URL https://www.mdpi.com/1099-4300/17/5/3253.

Pierre Baudot, Monica Tapia, Daniel Bennequin, and Jean-Marc Goaillard. Topological information data analysis. *Entropy*, 21(9), 2019. ISSN 1099-4300. https://doi.org/10.3390/e21090869. URL https://www.mdpi.com/1099-4300/21/9/869.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 1st edition, 2007. ISBN 0387310738.

Christopher M. Bishop and Hugh Bishop. *Deep Learning: Foundations and Concepts*. Springer, 1st edition, 2023. ISBN 9783031454684. https://doi.org/10.1007/978-3-031-45468-4.

Gregory. J. Chaitin. *Algorithmic Information Theory*. Cambridge University Press, Oct. 1987. ISBN 9780521343060. https://doi.org/10.1017/CBO9780511608858.

Simosa Cocco and Rémi Monasson. Adaptive cluster expansion for the inverse Ising problem: Convergence, algorithm and tests. *Journal of Statistical Physics*, 147(2):252–314, Apr. 2012. https://doi.org/10.1007/s10955-012-0463-4.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006. ISBN 9780471241959.

Alexander P. Dawid. Separoids: A mathematical framework for conditional independence and irrelevance. *Annals of Mathematics and Artificial Intelligence*, 32(1):335–372, 2001. ISSN 1573-7470. https://doi.org/10.1023/A:1016734104787.

Patrick Forré. Transitional conditional independence. arXiv preprint arXiv:2104.11547, 2021. URL https://arxiv.org/abs/2104.11547.

Michèle Giry. A categorical approach to probability theory. In B. Banaschewski, editor, *Categorical Aspects of Topology and Analysis*, pages 68–85. Springer Berlin Heidelberg, 1982. ISBN 9783540390411. https://doi.org/10.1007/BFb0092872.

John M. Hammersley and Peter Clifford. Markov fields on finite graphs and lattices, 1971. URL https://www.statslab.cam.ac.uk/~grg/books/hammfest/hamm-cliff.pdf. Unpublished Manuscript.

Te S. Han. Nonnegative entropy measures of multivariate symmetric correlations. *Information and Control*, 36(2):133–156, 1978. ISSN 0019-9958. https://doi.org/10.1016/S0019-9958(78)90275-9.

Gerhard Hochschild. On the cohomology groups of an associative algebra. *Annals of Mathematics*, 46(1):58–67, 1945. ISSN 0003-486X. https://doi.org/10.2307/1969145.

Kuo T. Hu. On the amount of information. *Theory of Probability and Its Applications*, 7(4):439–447, 1962. https://doi.org/10.1137/1107041.

Tsutomu Kawabata and Raymond W. Yeung. The structure of the I-measure of a Markov chain. *IEEE Transactions on Information Theory*, 38(3):1146–1149, 1992. https://doi.org/10.1109/18.135658.

Leon Lang, Pierre Baudot, Rick Quax, and Patrick Forré. Information decomposition diagrams applied beyond Shannon entropy: A generalization of Hu's Theorem. *Compositionality*, Volume 7:1, Jan 2025. ISSN 2631-4444. https://doi.org/10.46298/compositionality-7-1. URL https://compositionality.episciences.org/14181.

Ming Li and Paul Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Cham, 4th edition, 2019. ISBN 9783030112974. https://doi.org/10.1007/978-3-030-11298-1.

Anibal M. Medina-Mardones, Fernando E. Rosas, Sebastián E. Rodríguez, and Rodrigo Cofré. Hyperharmonic analysis for the study of high-order information-theoretic signals. *Journal of Physics: Complexity*, 2(3), 2021. ISSN 2632-072X. https://doi.org/10.1088/2632-072X/abf231.

Judea Pearl. Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*, pages 329–334, Irvine, CA, 1985. URL http://ftp.cs.ucla.edu/tech-report/198_-reports/850017.pdf.

Christopher Preston. *Random Fields*, volume 534 of *Lecture Notes in Mathematics*. Springer, 1976. ISBN 9783540078524.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, July 2021. URL https://proceedings.mlr.press/v139/ramesh21a.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022. https://doi.org/10.1109/CVPR52688.2022.01042.

Fernando E. Rosas, Pedro A.M. Mediano, Michael Gastpar, and Henrik J. Jensen. Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Physical Review E*, 100(3):1–17, 2019. ISSN 2470-0053. https://doi.org/10.1103/PhysRevE.100.032305.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 36479–36494. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, July 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Frank L. Spitzer. *Random Fields and Interacting Particle Systems*. Mathematical Association of America, 1971.

Bastian Steudel, Dominik Janzing, and Bernhard Schölkopf. Causal Markov condition for submodular information measures. *Conference on Learning Theory*, pages 464–476, 2010. URL https://www.learningtheory.org/colt2010/conference-website/papers/COLT2010proceedings.pdf.

Juan P. Vigneaux. *Topology of statistical systems : A cohomological approach to information theory*. PhD dissertation, Université Sorbonne Paris Cité, June 2019. URL https://theses.hal.science/tel-02951504.

Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4(1):66–82, 1960. https://doi.org/10.1147/rd.41.0066.

Raymond W. Yeung. A new outlook on Shannon's information measures. *IEEE Transactions on Information Theory*, 37(3):466–474, 1991. https://doi.org/10.1109/18.79902.

Raymond W. Yeung. *Information Theory and Network Coding*, volume 46. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387792333.

Raymond W. Yeung, Tony T. Lee, and Zhongxing Ye. Information-theoretic characterizations of conditional mutual independence and Markov random fields. *IEEE Transactions on Information Theory*, 48(7):1996–2011, 2002. ISSN 0018-9448. https://doi.org/10.1109/TIT.2002.1013139.

Raymond W. Yeung, Ali Al-Bashabsheh, Chao Chen, Qi Chen, and Pierre Moulin. On information-theoretic characterizations of markov random fields and subfields. *IEEE Transactions on Information Theory*, 65(3):1493–1511, 2019. ISSN 00189448. https://doi.org/10.1109/TIT.2018.2866564.