Article

Not peer-reviewed version

Probability via Expectation Measures

Peter Harremoës *

Posted Date: 24 October 2024

doi: 10.20944/preprints202410.1897.v1

Keywords: Category; double slit experiment; expectation measure; extended probabilistic power domain; expected value; Gaussian approximation; information divergence; monad; point process; Poisson distribution; Poisson point process; quantum information theory; thinning; valuation

 doi:10.20944/preprints202410.1897.v1

Article

Probability via Expectation Measures

Peter Harremoës †

Niels Brock, Copenhagen Business College; harremoes@ieee.org
† Current address: Nørre Voldgade 34, Copenhagen, Denmark.

Abstract: Since the seminal work of Kolmogorov, probability theory has been based on measure theory, where the central components are so-called probability measures, defined as measures with total mass equal to 1. In Kolmogorov's theory, a probability measure is used to model an experiment with a single outcome that will belong to exactly one out of several disjoint sets. In this paper, we present a different basic model where an experiment results in a multiset, i.e., for each of the disjoint sets we get the number of observations in the set. This new framework is consistent with Kolmogorov's theory, but the theory focuses on expected values rather than probabilities. We present examples from testing Goodness-of-Fit, Bayesian statistics, and quantum theory, where the shifted focus gives new insight or better performance. We also provide several new theorems that address some problems related to the change in focus.

Keywords: category; double slit experiment; expectation measure; extended probabilistic power domain; expected value; Gaussian approximation; information divergence; monad; point process; Poisson distribution; Poisson point process; quantum information theory; thinning; valuation

MSC: 60A05, 60G55

---

1. Introduction

Throughout the history of probability theory, some mathematicians have focused on probabilities, and others have focused on expectations. In his seminal paper from 1933, Kolmogorov focused on probabilities [1], but in the present paper, we will develop *expectation theory* as an alternative to probability theory, focusing on expectations. Expectation theory has been developed to handle technical problems in several applications in information theory, statistics (both frequentist and Bayesian), quantum information theory, and probability theory itself. We will present the basic definitions of our theory and provide interpretations of the core concepts. Standard probability theory, as developed by Kolmogorov, can be embedded in the present theory. Similarly, our theory can be embedded into standard probability theory. The focus is on discrete measures to keep this paper at a moderate length. Since there is no inconsistency between the two theories, there are many cases where expectation theory, as developed here for discrete measures, can also be used for more general measures. Some readers may be more interested in theory, and others may be more interested in applications. The paper is written hoping that the different sections can be read quite independently.

1.1. Organization of the Paper

In Section 2 we point out some significant ideas related to the work of Kolmogorov and we will review some later developments that are important for our topic. The notion of a probability monad will be presented in Section 2.3. This approach allows us to focus on which basic operations are needed to define a well-behaved class of models.

The theoretical framework we will develop may be viewed as a part of the theory of point processes. Usually, the topic of point processes is considered an advanced part of probability theory. Still, we will consider some aspects of the theory of point processes as quite fundamental for modeling randomness. In Section 2.4 we will give a short overview of the relevant concepts of the theory of point processes. For readers familiar with the theory of point processes, it puts our contribution in a well-known context. It will also provide a framework to ensure the rest of the paper provides a consistent theory. In the subsequent sections, our discussions will often focus on finite samples, but the

 doi:10.20944/preprints202410.1897.v1

conclusions will also hold for more general samples, which is easy to see with some general knowledge of point processes.

In Section 3 we develop the theory of finite empirical measures. Such finite empirical measures are formally equivalent to multisets. Some basic properties of empirical measures are established. It is pointed out that many problems in information theory can be formulated for empirical measures without reference to randomness.

In Section 4 we introduce finite expectation measures. In Section 4.3 we introduce the Poisson interpretation that allows us to translate between results for expectation measures and results for Poisson point processes with probability measures. The cost of this translation is that the outcome space of the process is infinite, even if the expectation measure is finite. The Poisson interpretation enables us to give probabilistic interpretations of conditioning and independence for general measures, as discussed in Sections 4.4 and 4.5. In [2] it was demonstrated that the reverse information projection of a probability distribution into a convex set of probability distributions may not be a probability distribution. This has been an important motivation for studying information divergence for general measures, as done in Section 4.6.

In Section 5, we will provide examples of how the present theory gives alternative interpretations and improved results to some familiar problems in decision theory, Bayesian statistics, Testing Goodness-of-Fit, and information theory.

We end the paper with a short conclusion, including a list of concepts in probability theory and the corresponding concepts in expectation theory.

1.2. Terminology

A measure with a total mass of 1 is usually called a probability measure or a normalized measure. We will deviate from this terminology and use the term *unital measure* for a measure with total mass 1. The term *normalized measure* will only be used when a unital measure is the result of dividing a finite measure by the total mass of the measure. We will reserve the word *probability measure* to situations where the weights of a unital measure are used to quantify uncertainty, and it is known that precisely one observation will be made and one can decide which event the observation belongs to in a system of mutually exclusive events that cover the whole outcome space. Similarly, we will talk about an *expectation measure* if our interpretation of its values are given in terms of expected values of some random variables. Other classes of measures are *coding measures* that are used in information theory and *mixing measures* that are unital measures used for barycentric decompositions of points in convex sets.

In standard probability theory, the probability measures lives on a space often called a sample space, but we will use the alternative term, *an outcome space*. The word *sample* will be used informally about the result of a sampling process. The result of a point process will be called *an instance* of the process.

2. Probability Theory Since Kolmogorov

2.1. Kolmogorov's Contribution

The modern foundation of probability theory is due to Kolmogorov. He contributed in many ways, and here we shall only focus on the aspects that are most relevant for the present paper. His 1933 paper [1] was in line with two ideas earlier mathematicians developed.

The first idea is symbolic logic, as developed by G. Boole. In this approach to logic, the propositions form a Boolean algebra. A *truth assignment function* is a binary function that assigns one of the values 0 (false) and 1 (true) to any proposition consistently. In particular, if A is a proposition, either A is assigned the value 1 or $\neg A$ is assigned the value 1. Kolmogorov's work may be seen as an extension where statements are replaced by events, i.e., measurable sets, and the events A and $\complement A$ are assigned probabilities in $[0,1]$ in such a way that $P(A) + P(\complement A) = 1$. Thus, probability theory can be

 doi:10.20944/preprints202410.1897.v1

described as an extension of logic where the functions can take values in $[0,1]$ rather than values in $\{0,1\}$. Such extensions have been formalized as monads to be discussed in Section 2.3, but the theory of monads was only developed much later as part of category theory. See [3,4] for a general discussion of probability monads.

Lebesgue's theory of measures inspired the second main idea in Kolmogorov's 1933 paper. Measure theory was used to extend the previous definitions of integrals. The basic idea is that a measure is defined on a set of measurable sets. Such a measure should be countable additive like the notion of an area. Measure theory has been beneficial for the theory of integration, and in particular, it leads to nice convergence theorems like Lebesgue's theorem on dominated convergence. Kolmogorov used measure theory to allow similar general convergence results in probability theory.

Basic probability theory is defined on measurable spaces, but several essential theorems do not hold for all measurable spaces. Therefore, it is often assumed that the outcome space is a standard Borel space. Such a space emerges if the measurable sets are the Borel sets of a topology defined by a complete separable metric space. This assumption will cover most applications. A standard Borel space has a one-to-one measurable mapping from the outcome space to the unit interval equipped with the Borel σ-algebra.

The primary objects in Kolmogorov's probability theory are the outcome space and a probability measure on the outcome space. All probabilities are with respect to this outcome space and this probability measure. This assumption leads to a consistent theory, but it is just assumed that an outcome space and a probability measure exist. Theorems like the Daniell-Kolmogorov consistency theorem (also called Kolmogorov's extension theorem [[5] page 246, Theorem 1]) make it quite explicit that the existence of an outcome space is a consistency assumption. For a random process $\mathbf{X} = (\xi_t)_{t\in T}$ where $T \subseteq \mathcal{R}$ we define the *finite-dimensional distribution functions* by

$$F_{t_1,\ldots,t_n}(x_1,\ldots,x_n) = P[\omega : \xi_{t_1} \leq x_1,\ldots,\xi_{t_n} \leq x_n] \tag{1}$$

defined for all sets with $t_1 < t_2 < \cdots < t_n$. For the random $X = (\xi_t)_{t\in T}$ the finite-dimensional distributions function satisfies the *Chapman-Kolmogorov equations* stated below.

$$\lim_{x_k\to\infty} F_{t_1,\ldots,t_n}(x_1,\ldots,x_n) = F_{t_1,\ldots,\hat{t}_k,\ldots,t_n}(x_1,\ldots,\hat{x}_k,\ldots,x_n) \tag{2}$$

where $\wedge$ indicates an omitted coordinate

Theorem 1 (Kolmogorov's Theorem on the existence of a process)**.** *Let* $\{F_{t_1,\ldots,t_n}(x_1,\ldots x_n)\}$*, with* $t_i \in T \subseteq \mathbb{R}, t_1 < t_2 < \cdots < t_n$ *,* $n \geq 1$*, be a given family of finite- dimensional distributions, satisfying the Chapman-Kolmogorov Equations (2). Then , there exists a probability space* $(\Omega, \mathcal{F}, \phi)$ *and a random process* $\mathbf{X} = (\xi_t)_{t\in T}$ *such that*

$$P[\omega : \xi_{t_1} \leq x_1,\ldots,\xi_{t_n} \leq x_n] = F_{t_1,\ldots,t_n}(x_1,\ldots,x_n) \tag{3}$$

Later, category theory was developed, and commutative diagrams in category theory are exactly a language for expressing this type of consistency.

2.2. Probabilities or Expectations?

To a large extent, the present paper may be viewed as an extension of the point of view presented by Wittle [6]. By identifying an event with its indicator function, his exposition is formally equivalent to Kolmogorov's probability theory.

If $X = (\Omega, \mathcal{F})$ is a measurable space, we may define $F(\Omega, \mathcal{F})$ as the set of bounded $\mathcal{F}$-measurable functions $\Omega \to \mathbb{R}$. For any unital measure μ on $F(\Omega, \mathcal{F})$ we may define the functional $E_\mu : F(\Omega, \mathcal{F}) \to \mathbb{R}$ by

$$E_\mu(f) = \int_\Omega f \, \mathrm{d}\mu. \tag{4}$$

 doi:10.20944/preprints202410.1897.v1

The functional satisfies that

$$E_\mu(f) \geq 0, \text{ when } f \geq 0; \tag{5}$$
$$E_\mu(1) = 1. \tag{6}$$

Any functional that satisfies these two conditions may be identified with a unital measure.

To describe weak convergence, we are interested in the outcome space as a topological space rather than as a measurable space. A second countable space is also a Lindelöf space, i.e., any open covering has a countable sub-covering. If the measure μ is locally finite, then the whole space has a covering by open sets of finite measures. In particular, the measure μ is σ-finite.

If the space is a locally compact Hausdorff space, then a locally finite measure is the same as a finite measure on compact sets. For such spaces, the integral

$$\int_\Omega f(\omega)\,\mathrm{d}\mu\omega \tag{7}$$

is well-defined for any function f that is continuous with compact support. Radon measures can be identified with positive functionals on $C_c(\Omega)$. With these conditions, the duality between Radon measures and continuous functions with compact support works perfectly without any pathological problems.

On a locally compact Hausdorff space, the finite measures can be identified with positive functionals on the continuous functions on the one-point compactification of the space.

2.3. Probability Theory and Category Theory

For the categorical treatment of probability theory, we first have to recall the definition of a transition kernel [[7] Chapter 1].

Definition 1. *Let $(\Omega, \mathcal{F})$ and $(S, \mathcal{S})$ be two measurable spaces. A* transition kernel $\omega \to \mu_\omega$ *from Ω to S is a function*

$$\mu : \Omega \times \mathcal{S} \to [0, \infty] \tag{8}$$

such that:

- *For any fixed $B \in \mathcal{S}$ the mapping*
$$\omega \to \mu_\omega(B) \tag{9}$$
is measurable.
- *For every fixed $\omega \in \Omega$, the mapping*
$$B \to \mu_\omega(B) \tag{10}$$
is a measure on $(S, \mathcal{S})$.

Let $M_+(\Omega, \mathcal{F})$ denote the measures on $(\Omega, \mathcal{F})$. If $P \in M_+(\Omega, \mathcal{F})$, then a measure on $(S, \mathcal{S})$ is given by

$$B \to \int_\Omega \mu_\omega(B)\,\mathrm{d}P\omega. \tag{11}$$

Thus, a transition kernel may be identified with a mapping $M_+(\Omega, \mathcal{F}) \to M_+(S, \mathcal{S})$ that we will call a *transition operator*.

If the measure μ_ω is a unital measure for any $\omega \in \Omega$, then the transition kernel is called a *Markov kernel*. The key observation for the categorical treatment of probability theory is that Markov kernels can be composed.

The first to put probability theory into the framework of category theory seems to be Lawvere [8]. He defined a category Pro that has measurable spaces as objects and Markov kernels as morphisms.

 doi:10.20944/preprints202410.1897.v1

The category Pro contains the singleton sets as initial objects. A probability measure on $(\Omega, \mathcal{F})$ can then be identified with a morphism from an initial object to $(\Omega, \mathcal{F})$.

Later, the theory of monads was introduced in category theory, and monads have had a significant impact on functional programming [9]. The first to describe the category Pro in terms of monads was Giry [10]. First, we consider the category of measurable spaces Maes. It has measurable spaces as objects and measurable maps between measurable spaces as morphisms.

To the measurable space $X = (\Omega, \mathcal{F})$ we associate the set $M_+^1(X)$ of probability measures on X. The set $M_+^1(X)$ is equipped with the smallest σ-algebra such that the maps $f \to E_\mu(f)$ are all measurable where $E_\mu(f) = \int f \, \mathrm{d}P$. If g is a measurable map from X_1 to X_2 then a measurable map $M_+^1(g)$ from $M_+^1(X_1)$ to $M_+^1(X_2)$ is defined by

$$M_+^1(g)(E_\mu)(f) = E_\mu(f \circ g), \tag{12}$$

which can also be written as $M_+^1(g)(\phi) = E_\mu \circ F(g)$. If f is the indicator function of the measurable set B and the functional ϕ is given by the probability measure μ, then we get

$$M_+^1(g)(\mu)(B) = \mu\Big(g^{-1}(B)\Big), \tag{13}$$

which is the usual definition of an *induced probability measure*. In this way M_+^1 is a functor from the category Maes to itself, i.e., an *endofunctor*.

The morphisms in the category Pro are Markov operators $M_+^1(X_1) \to M_+^1(X_2)$, but Markov operators are given by Markov kernels. It is useful to describe in detail how one can switch between Markov operators and Markov kernels. For this purpose, we introduce a natural transformation δ that translates Markov operators into Markov kernels, and and we introduce a natural transformation π that translates Markov kernels into Markov operators.

A measurable space $X = (\Omega, \mathcal{F})$ can be embedded into $M_+^1(X)$ by mapping $\omega \in \Omega$ into the Dirac measure δ_ω. In this way $\delta_\omega(f) = f(\omega)$. Now δ may be considered as a measurable map, i.e., a morphism in the category Maes, and the following diagram commutes.

$$\begin{array}{ccc} X_1 & X_1 & \xrightarrow{\delta_{X_1}} & M_+^1(X_1) \\ \downarrow{\scriptstyle g} & \downarrow{\scriptstyle g} & & \downarrow{\scriptstyle M_+^1(g)} \\ X_2 & X_2 & \xrightarrow{\delta_{X_2}} & M_+^1(X_2) \end{array} \tag{14}$$

Thus, δ is a natural transformation from the identity functor in the category Maes to the functor M_+^1. If $\Psi : M_+^1(X_1) \to M_+^1(X_2)$ is a Markov operator, then the corresponding Markov kernel $\psi : X_1 \to M_+^1(X_2)$ is given by $\psi_x = \Psi(\delta_x)$.

We will use $\big(M_+^1\big)^2$ to denote the functor $M_+^1 \circ M_+^1$. We have a measurable map π from $\big(M_+^1\big)^2(X)$ to $M_+^1(X)$ that maps $\mu \in \big(M_+^1\big)^2(X)$ to

$$\pi(E_\mu)(f) = \int_{M_+^1(X)} E_\nu(f) \, \mathrm{d}\mu(\nu). \tag{15}$$

Then we have the following commutative diagram

$$\begin{array}{ccc} X_1 & \big(M_+^1\big)^2(X_1) & \xrightarrow{\pi_{X_1}} & M_+^1(X_1) \\ \downarrow{\scriptstyle g} & \downarrow{\scriptstyle \big(M_+^1\big)^2(g)} & & \downarrow{\scriptstyle M_+^1(g)} \\ X_2 & \big(M_+^1\big)^2(X_2) & \xrightarrow{\pi_{X_2}} & M_+^1(X_2) \end{array} \tag{16}$$

 doi:10.20944/preprints202410.1897.v1

so that π is a natural transformation from $\left(M_+^1\right)^2$ to M_+^1. If $\psi : X_1 \to M_+^1(X_2)$ is a Markov kernel, then the corresponding Markov operator Ψ is given by $\Psi = \pi \circ M_+^1(\psi)$. Thus, the following diagram commutes.

$$\begin{array}{ccc} & & \left(M_+^1\right)^2(X_2) \\ & \overset{M_+^1(\psi)}{\nearrow} & \big\downarrow \pi \\ M_+^1(X_1) & \xrightarrow{\Psi} & M_+^1(X_2) \\ \big\uparrow \delta_X & \nearrow_{\psi} & \\ X_1 & & \end{array} \tag{17}$$

A Markov kernel from X_1 to X_2 may now be described as a measurable map ψ_1 from X_1 to $M_+^1(X_2)$. Composition of Markov kernels is given by

$$\psi_2 \odot \psi_1 = \pi \circ M_+^1(\psi_2) \circ \psi_1 \tag{18}$$

and we have the identities

$$\delta \odot \psi = \psi, \tag{19}$$

$$\psi \odot \delta = \psi, \tag{20}$$

$$(\psi_1 \odot \psi_2) \odot \psi_3 = \psi_1 \odot (\psi_2 \odot \psi_3). \tag{21}$$

The first two identities can be translated into the following commutative diagram

$$\begin{array}{ccc} M_+^1(X) & \xrightarrow{\delta_{M_+^1(X)}} & \left(M_+^1\right)^2(X) \\ \big\downarrow M_+^1(\delta) & \searrow & \big\downarrow \pi_X \\ \left(M_+^1\right)^2(X) & \xrightarrow[\pi_X]{} & M_+^1(X) \end{array} \tag{22}$$

Whenever this diagram commutes, we say that δ *acts as an identity*. Associativity means that the following diagram commutes.

$$\begin{array}{ccc} \left(M_+^1\right)^3(X) & \xrightarrow{\pi_{M_+^1(X)}} & \left(M_+^1\right)^2(X) \\ \big\downarrow M_+^1(\pi_X) & & \big\downarrow \pi_X \\ \left(M_+^1\right)^2(X) & \xrightarrow{\pi_X} & M_+^1(X) \end{array} \tag{23}$$

and we say that the functor M_+^1 is *associative*.

When an endofunctor M_+^1 together with two natural transformations δ and π satisfies associativity and δ acts as the an identity, we say that $\left(M_+^1, \delta, \pi\right)$ forms a *monad*. For any monad, morphisms from X to $M_+^1(X)$ can be composed by Equations (18) leading to the *Kleisli composition* of morphisms. For a category with a monad, the Kleisli category of the monad has the same objects as the original category, and as morphisms, it has the Kleisli morphisms. In this way, the category of Markov kernels can be identified with the Kleisli morphisms associated with the monad $\left(M_+^1, \delta, \pi\right)$. The Kleisli category is equivalent to the category introduced by Lawvere. The equivalence is established by a functor that maps the object X into $M_+^1(X)$ and maps Kleisli morphisms into their extensions.

 doi:10.20944/preprints202410.1897.v1

2.4. Preliminaries on Point Processes

First, we will define a point process with points in S. Typically, S will be a d-dimensional Euclidean space, but S could, in principle, denote any complete separable metric space. Let $\mathcal{S}$ denote the Borel σ-algebra on S. Let $(\Omega, \mathcal{F}, P)$ denote a probability space. A transition kernel $\omega \to \mu_\omega$ from $(\Omega, \mathcal{F})$ to $M_+(S, \mathcal{S})$ is called a *point process* if

- For all $\omega \in \Omega$ the measure $\mu_\omega(\cdot) : \mathcal{S} \to \mathbb{R}_{0,+}$ is locally finite.
- For all bounded sets $B \in \mathcal{S}$ the random variable $\omega \to \mu_\omega(B) : \Omega \to \mathbb{R}_{0,+}$ is a count variable.

For further details about point processes, see [11] or [[7] Chapter 3].

The interpretation is that if the outcome is ω then μ_ω is a measure that counts how many points there are in various subsets of S, i.e., $\mu_\omega(B)$ is the number of points in the set $B \in \mathcal{S}$. Each measure μ_ω will be called *an instance* of the point process. In the literature on point processes, one is often interested in *simple point processes* where $\mu_\omega(B) = 0$ when B is a singleton. However, point processes that are not simple are also crucial for the problems that will be discussed in this paper.

The definition of a point process follows the general structure of probability theory, where everything is based on a single underlying probability space. This will ensure consistency, but often this probability space has to be quite large if several point processes or many random variables are considered simultaneously.

The measure μ is called the *expectation measure* of the process $\omega \to \mu_\omega$ if for any $B \in \mathcal{S}$ we have

$$\mu(B) = \int_\Omega \mu_\omega(B) \, \mathrm{d}P\omega. \tag{24}$$

The expectation measure gives the mean value of the number of points in the set B. Different point processes may have the same expectation measure.

A one-point process is a process that outputs precisely one point with probability 1. For a one-point process the expectation measure of the process is simply a probability measure on S. Thus, probability measures can be identified with one-point processes. It is possible to define a monad for point processes [12]. The monad defined in [12] is also related to the observation that the Giry monad is distributive over the multiset monad as discussed in [13]. These results are closely related to the results we will present in the following sections.

A point process can be *thinned*, meaning each point in an instance μ_ω is kept or deleted according to some random procedure. We can do α-thinning for $\alpha \in [0, 1]$ by keeping a point in the sample with probability α and deleting it from the sample with probability $1 - \alpha$. This is done independently for each point in the sample. Thus $\omega \to \nu_\omega$ is an α-thinning of $\omega \to \mu_\omega$ if

$$P(\nu_\omega(B) = \nu(B)) = \binom{\mu(B)}{\nu(B)} \alpha^{\nu(B)} (1 - \alpha)^{\mu(B) - \nu(B)} \tag{25}$$

for any measurable set B. If $\omega \to \nu_\omega$ is an α-thinning of $\omega \to \mu_\omega$ we write $\nu_\omega = T_\alpha \mu_\omega$.

2.5. Poisson Distributions and Poisson Point Processes

For $\lambda \in [0, \infty)$ the Poisson distribution $Po(\lambda)$ is the probability distribution on $\mathbb{N}_0$ with point probabilities

$$Po(j, \lambda) = \frac{\lambda^j}{j!} \exp(-\lambda). \tag{26}$$

The Poisson distribution may be viewed as a point process on a singleton set. The set of Poisson distributions is closed under addition and thinning.

$$Po(\lambda_1) * Po(\lambda_2) = Po(\lambda_1 + \lambda_2), \tag{27}$$

$$T_\alpha(Po(\lambda)) = Po(\alpha \cdot \lambda). \tag{28}$$

 doi:10.20944/preprints202410.1897.v1

For any locally finite measure μ on $(S, \mathcal{S})$ that has density Λ with respect to the Lebesgue measure, can define a *Poisson point process* with *intensity* Λ as a point process $\omega \to \mu_\omega$ with expectation measure μ [[7] Chapter 3]. The following two properties characterize the Poisson point process.

- For all $B \in \mathcal{S}$ the random variable $\omega \to \mu_\omega(B)$ is Poisson distributed with mean value $\mu(B)$.
- If $B_1, B_2 \in \mathcal{S}$ are disjoint, then the random variables $\omega \to \mu_\omega(B_1)$ and $\omega \to \mu_\omega(B_2)$ are independent.

The Equations (27) and (28) also hold if the numbers λ are replaced by measures. The following result was essentially proved by Rényi [14] and Kallenberg [15].

Theorem 2. *Let P be a unital measure and let $\omega \to \mu_\omega$ be independent point processes with expectation measure $\pi(P) = \mu$. Then $T_{1/n}\left(*_{i=1}^{n} \mu_{\omega_i}\right)$ converges to $Po(\mu)$ in total variation.*

2.6. Valuations

Until now, we have presented the results in terms of measure theory. For a more general theory, it is helpful to switch to from measures to *valuations* [16]. Any measure μ satisfies the following properties.

Strictness $\mu(\emptyset) = 0$.
Monotonicity For all subsets $A \subseteq B$, implies $\mu(A) \leq \mu(B)$.
Modularity For all subsets A, B,

$$\mu(A) + \mu(B) = \mu(A \cup B) + \mu(A \cap B). \tag{29}$$

For any lattice with a bottom element a valuation is defined as a function that satisfies strictness, monotony, and modularity. The notion of a valuation makes sense in any lattice with a bottom element.

We are mainly interested in valuations that are continuous in the following sense. The possible values of a valuation are lower reals in $\overrightarrow{[0,\infty]}$, which are elements in the set $[0,\infty]$ with a topology of lower bounded open intervals. Thus, a function into $\overrightarrow{[0,\infty]}$ is continuous, if it is lower semicontinuous when $[0,\infty]$ the usual topology.

Continuity $\mu\left(\sup_\lambda A_\lambda\right) = \sup_\lambda \mu(A_\lambda)$ for any directed net A_λ.

This notion of continuity captures both the inner regularity of a measure and it captures σ-additivity.

A Borel measure restricted to the open sets of a topological space is a valuation. On any complete separable metric space any continuous valuation on the open sets is the restriction of a Borel measure. It will not make any difference whether we speak of measures or valuations for the applications that we will discuss in Section 5

If X is a topological space, then $\mathcal{V}(X)$ denotes the set of continuous valuations on X. The set of valuations has a structure as a topological space, so that $\mathcal{V}$ is an endofunctor in the category Top of topological spaces. The functor $\mathcal{V}$ defines a monad that is called *the extended probabilistic power domain* [17]. It is defined in much the same way as the monads defined by Giry.

3. Observations

The outcome space plays a central role in Kolmogorov's approach to probability. The basic experiment in his theory will result in a single point in the outcome space. The measurable subsets of the outcome space play the same role as the propositions play in a Boolean algebra in logic. The principle of excluded middle in logic states that any proposition is either true or false. Similarly, the basic experiment in Kolmogorov's theory results in a single point, and these points exclude each other. In this paper, the result of a basic experiment will be a multiset rather than a single point.

 doi:10.20944/preprints202410.1897.v1

3.1. Observations as Multiset Classifications

In computer science, we operate with different data types. *A set* is a collection of objects, where repetition and order are not relevant. *A list* is a collection of objects, where order and repetition are relevant. *Multisets* are unordered, but repetitions are relevant, so objects are counted with multiplicity. We shall discuss the relation between these data types in this subsection.

A review of the theory of multisets can be found in [18]. Monro [19] discusses two ways of defining a multiset, and the distinction between these definitions is important for the present work.

The following example is similar to what can be found in basic textbooks on statistics.

Example 1. *Consider a list of observations of five animals like* $(cow, horse, bee, sheep, flie)$*. The list may be represented by a table with a number as a unique key that indicates the order in which we have made the observations. In the present example all the animals are different and if we are not interested in the order in which*

Key	Animal
1	cow
2	horse
3	bee
4	sheep
5	flie

we have made the observations, we may represent the observations by the set $\Omega = \{bee, cow, flie, horse, sheep\}$ *where the animals have been displayed in alphabetic order, but in a set the order does not matter.*

These animals can be classified as either mammals or insects, leading to the list of observations $(mammal, mammal, insect, mam$ *or, equivalently, to the table. Since the list contains repetitions, we cannot represent it by the set* $\mathbb{A} =$

Key	Animal
1	mammal
2	mammal
3	insect
4	mammal
5	insect

$\{insect, mammal\}$*. Instead, we may represent it by the multiset* $[insect, insect, mammal, mammal, mammal]$.

According to the first definition of *a multiset* by Monro [19], a multiset is a set Ω with an equivalence relation $\simeq$. If $\mathbb{A}$ denotes the set of equivalence classes, then we get a mapping $g : \Omega \to \mathbb{A}$. Alternatively, any mapping $g : \Omega \to \mathbb{A}$ leads to equivalences classes on Ω. Dedekind was the first to identify a multiset with a function [18]. To each equivalence class we assign the number of elements in the equivalence class. This is called the *multiplicity* of the equivalence class.

A category Mul with multisets as objects was defined by Monro [19]. An object in the category Mul is a set Ω with an equivalence relation $\simeq$. If $(\Omega_1, \simeq_1)$ and $(\Omega_2, \simeq_2)$ are objects in the category Mul, then a morphism from $(\Omega_1, \simeq_1)$ to $(\Omega_2, \simeq_2)$ is a mapping $f : \Omega_1 \to \Omega_2$ that respects the equivalence relations, i.e., if $x \simeq_1 y$ in $(\Omega_1, \simeq_1)$ then $f(x) \simeq_2 f(y)$ in $(\Omega_2, \simeq_2)$. This category has been studied in more detail in [20].

An equivalence is often based on a partial preorder. Let $\preceq$ denote a partial preordering on Ω. Then, an equivalence relation $\simeq$ is defined by $a \simeq b$ if and only $a \preceq b$ and $b \preceq a$. If $\mathbb{A} = \omega / \simeq$ then $\preceq$ induces an ordering on $\mathbb{A}$, that we will also denote $\preceq$.

The downsets (hereditary sets) in $(\Omega, \preceq)$ form a distributive lattice with $\cap$ and $\cup$ as lattice operations. If the set $\mathbb{A}$ is finite, then the lattice is finite, and if $(\mathbb{A}, \preceq)$ satisfies the *decending chain condition* (DCC) then so does the lattice of downsets. Any finite distributive lattice can be represented by a finite poset [[21] Thm. 9], and a distributive that satisfies DCC can be represented by a poset that satisfies DCC. There is a one-to-one correspondence between the elements of $\mathbb{A}$ and the irreducible

 doi:10.20944/preprints202410.1897.v1

elements of the lattice. This construction can be viewed as a *concept lattices* [22] based on the partial ordering $\preceq$.

The partial preordering $\preceq$ is an equivalence relation if and only if the lattice generated is a Boolean lattice. Therefore, we get a lattice where one cannot form complements except if $\preceq$ is an equivalence relation. The shift from Boolean lattices to more general lattices corresponds to shifting from logic with a law of excluded middle to more general classification systems.

The set of downset in $(\Omega, \preceq)$ is closed under finite intersections and under arbitrary unions because any union is a finite union. Thus, the downsets in $(\Omega, \preceq)$ form a *topology*. The continuous functions for this topology are monotone functions for the ordering. Topological spaces and continuous functions form a category. If some of the equivalence classes in $(\Omega, \preceq)$ has more than one element, then the topology does not satisfy the separation condition T_0, but the topology on equivalence classes always satisfies T_0.

Multisets can also be described using σ-algebras as it is done in probability theory. A classification on Ω given by an equivalence relation $\simeq$ or by a partial preordering $\preceq$ leads to a topology on Ω. This topology generates a Borel σ-algebra on Ω. For any two outcomes $\omega_1, \omega_2 \in \Omega$ we have $\omega_1 \simeq \omega_2$ if and only if $\omega_1 \in B \Leftrightarrow \omega_2 \in B$ for all Borel measurable sets B.

3.2. Observations as Empirical Measures

According to the second definition of *a multiset* discussed by Monro [19], a multiset is a mapping of a set $\mathbb{A}$ into $\mathbb{N}_0$ that gives the multiplicity of each of the elements. This corresponds to the data types in statistics that are called *count data*. Here, we will represent such multisets by *finite empirical measures*.

Example 2. *The list* $(mammal, mammal, insect, mammal, insect)$ *can be written as a multiset* $[insect, insect, mammal, mammal, mammal]$ *or, equivalently, we may represent the multiset by the measure* $2 \cdot \delta_{insect} + 3 \cdot \delta_{mammal}$*. Alternatively, the multiset can be represented by a table of frequencies.*

Classification	Frequency
insect	2
mammal	3

The mapping from lists to an empirical measure is called the *accumulation map*, and it is denoted *Acc*.

Definition 2. *Let* $(\mathbb{A}, \tau)$ *be a topological space. By a* finite empirical measure*, we understand a finite sum of Dirac measures on the Borel* σ*-algebra of* $(\mathbb{A}, \tau)$*. The set of finite empirical measures on* $(\mathbb{A}, \tau)$ *will be denoted* $\mathcal{M}(\mathbb{A}, \tau)$ *or* $\mathcal{M}(\mathbb{A})$ *for short.*

With these definitions, $\mathcal{M}$ is an endofunctor on the category Maes of measurable spaces.

When the notion of observation is based on an equivalence, there is an implicit assumption that any two elements each has its own identity, but at the same time, they are equivalent according to a classification. In quantum physics, particles may be indistinguishable in a way that one would not see in classical physics. The following example shows that there are data sets that the first definition of a multiset cannot handle.

Example 3. *Young's experiment, also called the double slit experiment, was first used by Young as a strong argument for the wave-like nature of light. Nowadays, it is often taken as an illustration of how quantum physics fundamentally differs from classical physics. The observations are often described as a paradox, but at least part of the paradox is related to a wrong presentation of the observations.*

In its modern form, the experiment uses monochromatic light from a laser. The laser beam is first sent through a slit in a screen. The electromagnetic wave spreads like concentric circles after passing through the first slit, as illustrated in Figure 1. The wave then hits a second screen with two slits. After passing the two

 doi:10.20944/preprints202410.1897.v1

slits we get two waves that spread like concentric circles until they hit a photographic film that will display an interference pattern created by the two waves.

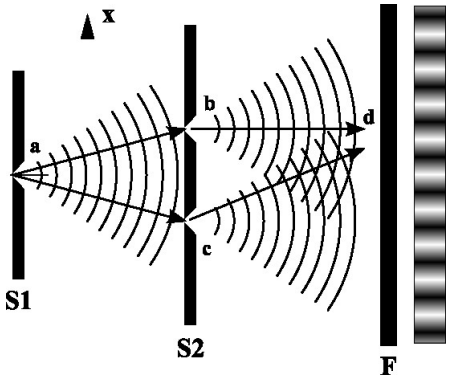

Figure 1. When coherent light is sent through first a single slit in screen S1 and then through the two slits b and c on screen S2, then an interference pattern emerges at the photographic film F.

Now follows a "paradox" as it is described in many textbooks. The electromagnetic wave is quantized into photons, so if the intensity of the light is low we will only get separate points on the photographic film, but we get the same interference pattern as before. This may be explained as interference between photons passing through the left and the right slit respectively. Now we lower the intensity so much that the photons arrive at the photographic film one by one. After a long time of exposure, we get the same interference pattern as before. This apparently gives a paradox because it is hard to understand how a single photon should pass through both slits and have interference with itself.

A problem with the above description is that the number of photons emitted from a laser is described by a Poisson point process. Even if the intensity is low, one cannot emit a single photon for sure. If the mean number of photons emitted is say 1 then there is still a probability that 0 or 2 (or more) photons are emitted and this will hold even if the intensity is very low. What we observe is a number of dots on the photographic film, which can be described as by a multiset if the photographic film is divided into regions. If we really want to send single photons, *it could be done using* a single-photon source. *For a single-photon source there is no uncertainty in the number of photons emitted, but according to the time-energy uncertainty relation there will still be uncertainty in the energy. The energy E of a photon is given by*

$$E = hf = h\frac{c}{\lambda} \tag{30}$$

where h is Planck's constant, f is the frequency, c is the speed of light, and λ is the wave length. Thus, uncertainty in the energy means uncertainty in frequency and wave length. Since the interference pattern observed on the photographic film depends on the wave length, one would not get the same interference pattern if a single-photon emitter was used instead of coherent light from a laser.

We see that the idea that the observation in terms of a multiset comes from observation of a large number of individual *photons is simply wrong and leads to an inconsistent description of the experiment.*

We can do the following operations with empirical measures.

- Addition.
- Restriction.
- Inducing.

The first operation we will look at is *addition*. Let ℓ_1 and ℓ_2 denote two lists of observations from the same alphabet $\mathbb{A}$, and let $\ell_1 * \ell_2$ denote the concatenation of the lists. Then $Acc(\ell_1 * \ell_2) = Acc(\ell_1) + Acc(\ell_2)$. Thus, the sum $\mu_1 + \mu_2$ of two empirical measures has an interpretation via merging two datasets together. The corresponding operation for point processes is called *superposition*. Addition of

 doi:10.20944/preprints202410.1897.v1

empirical measures is a way of obtaining an empirical measure without using the accumulation map on a single list.

The next operation we will look at is *restriction*. If data is described by the empirical measure μ on $\mathbb{A}$ and B is a subset of $\mathbb{A}$ then the restriction of μ to subsets of B is an empirical measure on B that we will denote by $\mu_{\cap B}$. In probability theory all measures should be unital so we need the notion of conditional probability, but multisets cannot be normalized, and the concept simplifies to the notion of restriction.

Assume that $g : \mathbb{A} \to \mathbb{B}$ is a continuous (or measurable) mapping. Then the induced measure $g(\mu)$ is defined by

$$g(\mu)(B) = \mu\Big(g^{-1}(B)\Big), \tag{31}$$

where μ is an empirical measure on $\mathbb{A}$ and B is a measurable subset of $\mathbb{B}$. The measure $\mathcal{M}(g)(\mu)$ is called the *induced measure* and is often denoted $g(\mu)$ rather than $\mathcal{M}(g)(\mu)$.

One can easily prove that inducing is additive in the measure and similar basic results regarding the interaction between addition, restriction and creation of induced measures.

3.3. Categorical Properties of the Empirical Measures and Some Generalizations

Addition, restriction, and inducing can be described in the language of category theory. The existence of addition means that the category of multisets is *commutative*. Restriction can be characterized as a retraction that is additive. The existence of induced measures means that $\mathcal{M}$ is a functor from topological spaces to measure spaces.

If $(\mathbb{A}, \tau)$ is a topological space, then we equip $\mathcal{M}(\mathbb{A}, \tau)$ with the smallest topology such that

$$\mu \to \int f \, \mathrm{d}\mu \tag{32}$$

are continuous for all continuous functions f. Here, the integral $\int f \, \mathrm{d}\mu$ simplifies to a sum $\sum_{i=1}^{n} f(a_i)$ when $\mu = \sum_{i=1}^{n} \delta_{a_i}$. If we define $\mathcal{M}(g)(\mu) = g(\mu)$ then $\mathcal{M}$ becomes an endofunctor in the category of topological spaces and a monad can be defined in exactly the same way as for the Giry monad. Kleisli morphisms map points in a topological space into empirical measures on the topological space. Empirical measures will often be denoted μ_ω in order to emphasize that an empirical measure typically is the results of some sampling process.

One can generalize finite empirical measures on a topological space to continuous integer valued valuations on a topological space. These integer valued valuations form a sub-monad of the monad of all continuous valuations.

As we have seen in Section 3.1 classifications may lead to other lattices than Boolean lattices, so it is relevant to discuss integer-valued valuations on more general distributive lattices. We shall work out the theory for finite lattices in all details below. If (Ω, τ) is a finite topological space, then we can define an integer-valued valuation v on the topology τ by

$$v(B) = \#(B), \tag{33}$$

where $\#(B)$ means the number of elements in the open set $B \in \tau$.

Lemma 1. *Let v be a continuous valuation on a topological space (Ω, τ). Let $c \in L$ be some element. Then the restriction $v_{\cap c}$ given by $v_{\cap c}(a) = v(a \cap c)$ is an integer valued valuation. If $v(c) < \infty$ then $v_{\setminus c}$ given by $v_{\setminus c}(a) = v(a) - v(a \cap c)$ is also an integer valued valuation and $v = v_{\cap c} + v_{\setminus c}$. If v is continuous, then $v_{\cap c}$ and $v_{\setminus c}$ are continuous.*

 doi:10.20944/preprints202410.1897.v1

Proof. The strictness of $v_{\cap c}$ and $v_{\setminus c}$ are obvious. The monotonicity of $v_{\cap c}$ is obvious. To see that $v_{\setminus c}$ is monotone, let $a \subseteq b$ be some elements in the lattice. Then

$$v(a \cup c) \leq v(a \cup c), \tag{34}$$
$$v(a) + v(c) - v(a \cap c) \leq v(b) + v(c) - v(b \cap c), \tag{35}$$
$$v(a) - v(a \cap c) \leq v(b) - v(b \cap c), \tag{36}$$
$$v_{\setminus c}(a) \leq v_{\setminus c}(b). \tag{37}$$

Modularity of $v_{\cap c}$ is a simple calculation. Modularity of $v_{\setminus c}$ follows because $v_{\setminus c} = v - v_{\cap c}$. Continuity is obvious. □

Let v denote a valuation on a topological space (Ω, τ). Then $a \in L$ is called an *atom* of the valuation if $b \subseteq a$ implies that $v(b) = 0$ or $v(b) = v(a)$. An *atomic valuation* is a valuation that is a sum of valuations for which L is an atom.

Proposition 1. *Any integer valued valuation on a topological space (Ω, τ) is atomic.*

Proof. The proof is by induction on $n = v(\Omega)$. First, the results hold for the trivial valuation with $v(\Omega) = 0$. Assume that the result holds for any valuation with $v(\Omega) \leq n$. Let v be a valuation with $v(\Omega) = n + 1$. If L is an atom, then v is atomic. If Ω is not an atom, then there exists $c \in L$ such that $0 < v(c) < v(\Omega) = n$. Then $v_{\cap c}$ and $v_{\setminus c}$ are atomic valuations and $v = v_{\cap c} + v_{\setminus c}$ implying that v is atomic. □

Theorem 3. *Let $(\mathbb{A}, \tau)$ denote a finite topological space. If the topology satisfies the separation condition T_0, then for any integer valued valuation v on the lattice of open sets there exists a uniquely determined empirical measure μ on $\mathbb{A}$ such that for any open set B we have*

$$v(B) = \mu(B). \tag{38}$$

Proof. We have to prove that if $\mathbb{A}$ is an atom for the valuation v, then v is given by a uniquely determined Dirac valuation. Let A denote a minimal atom. Then

$$\begin{aligned} v(B) &= v(B \cup A) + v(B \cap A) - v(A) \\ &= v(A) + v(B \cap A) - v(A) \\ &= v(B \cup A) \\ &= \begin{cases} v(A), & \text{if } A \subseteq B; \\ 0, & \text{else.} \end{cases} \end{aligned} \tag{39}$$

For $a \in A$ let $a_{\downarrow}$ denote the smallest open set that contains a. Then there must exist $a \in A$ such that $v(a_{\downarrow}) = v(V)$, since, otherwise, one would have $V(A) = 0$. Hence we have $A \subseteq a_{\downarrow}$, which implies that $A = a_{\downarrow}$. Hence,

$$\begin{aligned} v(B) &= \begin{cases} v(A), & \text{if } a \in B; \\ 0, & \text{else.} \end{cases} \\ &= \delta_a(B). \end{aligned} \tag{40}$$

Assume that $\delta_{a_1} = \delta_{a_1}$. Then, $\delta_{a_1}(B) = \delta_{a_1}(B)$ for all open sets B. Hence, $a_1 \in B$ if and only if $a_2 \in B$. Since the topology satisfies T_0 , we must have $a_1 = a_2$. □

 doi:10.20944/preprints202410.1897.v1

As a consequence of the theorem, any integer-valued valuation v on a finite distributive lattice can be represented by a finite set Ω with a topology such that the multiset obtained by mapping Ω to equivalence classes $\mathbb{A}$ satisfies (33).

For applications in statistics, the main example of a lattice is the topology of a complete separable metric space.

Theorem 4. *Let μ denote an integer valued valuation on the topology of a complete separable metric space. Then μ is a simple valuation, i.e., there exists integers $s_1, s_2, \ldots, s_n \in \mathbb{N}_0$ and letters $a_1, a_2, \ldots, a_n \in \mathbb{A}$ such that*

$$\mu = \sum_{i=1}^{n} s_i \cdot \delta_{a_i}. \tag{41}$$

Proof. Based on a result of Topsøe [[23] Thm. 3] Manilla has proved that any valuation on the topology of a metric space can be extended to a measure μ on the σ-algebra of Borel sets of the metric space [[24] Cor. 4.5]. When the metric space is separable and complete, we may, without loss of generality, assume that the metric space is $\mathbb{B} = [0, 1]$ and let $\mathbb{B}$ denote the identity on $[0, 1]$. Then $F(x) = \mu(X \leq x)$ is increasing and integer-valued. Therefore, F is a staircase function with a finite number of steps. The measure P will have a point mass at each step and no mass in between. Hence, μ is simple. □

3.4. Lossless Compression of Data

Bayesian statistics focus on single outcomes of experiments and the frequentist interpretation focus on infinite i.i.d. sequences. Information theory takes a position in between. In information theory, the focus is on extendable sequences rather than on finite or infinite sequences [25]. In lossless source coding, we consider a sequence of observations in the source alphabet $\mathbb{A}$, i.e., an observation is an element in $\mathbb{A}^n$ where n is some natural number. We want to encode the letters in the source alphabet by sequences of letters in an output alphabet $\mathbb{B}$ of length β. In lossless coding, the encoding should be uniquely decodable. Further, the encoding should be so that as concatenation of source letters is encoded as the corresponding concatenation of output letters. We require that not only A^n is uniquely decodable, but any (finite) string in $\mathbb{A}^*$ should be encoded into a string in $\mathbb{B}^*$ in a unique way. If the code $\kappa : \mathbb{A} \to \mathbb{B}^*$ is uniquely decodable then it satisfies Kraft's inequality

$$\sum_{a \in \mathbb{A}} \beta^{-\ell(\kappa(a))} \leq 1 \tag{42}$$

where $\ell(\kappa(a))$ denotes the length of the code word $\kappa(a)$ [[26] Thm. 5.2.1]. Instead of encoding single letters in $\mathbb{A}$ into $\mathbb{B}^*$ we may do it as block coding where a block in $\mathbb{A}^n$ is mapped as a string $\mathbb{B}^*$ via a mapping κ. The following theorem is a kind of reverse of Kraft's inequality for block coding.

Theorem 5 ([25])**.** *Let $\ell : \mathbb{A} \to \mathbb{R}$ denote a function. Then the function ℓ satisfies Kraft's inequality 42 if and only if for all $\epsilon > 0$ there exists an integer n and a uniquely decodable code $\kappa : \mathbb{A}^n \to \mathbb{B}^*$ such that*

$$\left| \bar{\ell}_\kappa(a_1^n) - \frac{1}{n} \sum_{i=1}^{n} \ell(a_i) \right| \leq \epsilon \tag{43}$$

where $\bar{\ell}_\kappa(a_1^n)$ denotes the length $\ell_\kappa(a_1^n)$ divided by n.

Using this theorem, we can identify uniquely decodable codes with code length functions that satisfy Kraft's inequality. There is a correspondence between codes and sub-unital measures given by

$$\mu(a) = \beta^{-\ell(\kappa(a))}. \tag{44}$$

 doi:10.20944/preprints202410.1897.v1

Now, the goal in lossless source coding is to code with a code-length that is as short as possible. If a code word has empirical measure μ and the function ℓ is used then the total code length is

$$\sum_a \ell(a)\mu(a). \tag{45}$$

Our goal is to minimize the total code length and this is achieved by the code length function

$$\ell(a) = -\log_\beta\left(\frac{\mu(a)}{n}\right). \tag{46}$$

If a code with this code length function is used then the total code length will be $n \cdot H(\frac{\mu}{n})$ where H denotes the Shannon entropy of a probability measure. We can define the entropy of any finite discrete measure by

$$H(\mu) = \mu(\mathbb{A})H\left(\frac{\mu}{\mu(\mathbb{A})}\right). \tag{47}$$

With this definition one can easily prove that if $g : \mathbb{A} \to \mathbb{B}$ is a measurable mapping and μ is a measure on $\mathbb{A}$ then the following *chain rule* holds

$$H(\mu) = H(g(\mu)) + \sum_{b\in\mathbb{B}} H\Big(\mu_{|g^{-1}(b)}\Big). \tag{48}$$

The chain rule reflects that coding the result in $\mathbb{A}$ can be done by first coding the result in $\mathbb{B}$ and then code the result in $\mathbb{A}$ restricted to subset of letters in $\mathbb{A}$ that maps into the observed letter in $\mathbb{B}$.

One could proceed on exploring the correspondence between measures and codes as is done in the minimum description length (MDL) approach to statistics, and much of the content of Section 4 could be treated as using MDL. For instance, the number of source letters of length n is α^n and it grows exponentially, but the number of different multisets of size n is upper bounded by $(n+1)^\alpha$ and it grows like a polynomial in n. This fact has important consequences related to the maximum entropy principle and in the information theory literature [27] it is called the *method of types* where type is another word for multiset.

In order to emphasize the distinction between probability measures and expectation measures, we will not elaborate on this approach in the present exposition. Here we will just emphasize that Kraft's inequality and Equations (46) are some of the few examples where a theorem states that unital measure play a special role beyond the fact that all finite measures can be normalized.

3.5. Lossy Compression of Data

If an information source is compressed to a rate below the entropy of the source letter cannot be reconstructed from the output letters. In this case one will experience a loss in description of the source and methods for minimizing the loss are handled by rate distortion theory. In rate distortion theory, we introduce a distortion function $d : \mathbb{A} \times \hat{\mathbb{A}} \to \mathbb{R}$ that quantified how much is lost if $a \in \mathbb{A}$ was sent, and it was reconstructed as $\hat{\mathbb{A}}$. If $\hat{\mathbb{A}} \subseteq \mathbb{A}$ then d may be a metric or a function of a metric. As demonstrated in the papers [28–30] many aspects of statistical analysis can be handled by rate distortion theory. This involves cluster analysis, outlier detection, testing Goodness-of-Fit, estimation, and regression. Important aspects of statistics can be treated using rate distortion theory, but in order to keep the focus on the distinction between probability measures and expectation measures, we will not go into further details regarding rate distortion theory.

4. Expectations

Multisets, empirical measures, and integer valued valuations are excellent for descriptive statistics, but these concepts neither describe randomness, sampling, nor expectations. In this section, we will discuss more general categories where these concepts are incorporated.

 doi:10.20944/preprints202410.1897.v1

4.1. Simple Expectation Measures

Let Ω denote a large outcome space and let μ_ω denote the empirical measure on the set $\mathbb{A}$ if the outcome is ω. The empirical measure μ_ω can be described as a list of frequencies or as a multiset. Assume that the size of the multiset is $N = \mu_\omega(\mathbb{A})$. The measure μ_ω may describe a sample from a population, but it may also describe the whole population, in which case the subscript ω is not needed. In modern statistics various resampling techniques play an important role, and for this reason we will keep the subscript in order both to describe sampling and resamling. Now we take a sample of size n from the population.

The simplest situation is when $n = 1$. If $B \subseteq \mathbb{A}$ then $\frac{1}{N} \cdot \mu_\omega(B)$ is the probability that a randomly selected point from the multiset described by μ_ω belongs to B. The unital measure $\frac{1}{\mu_\omega(\mathbb{A})} \cdot \mu_\omega$ is the *empirical distribution*. The empirical measure μ_ω gives a table of frequencies and the empirical distribution gives a table of relative frequencies.

Next, consider the situation when we take a sample of size $n > 1$ from the population. There are different ways of taking a sample of size n. One may sample with replacement or without replacement. These two basic sampling schemes are described by the multinomial distribution and by the multivariate hyper-geometric distribution respectively. For both sampling methods, the mean number of observations in a set B is given by $\frac{n}{N} \cdot \mu_\omega(B)$. Thus, the expected values are described by the measure $\alpha \cdot \mu$ where $\alpha = n/N$. Here we have scaled the measure μ_ω by a factor $\alpha \in [0, 1]$, and this leads to a measure that is not given by a multiset.

Consider a sample described by an empirical measure μ_ω with sample size $N = \mu_\omega(\mathbb{A})$. For cross validation, we may randomly partition the sample into a number of subsamples. One may then check if a conclusion based on a statistical analysis of one of the subsamples is the same as if another subsample had been taken. If there are k random subsamples, then the expected number of observations in B is $\frac{1}{k} \cdot \mu_\omega(B)$ and the random subsample may be described by the measure $\alpha \cdot \mu_\omega$ where $\alpha = 1/k$.

In bootstrap re-sampling one selects n objects from a sample of size n but this is done with replacement. If the sample is described by the measure μ_ω, then the mean number of observations in B is $\mu_\omega(B)$. Thus, bootstrap re-sampling corresponds to the measure $\alpha \cdot \mu_\omega$ where $\alpha = 1$. We see that although multiplying by 1 does not change the measure, it may reflect a non-trivial re-sampling procedure.

In general, we may perform α-thinning of a multiset. This is done by keeping each point with probability α and deleting it with probability $1 - \alpha$. The preservation/deletion of observations is done independently for each observation. For values of α in $[0, 1] \cap \mathbb{Q}$ we can implement α-thinning using a random number generator that gives a uniform distribution on a finite set. Such random numbers can be created by rolling a die, draw a card from a deck, or a similar physical procedure. The set $[0, 1] \cap \mathbb{Q}$ is not complete, which is inconvenient for formulating various theorems. Therefore, we will also allow irrational values of α so that α can assume any value in $[0, 1]$.

We discussed addition of measures in the previous section. In particular, we may add n copies of the measure μ_ω together to obtain the measure $n \cdot \mu_\omega$. Then we may sample from $n \cdot \mu_\omega$ by thinning by some factor $\alpha \in [0, 1]$ so that we obtain the measure $\alpha \cdot (n \cdot \mu_\omega) = (\alpha \cdot n) \cdot \mu_\omega$. In this way, we may multiply a measure by any positive number. In general, there will be many different ways of implementing a multiplication by the positive number t:

- There are many ways of writing t as a product $\alpha \cdot n$ where $\alpha \in [0, 1]$ and n is an integer.
- There are many different sampling schemes that will lead to a multiplication be α.
- There are many ways of generating the randomness that is needed to perform the sampling.

Although there are many ways of implementing a multiplication of the measure μ_ω by the number $t \geq 0$, it is often sufficient to know the resulting measure $t \cdot \mu_\omega$ rather than details about how the multiplication is implemented. In Section 4.3 we will introduce a kind of default way of implementing the multiplication.

 doi:10.20944/preprints202410.1897.v1

By allowing multiplication by positive numbers we can obtain any finite measure concentrated on a finite set, i.e., measures of the form

$$\mu = \sum s_a \cdot \delta_a. \tag{49}$$

The set $M_+^{fin}(\mathbb{A})$ is defined as the finitely supported finite measures on $\mathbb{A}$. The finitely supported finite measures are related to the empirical measures in exactly the same way as probability measures are related to truth assignment functions in logic.

4.2. Categorical Properties of the Expectation Measures and Some Generalizations

We want to study a monad that allow us to work with both empirical measures and unital measures as done in probability theory. The set $M_+^{fin}(\mathbb{A})$ is defined as the finitely supported finite measures on $\mathbb{A}$. If $P \in M_+^{fin}(\mathcal{M}(\mathbb{A}))$ is a probability measure and the outcome space is $\Omega = \mathcal{M}(\mathbb{A})$ then $\omega \to \mu_\omega$ is formally a point process with points in $\mathbb{A}$. The expectation measure $\pi(P)$ of the point process $\omega \to \mu_\omega$ is given by $M_+^{fin}(\mathcal{M}(\mathbb{A}))$ to $M_+^{fin}(\mathbb{A})$ by

$$\pi(P)(f) = \int_{\mathcal{M}(\mathbb{A})} \sum_{\mathbb{A}} f(a)\mu_\omega(a)\, \mathrm{d}P\omega. \tag{50}$$

By linearity the transformation π defined by Equations (50) extends to a natural transformation from $M_+^{fin}\left(M_+^{fin}(\mathbb{A})\right)$ to $M_+^{fin}(\mathbb{A})$.

As before, we let δ denote the natural transformation that maps $a \in \mathbb{A}$ into δ_a, i.e., the Dirac measure concentrated in a. It is straight forward to check that $\left(M_+^{fin}, \delta, \pi\right)$ forms a monad. The Kleisli morphisms generated by this monad will generate a category that we will call *the finite expectation category*, and we will denote it by Fin.

Finite lattices can be represented by finite topological spaces, and for these spaces the theory is simple.

Theorem 6. *Let $(\mathbb{A}, \tau)$ denote a finite topological space. If the topology satisfies the separation condition T_0, then for any finite valuation v on τ there exists a finite expectation measure μ on the Borel σ-algebra $\mathbb{A}$ such that for any open set B we have*

$$v(B) = \mu(B). \tag{51}$$

Proof. The proof is an almost a word by word repetition of the proof of Theorem 3. □

All finite expectation measures are continuous valuations on a topological space. The monad of continuous valuations on topological spaces is important because it includes all probability measures on complete separable metric spaces, that is the most used model in probability theory.

4.3. The Poisson Interpretation

Let μ denote a discrete measure such that

$$\mu = \sum_{a \in \mathbb{A}} s_a \cdot \delta_a \tag{52}$$

where $\mathbb{A} = \{a \mid \mu(a) > 0\}$. Then the Poisson point process $Po(\mu)$ given by the product

$$\bigotimes_{a \in \mathbb{A}} Po(s_a) \tag{53}$$

is a point process with expectation measure μ. Thus, any discrete measure has an interpretation as the expectation measure of a discrete Poisson point process (see Section 2.5). This interpretation will be

 doi:10.20944/preprints202410.1897.v1

called *the Poisson interpretation*, and it can be used to translate properties and results for (non-unital) expectation measure into properties and results for probability measures.

A. Rényi was the first to point out that a Poisson process has extreme properties related to information theory [31] and entropy for point processes were later studied by Mc Fadden [32]. Their results were formulated for simple point processes. Here we will look at some results for processes supported on a finite number of points. If we thin a one-point process, We will get a process given by the expectation measure $P = \sum_i p_i \cdot \delta_i$ where $\sum_i p_i \leq 1$ and $1 - \sum_i p_i$ is *the void probability*, i.e., the probability of getting no point. Here we shall just present two results that are generalizations of similar results in [33–35]. We say that a point process is a *multinomial sum* if it is a sum of independent thinned one point processes. Let $Be(\mu)$ denote the set of sums of independent thinned one-point processes with expectation measure μ and let $Be^*(\mu)$ denote the total variation closure of $Be(\mu)$. As a consequence of Theorem 2 the distribution $Po(\mu)$ lies in $Be^*(\mu)$. The following results can be proved in the same way as a corresponding result in [33] was proved.

Theorem 7. *The maximum entropy process in $Be^*(\mu)$ is the Poisson point process $Po(\mu)$.*

The following result states that a homogeneous process has greater entropy than an inhomogeneous process with the same mean number of points. This is a point process version of the result that the uniform distribution is the distribution with maximal entropy on a finite set.

Theorem 8. *Let $\mathbb{A}$ be a set with m elements. Then the Poisson point process $Po(\mu)$ that has maximum entropy under the condition that $\mu(\mathbb{A}) = \lambda$, is the process where $\mu(a) = \lambda/m$ for all $a \in \mathbb{A}$.*

Proof. We note that

$$\begin{aligned} H(Po(\mu)) &= H\left(\bigotimes_a Po(\mu(a))\right) \\ &= \sum_a H(Po(\mu(a))) \end{aligned} \tag{54}$$

so it is sufficient to prove that the sum is Shur concave under the condition that $\sum_a \mu(a) = \lambda$.

Let $\mu_1, \mu_2 > 0$ and let $X_i \sim Po(\mu_i)$ then

$$\begin{aligned} H(X_1) + H(X_2) = H((X_1, X_2)) &= H(X_1 + X_2) + H((X_1, X_2)|X_1 + X_2) \\ &= H(Po(\mu_1 + \mu_2)) + \sum_{n=0}^{\infty} H\left(bin\left(n, \frac{\mu_1}{\mu_1 + \mu_2}\right)\right) \cdot \frac{(\mu_1 + \mu_2)^n}{n!} \exp(-(\mu_1 + \mu_2)), \end{aligned} \tag{55}$$

where $bin(n, p)$ means the binomial distribution with number parameter n and success probability p. For a fixed value of $\mu_1 + \mu_2$ we have to maximize $H\left(bin\left(n, \frac{\mu_1}{\mu_1+\mu_2}\right)\right)$. The maximum is achieved when $\frac{\mu_1}{\mu_1+\mu_2} = \frac{1}{2}$, i.e., $\mu_1 = \mu_2$. To see that $H(bin(n, p))$ is maximal for $p = 1/2$ it is sufficient to note that $p \to H(bin(n, p))$ is a concave function which follows from [36]. □

The Poisson interpretation does not only work for finite expectation measure. The following result is relevant for valuations on topological spaces.

Theorem 9. *Let μ denote a σ-finite measure on a complete separable metric space. Then μ is an expectation measure of a Poisson point process.*

Proof. Since the measure μ is σ-finite, it can be written as $\mu = \sum_{i=1}^{\infty} \mu_i$, where μ_i is a finite measure for all i. Each μ_i can be written as a sum of a discrete measure and a continuous measure. The discrete measure is the expectation measure of a discrete product of Poisson distributions. The continuous

 doi:10.20944/preprints202410.1897.v1

measure is the expectation measure of a Poisson point process. Each of these are examples of simple Poisson point processes. The result follows because a countable sum of Poisson point processes is a Poisson random process. □

It is also possible to construct Poisson point processes on finite topological spaces.

Theorem 10. *Let v denote a valuation on the topology of a finite set $\mathbb{A}$. Assume that the topology satisfies the separation condition T_0. Then there exists an outcome space Ω with a probability measure P and a transition kernel $\omega \to \mu_\omega$ from Ω to $\mathbb{A}$ such that*

- *$\mu_\omega(B)$ is Poisson distributed for any open set B.*
- *For any open sets A, B, C the random variable $\mu_\omega(A)$ is independent of the random variable $\mu_\omega(B)$ given the random variable $\mu_\omega(C)$ if and only if $A \cap B \subset C$.*

Further for any open set B we have

$$v(B) = \int_\Omega \mu_\omega(B)\, \mathrm{d}P\omega. \tag{56}$$

Proof. First we determine the measure μ on $\mathbb{A}$ such that $v(B) = \mu(B)$ for any open set B. Then we construct independent Poisson distributed random variables X_a such that $X_a \sim Po(\mu(a))$. Then, for any open set B, we define a random variable

$$Y_B = \sum_{a \in B} X_a. \tag{57}$$

With these definitions

$$\begin{aligned} E(Y_B) &= E\left(\sum_{a \in B} X_a\right) \\ &= \sum_{a \in B} E(X_a) \\ &= \sum_{a \in B} \mu(a) \\ &= \mu(B) \\ &= v(B). \end{aligned} \tag{58}$$

The conditional independence follows from the construction. □

It is worth noting that for any lattice, the relation $A \wedge B \leq C$ defines a separoid relation (abstract notion of independence) if and only the lattice is distributive. For a distributive lattice the relation $A \wedge B \leq C$ can also be written as $(A \vee C) \wedge (B \vee C) = C$, and this relation is separoid if and only the lattice is modular (see [37] and [[38] Cor. 2].

4.4. Normalization, Conditioning and Other Operations on Expectation Measures

Empirical measures can be added, one can take restrictions and one can find induced measures. Using the same formulas these operations can be performed on expectation measures, but we are not only interested in the formulas but also in probabilistic interpretations.

First we note that any σ-finite measure μ is the expectation measure of some point process. Assume that $\mu = \sum_{i=1}^{\infty} \mu_i$ for some finite measures μ_i. Then

$$\mu = \sum_{i=1}^{\infty} \frac{1}{2^i} \cdot \nu_i \tag{59}$$

where $\nu_i = 2^i \cdot \mu_i$ are finite measures. Thus, μ is a probabilistic mixture of finite measures. Therefore, we just have to prove that any finite measure ν is the expectation measure of a point process. The

 doi:10.20944/preprints202410.1897.v1

normalized measure $\nu/\nu(\mathbb{A})$ has an interpretation as a probability measure, which is the same as a one-point process. We may add n copies of this process to obtain a process with expectation measure $n \cdot \nu/\nu(\mathbb{A})$. If $n \geq \nu(\mathbb{A})$, then this process can be thinned to get a process with expectation measure ν. The following proposition gives probabilistic interpretations of addition, restriction, and inducing for expectation measures via the same operations applied to empirical measures. These equations are proved by simple calculations.

Proposition 2. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $\omega \to \mu_\omega$ and $\omega \to \nu_\omega$ denote point processes with expectation measures μ and ν and with points in $\mathbb{A}$. Let A be a subset of $\mathbb{A}$ and let $f : \mathbb{A} \to \mathbb{B}$ be some mapping. Then*

$$\mu + \nu = \int (\mu_\omega + \nu_\omega) \, \mathrm{d}P\omega, \tag{60}$$

$$\mu_{\cap A} = \int \mu_{\omega \cap A} \, \mathrm{d}P\omega, \tag{61}$$

$$f(\mu) = \int f(\mu_\omega) \, \mathrm{d}P\omega. \tag{62}$$

respectively.

If μ is an expectation measure on the set $\mathbb{A}$, then $\frac{1}{\mu(\mathbb{A})} \cdot \mu$ is a unital measure that we will call *the normalized measure*. Unital measures are normally called probability measures, and our aim is to give a probabilistic interpretation of the normalized measure $\frac{1}{\mu(\mathbb{A})} \cdot \mu$ by specifying an event that has probability equal to $\frac{\mu(B)}{\mu(\mathbb{A})}$.

Theorem 11. *Let B be a subset of $\mathbb{A}$. Let P denote a probability measure on Ω and assume that $\omega \to \mu_\omega$ is a Poisson point process with expectation measure μ. Then*

$$\frac{\mu(B)}{\mu(\mathbb{A})} = \int \frac{\mu_\omega(B)}{\mu_\omega(\mathbb{A})} \, \mathrm{d}P(\omega | \mu_\omega(\mathbb{A}) > 0). \tag{63}$$

Proof. We take the mean of the empirical distribution with respect to P.

$$\begin{aligned}
&\int \frac{\mu_\omega(B)}{\mu_\omega(\mathbb{A})} \, \mathrm{d}P(\omega | \mu_\omega(\mathbb{A}) > 0) \\
&\quad = \sum_{n=1}^{\infty} \left(\int \frac{\mu_\omega(B)}{\mu_\omega(\mathbb{A})} \, \mathrm{d}P(\omega | \mu_\omega(\mathbb{A}) = n) \right) \cdot P(\mu_\omega(\mathbb{A}) = n | \mu_\omega(\mathbb{A}) > 0) \\
&\quad = \sum_{n=1}^{\infty} \left(\int \frac{\mu_\omega(B)}{n} \, \mathrm{d}P(\omega | \mu_\omega(\mathbb{A}) = n) \right) \cdot P(\mu_\omega(\mathbb{A}) = n | \mu_\omega(\mathbb{A}) > 0) \\
&\quad = \sum_{n=1}^{\infty} \frac{\int \mu_\omega(B) \, \mathrm{d}P(\omega | \mu_\omega(\mathbb{A}) = n)}{n} \cdot P(\mu_\omega(\mathbb{A}) = n | \mu_\omega(\mathbb{A}) > 0).
\end{aligned} \tag{64}$$

Now the random variable $\mu_\omega(B)$ is Poisson distributed with mean $\mu_\omega(B)$ and the random variable $\mu_\omega(\complement B)$ is Poisson distributed with mean $\mu_\omega(\complement B)$ and these two random variables are independent.

 doi:10.20944/preprints202410.1897.v1

When we condition on $\mu(B) + \mu(\complement B) = n$ the distribution of $\mu_\omega(B)$ is binomial with mean $n \cdot \frac{\mu(B)}{\mu(\mathbb{A})}$. Hence

$$\begin{aligned}\int \frac{\mu_\omega(B)}{\mu_\omega(\mathbb{A})} \, \mathrm{d}P(\omega|\mu_\omega(\mathbb{A}) > 0) &= \sum_{n=1}^{\infty} \frac{n \cdot \frac{\mu(B)}{\mu(\mathbb{A})}}{n} \cdot P(\mu_\omega(\mathbb{A}) = n|\mu_\omega(\mathbb{A}) > 0) \\ &= \sum_{n=1}^{\infty} \frac{\mu(B)}{\mu(\mathbb{A})} \cdot P(\mu_\omega(\mathbb{A}) = n|\mu_\omega(\mathbb{A}) > 0) \\ &= \frac{\mu(B)}{\mu(\mathbb{A})}.\end{aligned} \tag{65}$$

□

The theorem states that $\frac{\mu(B)}{\mu(\mathbb{A})}$ is the probability of observing a point in B if one first observe a multiset of points as an instance of a point process and then randomly select one of the points from the multiset. This is not very different from random matrix theory, where one first calculate all eigenvalues of a random matrix and then randomly selected one of the eigenvalues. Wigner's semicircular law states that such a randomly selected eigenvalue from a large random matrix has a distribution that approximately follows a semicircular law.

Proposition 2 holds for all point processes, but in Theorem 11 it is required that the point process is a Poisson point process. The following example shows there are point processes where Equations (63) does not hold.

Example 4. *Let $\Omega = \{\omega_1, \omega_2\}$ and assume that $P(\omega_1) = P(\omega_2) = 1/2$. Let $\mathbb{A} = \{a, b\}$, and let $\omega \to \mu_\omega$ denote a process where $\mu_{\omega_1} = 1 \cdot \delta_b$ and $\mu_{\omega_2} = 2 \cdot \delta_a + 1 \cdot \delta_b$. The expectation measure of this process is $\mu = 1 \cdot \delta_a + 1 \cdot \delta_b$. Let $B = \{b\}$. Then the left-hand side of Equations (63) evaluates to*

$$\frac{\mu(B)}{\mu(\mathbb{A})} = \frac{1}{2}. \tag{66}$$

The right-hand side of Equations (63) evaluates to

$$\int \frac{\mu_\omega(B)}{\mu_\omega(\mathbb{A})} \, \mathrm{d}P(\omega|\mu_\omega(\mathbb{A}) > 0) = 1 \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{2} = \frac{2}{3}. \tag{67}$$

The Poisson interpretation of normalized expectation measures carries over to conditional measures.

Corollary 1. *Let P denote a probability measure on Ω and assume that $\omega \to \mu_\omega$ is a Poisson point process with expectation measure μ. Let A and B be subsets of $\mathbb{A}$ with $\mu(A) > 0$. Then*

$$\mu(B|A) = \int \mu_\omega(B|A) \, \mathrm{d}P(\omega|\mu_\omega(A) > 0). \tag{68}$$

Proof. A conditional measure is the normalization of an expectation measure restricted to a subset.

$$\mu(B|A) = \frac{\mu(B \cap A)}{\mu(A)} = \frac{\mu_{\cap A}(B)}{\mu_{\cap A}(\mathbb{A})}. \tag{69}$$

The corollary is proved by applying Theorem 11 to the measure $\mu_{\cap A}$. The condition $\mu(A) > 0$ will ensure that $P(\mu_\omega(A) > 0) > 0$. □

 doi:10.20944/preprints202410.1897.v1

Let μ be an expectation measure on $\mathbb{A}_1$ and let g be a map from $\mathbb{A}_1$ to $\mathbb{A}_2$. Then the induced measure $g(\mu)$ is also an expectation measure. If $g(\mu)(a_2) > 0$ then a Markov kernel $\mu(\cdot|\cdot)$ from $\mathbb{A}_2$ to $\mathbb{A}_1$ is given by

$$\mu(a_1 \mid g(a) = a_2) = \frac{\mu_{\cap g^{-1}(a_2)}(a_1)}{g(\mu)(a_2)}. \tag{70}$$

With this Markov kernel the measure μ can be factored as

$$\mu(a_1) = \mu(a_1 \mid g(a) = a_2) \cdot g(\mu)(g(a_2)). \tag{71}$$

4.5. Independence

The notion of independence plays an important role in the theory of randomness, so we need to define this notion in the context of expectation measures.

Definition 3. *Assume that μ is a measure on $\mathbb{A}$. For $i = 1, 2$ let g_i denote the mappings $\mathbb{A} \to \mathbb{A}_i$. Then we say that g_1 is independent of g_2 if $g_1\left(\mu_{\cap g_2(a)=a_2}\right)$ does not depend on $a_2 \in \mathbb{A}_2$.*

Theorem 12. *Let μ be a measure on $\mathbb{A} = \mathbb{A}_1 \times \mathbb{A}_2$ with projections g_1 and g_2. Then g_1 is independent of g_2 if and only if*

$$\mu(\omega_1, \omega_2) = \frac{\mu_1(\omega_1) \cdot \mu_2(\omega_2)}{\mu(\mathbb{A})} \tag{72}$$

where μ_1 and μ_2 are the marginal measures on $\mathbb{A}_1$ and $\mathbb{A}_2$ respectively.

Proof. We have

$$\begin{aligned} \mu(\omega_1, \omega_2) &= \mu(\omega_1 \mid g(\omega) = \omega_2) \cdot \mu(g(\omega) = \omega_2) \\ &= \mu(\omega_1 \mid g(\omega) = \omega_2) \cdot \mu_2(\omega_2). \end{aligned} \tag{73}$$

Let $\tilde{\mu}$ denote the unital measure on $\mathbb{A}_1$ given by $\tilde{\mu}(\omega_1) = \mu(\omega_1 \mid g(\omega) = \omega_2)$. Then

$$\mu(\omega_1, \omega_2) = \tilde{\mu}(\omega_1) \cdot \mu_2(\omega_2), \tag{74}$$

$$\begin{aligned} \mu_1(\omega_1) &= \sum_{\omega_2 \in \mathbb{A}_2} \mu(\omega_1, \omega_2) \\ &= \sum_{\omega_2 \in \mathbb{A}_2} \tilde{\mu}(\omega_1) \cdot \mu_2(\omega_2) \\ &= \tilde{\mu}(\omega_1) \cdot \mu_2(\mathbb{A}_2) \\ &= \tilde{\mu}(\omega_1) \cdot \mu(\mathbb{A}), \end{aligned} \tag{75}$$

$$\frac{\mu_1(\omega_1)}{\mu(\mathbb{A})} = \tilde{\mu}(\omega_1). \tag{76}$$

□

Note that Equations (72) is the standard way of calculating expected counts in a contingency table under the assumption of independence. Note also that Equations (72) can be rewritten as

$$\frac{\mu(\omega_1, \omega_2)}{\mu(\mathbb{A})} = \frac{\mu_1(\omega_1)}{\mu_1(\mathbb{A}_1)} \cdot \frac{\mu_2(\omega_2)}{\mu_2(\mathbb{A}_2)}, \tag{77}$$

which is the well-known equation that states that for independent variables the joint probability is the product of the marginal probabilities.

 doi:10.20944/preprints202410.1897.v1

4.6. Information Divergence for Expectation Measures

Let P and Q be discrete probability measures. Then *Kullback-Leibler divergence* is defined by

$$D(P\|Q) = \begin{cases} \sum_i P(i) \ln\left(\frac{P(i)}{Q(i)}\right), & \text{if } P \preceq Q; \\ \infty, & \text{else.} \end{cases} \tag{78}$$

For arbitrary discrete measures μ and ν we define *information divergence* by extending Equations (78) via the following formula

$$D(\mu\|\nu) = \sum_i \mu(i) \ln\left(\frac{\mu(i)}{\nu(i)}\right) - \mu(i) + \nu(i). \tag{79}$$

With this definition information divergence becomes a Csiszár f-divergence, and it gives a continuous function from the cones of discrete measures to the lower reals $\overrightarrow{[0,\infty]}$.

Proposition 3. *Let μ and ν denote two finite measures. Then*

$$D(Po(\mu)\|Po(\nu)) = D(\mu\|\nu). \tag{80}$$

Note that on the left-hand side is a KL-divergence for probability measures, while the right-hand side is an information divergence for expectation measures.

Proof. First assume that the outcome space is a singleton so that μ and ν are elements in $[0,\infty[$. Then

$$\begin{aligned} D(Po(\mu)\|Po(\nu)) &= \sum_{i=0}^{\infty} Po(i,\mu) \ln\left(\frac{Po(i,\mu)}{Po(i,\nu)}\right) \\ &= \sum_{i=0}^{\infty} Po(i,\mu) \ln\left(\frac{\frac{\mu^i}{i!}\exp(-\mu)}{\frac{\nu^i}{i!}\exp(-\nu)}\right) \\ &= \sum_{i=0}^{\infty} Po(i,\mu)\left(i \ln\left(\frac{\mu}{\nu}\right) - \mu + \nu\right) \\ &= \mu \ln\left(\frac{\mu}{\nu}\right) - \mu + \nu. \end{aligned} \tag{81}$$

The results follows because KL-divergence is additive on product measures. □

Information divergence is a Csiszár f-divergence, but it is also a Bregman divergence defined on the cone of discrete measures, and except for a constant factor it is the only Bregman divergence that is also a Csiszár f-divergence. On an alphabet with at least three letters KL-divergence may (except for a constant factor) also be characterized as the only Bregman divergence that satisfies a data processing inequality for Markov kernels of unital measures, and there are a number of equivalent characterizations [39], if the alphabet has at least three letters. Here we focus on the convex cone of measures rather than the simplex of probability measures. Therefore it is interesting to note that information divergence has a characterization on a one letter alphabet as a Bregman divergence that satisfies the following property called 1-homogenuity.

$$D(\alpha \cdot \mu\|\alpha \cdot \nu) = \alpha \cdot D(\mu\|\nu). \tag{82}$$

Theorem 13. *Assume that $d : \mathbb{R}_+^2 \to \mathbb{R}_{0,+}$ is a function that satisfies the following conditions.*

- *$d(x,y) \geq 0$ with equality when $x = y$.*
- *$\sum_{i=1}^n d(x_i, y)$ is minimal when $y = \bar{x}$.*
- *$d(\alpha \cdot x, \alpha \cdot y) = \alpha \cdot d(x,y)$ for all $\alpha > 0$.*

 doi:10.20944/preprints202410.1897.v1

Then there exists a positive constant c such that $d(x,y) = c \cdot \left(x \ln\left(\frac{x}{y}\right) - (x-y)\right)$.

Proof. The first two conditions imply that d is a Bregman divergence [[39] Proposition 4]so there exists a strictly convex function g such that $d(x,y) = g(x) - (g(y) + (x-y) \cdot g'(y))$. According to property 3. we have

$$d(\alpha \cdot x, \alpha \cdot y) = \alpha \cdot d(x,y), \tag{83}$$

$$g(\alpha \cdot x) - \big(g(\alpha \cdot y) + (\alpha \cdot x - \alpha \cdot y) \cdot g'(\alpha \cdot y)\big) = \alpha \cdot \big(g(x) - \big(g(y) + (x-y) \cdot g'(y)\big)\big). \tag{84}$$

We differentiate with respect to y and get

$$-\Big(\alpha \cdot g'(\alpha \cdot y) - \alpha \cdot g'(\alpha \cdot y) - \alpha^2 \cdot y \cdot g''(\alpha \cdot y)\Big) = \alpha \cdot \big(-g'(y) + g'(y) + y \cdot g''(y)\big), \tag{85}$$

$$\alpha^2 \cdot y \cdot g''(\alpha \cdot y) = \alpha \cdot y \cdot g''(y), \tag{86}$$

$$\alpha \cdot y \cdot g''(\alpha \cdot y) = y \cdot g''(y). \tag{87}$$

This holds for all $\alpha, y > 0$ so $g''(y) \cdot y = c$ for some constant $c > 0$.

$$g''(y) \cdot y = c, \tag{88}$$

$$g''(y) = \frac{c}{y}, \tag{89}$$

$$g'(y) = c \cdot \ln(y), \tag{90}$$

$$g(y) = c \cdot (y \ln(y) - y + k), \tag{91}$$

for some constant k. Hence

$$\begin{aligned} d(x,y) &= c(x \ln(x) - x + k) - (c(y \ln(y) - y + k) + (x-y) \cdot c \cdot \ln(y)) \\ &= c \cdot (x \ln(x) - x + k - (y \ln(y) - y + k + (x-y) \cdot \ln(y))) \\ &= c \cdot (x \ln(x) - x - y \ln(y) + y - (x-y) \cdot \ln(y)) \\ &= c \cdot \left(x \ln\left(\frac{x}{y}\right) - x + y\right). \end{aligned} \tag{92}$$

□

The following theorem can be proved in the same way as similar theorems in [34,35].

Theorem 14. *Let P be a Bernoulli sum on $\mathcal{M}(\mathbb{A})$ with expectation measure $\pi(P) = \mu$. Then*

$$D(T_{1/n}(P^{*n}) \| Po(\mu)) \to 0 \tag{93}$$

for $n \to \infty$.

Let $\mathcal{C}$ denote a convex set of measures. Then $D(\mathcal{C}\|\nu)$ is defined as $\inf_{\mu \in \mathcal{C}} D(\mu\|\nu)$. If $D(\mathcal{C}\|\nu) < \infty$ and the measure $\mu^* \in \mathcal{C}$ satisfies $D(\mu^*\|\nu) = D(\mathcal{C}\|\nu)$ then μ^* is called the *information projection* of ν on $\mathcal{C}$ [40–42].

Proposition 4. *Let ν be a measure of a finite alphabet $\mathbb{A}$, and let $\mathcal{C}$ be a convex set of measures on $\mathbb{A}$. If μ^* is the information projection of ν on $\mathcal{C}$ then $Po(\mu^*)$ is the information projection of $Po(\nu)$ on the convex hull of the probability measures of the form $Po(\mu)$ where $\mu \in \mathcal{C}$.*

 doi:10.20944/preprints202410.1897.v1

Proof. The measure μ^* is the information projection of ν if and only if the following Pythagorean inequality

$$D(\mu\|\nu) \geq D(\mu\|\mu^*) + D(\mu^*\|\nu) \tag{94}$$

is satisfied for all $\mu \in \mathcal{C}$ [[42] Theorem 8]. Now we have

$$\begin{aligned} D(Po(\mu)\|Po(\nu)) &= D(\mu\|\nu) \\ &\geq D(\mu\|\mu^*) + D(\mu^*\|\nu) \\ &= D(Po(\mu)\|Po(\mu^*)) + D(Po(\mu^*)\|Po(\nu)). \end{aligned} \tag{95}$$

Since a Pythagorean inequality is satisfied for $Po(\mu^*)$, it must be the information projection of $Po(\nu)$ on the convex hull of the distributions $Po(\mu)$ where $\mu \in \mathcal{C}$. □

The reversed information projection is defined similar to the definition of the information projection [2,43–45]. If $\mathcal{C}$ is a convex set, then $D(\mu\|\mathcal{C})$ is defined as $\inf_{\nu\in\mathcal{C}} D(\mu\|\nu)$. If $D(\mu\|\mathcal{C}) < \infty$ then $\hat{\nu} \in \mathcal{C}$ is said to be the reversed information projection of μ on $\mathcal{C}$ if $D(\mu\|\hat{\nu}) = D(\mu\|\mathcal{C})$.

Proposition 5. *Let μ be a measure of a finite alphabet $\mathbb{A}$, and let $\mathcal{C}$ be a convex set of measures on $\mathbb{A}$. Assume that $\hat{\nu} \in \mathcal{C}$ and that $D(\mu\|\mathcal{C}) < \infty$. Then the probability measure $Po(\hat{\nu})$ is the reverse information projection of $Po(\mu)$ on the convex hull of the set of probability measures $\{Po(\nu) \mid \nu \in \mathcal{C}\}$.*

Proof. According to the so-called four point property by Csiszár and Tusnády [43] the measure $\hat{\nu}$ is the reverse information projection of μ on $\mathcal{C}$ if and only if

$$\sum_{\omega\in\Omega} \left(\frac{\mu(\omega)}{\hat{\nu}(\omega)} \cdot \nu(\omega) - \mu(\omega) + \hat{\nu}(\omega) - \nu(\omega) \right) \leq 0. \tag{96}$$

for all ν in $\mathcal{C}$. The probability measure $Po(\mu)$ has outcome space $\mathbb{N}_0^\Omega$. For $\vec{j} \in \mathbb{N}_0^\Omega$ we will write $Po\left(\mu, \vec{j}\right)$ as short for $\prod_{\omega\in\Omega} Po(P(\omega), j_\omega)$. We calculate

$$\begin{aligned} \sum_{\vec{j}\in\mathbb{N}_0^\Omega} \frac{Po\left(\mu,\vec{j}\right)}{Po\left(\hat{\nu},\vec{j}\right)} \cdot Po\left(\nu,\vec{j}\right) &= \sum_{\vec{j}\in\mathbb{N}^\Omega} \prod_{\omega\in\Omega} \frac{Po(\mu(\omega), j_\omega)}{Po(\hat{\nu}(\omega), j_\omega)} \cdot Po(\nu(\omega), j_\omega) \\ &= \prod_{\omega\in\Omega} \sum_{j\in\mathbb{N}_0} \frac{Po(\mu(\omega), j)}{Po(\hat{\nu}(\omega), j)} \cdot Po(\nu(\omega), j) \\ &= \prod_{\omega\in\Omega} \sum_{j\in\mathbb{N}_0} \frac{\frac{\mu(\omega)^j}{j!}\exp(-\mu(\omega))}{\frac{\hat{\nu}(\omega)^j}{j!}\exp(-\hat{\nu}(\omega))} \cdot \frac{\nu(\omega)^j}{j!}\exp(-\nu(\omega)) \\ &= \prod_{\omega\in\Omega} \sum_{j\in\mathbb{N}_0} \frac{\left(\frac{\nu(\omega)}{\hat{\nu}(\omega)} \cdot \nu(\omega)\right)^j}{j!} \exp(-\mu(\omega) + \hat{\nu}(\omega) - \nu(\omega)) \\ &= \prod_{\omega\in\Omega} \exp\left(\frac{\mu(\omega)}{\hat{\nu}(\omega)} \cdot \nu(\omega)\right) \exp(-\mu(\omega) + \hat{\nu}(\omega) - \nu(\omega)) \\ &= \exp\left(\left(\sum_{\omega\in\Omega} \frac{\mu(\omega)}{\hat{\nu}(\omega)} \cdot \nu(\omega) - \mu(\omega) + \hat{\nu}(\omega) - \nu(\omega)\right)\right) \\ &\leq \exp(0) = 1. \end{aligned} \tag{97}$$

 doi:10.20944/preprints202410.1897.v1

Therefore

$$\sum_{\vec{j}\in\mathbb{N}_0^\Omega}\left(\frac{Po\left(\mu,\vec{j}\right)}{Po\left(\hat{v},\vec{j}\right)}\cdot Po\left(\nu,\vec{j}\right)-Po\left(\mu,\vec{j}\right)+Po\left(\hat{v},\vec{j}\right)-Po\left(\nu,\vec{j}\right)\right)\leq 1-1+1-1 \\ =0 \tag{98}$$

for all $\nu\in\mathcal{C}$. □

5. Applications

In this section we will present some examples of how expectation measures can be used to give a new way to handle some problems from statistics and probability theory.

5.1. Goodness-of-Fit Tests

Here we will take a new look at testing Good-of-Fit in one of the simplest possible setups. We will test if a coin is fair, and we perform an experiment where we count the number of heads and tails after tossing the coin a number of times. Let X denote the number of heads and let Y denote the number of tails. Our null hypothesis is that there is symmetry between heads and tails. Here we will analyze the case when we have observed $X=\ell$ and $Y=m$.

Typically, one will fix the number of tosses so that $X+Y=n$, and assume that X has a binomial distribution with success probability p. First we will look at an example where the null hypothesis states that $p=1/2$ and $n=20$.

The maximum likelihood estimate of p is ℓ/n. The divergence is

$$D(\,bin(n,\ell/n)\|bin(n,1/2))=n\cdot\left(\frac{\ell}{n}\ln\left(\frac{\ell/n}{1/2}\right)+\frac{m}{n}\ln\left(\frac{m/n}{1/2}\right)\right). \tag{99}$$

We introduce the signed log-likelihood as

$$G_n(x)=\begin{cases}-\left(2\cdot D(\,bin(n,x/n)\|bin(n,1/2))\right)^{1/2}, & \text{if } x<n/2;\\ +\left(2\cdot D(\,bin(n,x/n)\|bin(n,1/2))\right)^{1/2}, & \text{if } x\geq n/2.\end{cases} \tag{100}$$

so that

$$D(\,bin(n,\ell/n)\|bin(n,1/2))=\frac{1}{2}(G_n(x))^2 \tag{101}$$

In [[46] Cor 7.2] it is proved that

$$\Pr(X<k)\leq\Phi(G_n(k))\leq\Pr(X\leq k) \tag{102}$$

where Φ denotes the distribution function of a standard Gaussian distribution. A QQ plot with a Gaussian distribution on the first axis and the distribution of $G_n(X)$ on the second axis one gets a staircase function with horizontal steps each intersecting the line $x=y$ corresponding to a perfect match between the distribution of $G_n(X)$ and a standard Gaussian distribution as illustrated in Figure 2.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 October 2024 doi:10.20944/preprints202410.1897.v1

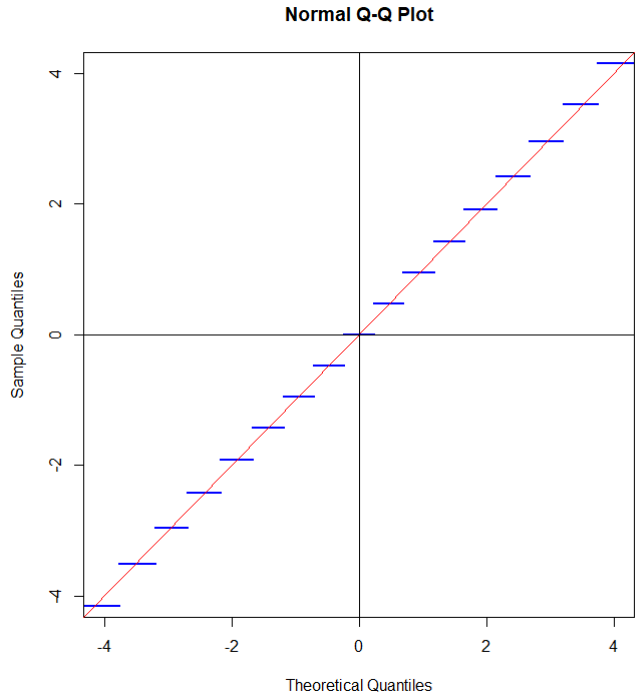

Figure 2. QQ-plot of a standard Gaussian distribution against the distribution of $G_{20}(X)$, where X has a binomial distribution with $n = 20$ and $p = 1/2$.

Instead of using the Gaussian approximation one could calculate tail probabilities exactly (Fisher's exact test), but as we shall see below one can even do better.

In expectation theory, it is more natural to assume that X and Y are independent Poisson distributed random variables with mean values λ and μ respectively. In our analysis, the null hypothesis states that $\lambda = \mu$.

Since $X + Y \sim Po(\lambda + \mu)$ the maximum likelihood estimate of $\lambda + \mu$ is $\ell + m$. Hence the estimate of λ and μ are $(\ell + m)/2$. Here we define the random variable $N = X + Y$ and $n = \ell + m$. We calculate the divergence

$$D\Big(Po(\ell) \otimes Po(m) \| Po\Big(\frac{n}{2}\Big) \otimes Po\Big(\frac{n}{2}\Big)\Big) = \ell \ln\Big(\frac{\ell}{n/2}\Big) + m \ln\Big(\frac{m}{n/2}\Big). \tag{103}$$

i.e., the same expression as in the classical analysis. Since X is binomial given that $X + Y = n$ we have

$$\Pr(X < k \mid N = n) \leq \Phi(G_n(k)) \leq \Pr(X \leq k \mid N = n), \tag{104}$$

Since the distribution of $G_N(X)$ is close to a standard Gaussian distribution under the condition $N = n$ the same is true for $G_N(X)$ when we take the mean value over the Poisson distributed variable N. Since each of the steps intersects the straight line near the mid-point of the step the effect of taking the mean value with respect to N is that the steps to a large extent cancel out as illustrated in Figure 3.

The only step that partly survive the randomization is the step $G(X) = 0$. For the binomial distribution, the length of this step is determined by $P(X = n/2) = 0.1762$. If the sample size is Poisson distributed then the length of the step is determined as

$$E[P(X = N/2)] = \sum_{n \text{ even}} \frac{20^n}{n!} \exp(-20) \cdot \frac{n!}{\left(\frac{n}{2}!\right)^2} 2^{-n} \tag{105}$$

$$= 0.0898\,, \tag{106}$$

 doi:10.20944/preprints202410.1897.v1

which is about half of the value for the binomial distribution. The reason for this is that the probability that a Poisson distributed random variable is even is approximately $1/2$. If we tested $p = 1/3$ we would get a step of length about one third of the corresponding step for the binomial distribution. In a sense, testing $p = 1/2$ gives the most significant deviation from a Gaussian distribution.

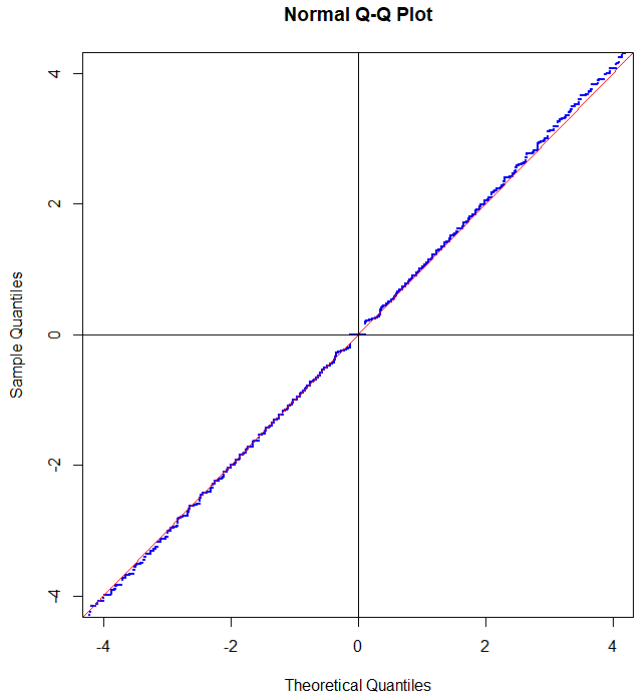

Figure 3. QQ-plot of a standard Gaussian distribution against the distribution of $G_N(X)$, where X has a Poisson distributed with mean $\lambda = 10$ and $N = X + Y$ where Y is also Poisson distributed with mean 10 and Y is independent of X.

If we square $G(X)$, we get 2 times divergence, which is often called the G^2-statistic. Due to symmetry between head and tail, the intersection property is also satisfied when the distribution of the G^2-statistic is compared with a χ^2-distribution [47]. This is illustrated in Figure 4.

For statistical analysis, one should not fix the sample size before sampling. A better procedure is to sample for a specific time so that the sample size becomes a random variable. In practice, this is often how sampling takes place and if the sample size is really random, it may even be misleading to analyze data as if n was fixed.

 doi:10.20944/preprints202410.1897.v1

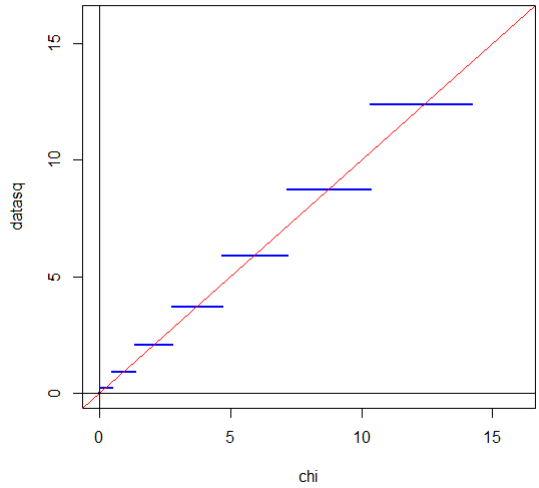

Figure 4. QQ-plot of the χ^2-distribution with $df = 1$ against the distribution of the G^2-statistic for testing $p = 1/2$ in $bin(20, p)$.

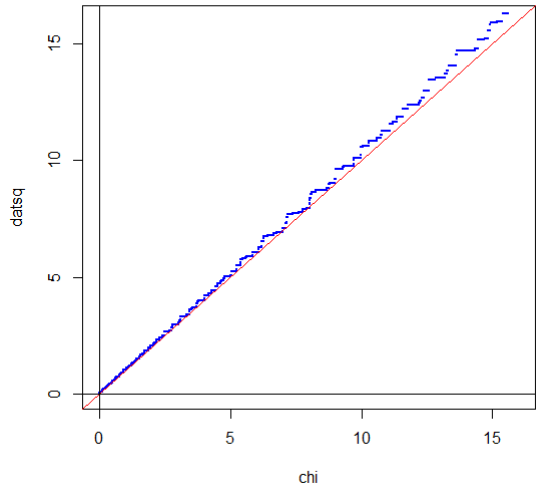

Figure 5. QQ-plot of the χ^2-distribution with $df = 1$ against the distribution of the G^2-statistic for testing $\lambda = \mu$ based on $\hat{\lambda} + \hat{\mu} = 20$. We see that there is a tiny but systematic deviation from the straight line in that the values of G^2 are a little larger than predicted by the χ^2-distribution. A continuity correction as described in [48] may handle this minor issue.

5.2. Improper Prior Distributions

Prior distributions play a major role in Bayesian statistics. A detailed discussion about how prior distributions can be determined in various cases is beyond the topic of this article. We will refer to [49] for a review of the subject including a long list of references. See [50] for a more information theoretic treatment of prior distributions.

In Bayesian statistics, a justification of how posterior distributions are calculated is normally based on a probabilistic interpretation of the prior distribution. It is well-known that proportional prior distributions will lead to the same posterior distribution. For this reason the total weight of the prior distribution is not important as long as the total weight is finite. With a finite total weight the total weight can always be normalized in order to obtain a unital measure, and unital measures allow a

 doi:10.20944/preprints202410.1897.v1

probabilistic interpretation within Bayesian statistics. A significant problem in Bayesian statistics is the use of improper prior distributions, i.e., prior distributions described by measures with infinite total mass. Such a prior is problematic in that it does not allow a literal interpretation in Bayesian statistics. If we replace probability measures by expectation measures, this problem disappears. We will just give a simple example of how improper prior distributions can be given probabilistic interpretation using the Poisson interpretation of expectation measures.

Consider a statistical model where a random variable X has distribution given by the Markov kernel

$$P(X = \mu - 1) = P(X = \mu + 1) = 1/2. \tag{107}$$

We assume that the mean value parameter μ is integer valued. We choose the counting measure times λ on $\mathbb{Z}$ as an improper prior distribution for the mean value parameter. There are a number of ways or arguing in favor of this prior distribution. For instance the counting measure is invariant under integer translations. Except for a multiplicative constant translation invariance uniquely determines the prior measure as a Haar measure. Combining λ times the counting measure on $\mathbb{Z}$ with the Markov kernel (107) we get a joint measure on $\mathbb{Z}^2$ supported on the points illustrated in Figure 6.

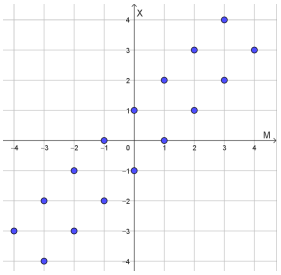

Figure 6. Support of the joint measure.

Each point supporting the joint measure will have weight $\lambda/2$. The marginal measure on X is λ times the counting measure on $\mathbb{Z}$ as discussed in Section 4.4. Now the joint measure can be factored into λ times the counting measure on $\mathbb{Z}$ and a Markov kernel

$$P(M = x - 1) = P(M = x + 1) = 1/2. \tag{108}$$

The Markov kernel (108) can be considered as the posterior distribution of M given that $X = x$.

These calculations can be justified within a probabilistic setting via point processes. First, the prior is represented by a Poisson point process with λ times the counting measure on $\mathbb{Z}$ as expectation measure. If the Markov kernel (107) is applied to this process, then we obtain a Poisson point process on the points in $\mathbb{Z}^2$ illustrated in Figure 6 with $\lambda/2$ times the counting measure as expectation measure. For an instance of this joint point process let N_x denote the number of points with x as second coordinate. We will condition on $X = x$ so we will assume that $N_x = n > 0$. Among these n points that all have second coordinate equal to x, we choose a random point according to a uniform distribution, i.e., each point is selected with probability $1/n$. The selected point has coordinates (M, x) where $M = x \pm 1$. We want to calculate the distribution of M for a given value of x. According to Corollary 1 we get $P(M = x \pm 1 \mid X = x) = 1/2$ and this is our posterior distribution.

 doi:10.20944/preprints202410.1897.v1

This derivation works for any value of $\lambda > 0$. In particular we will get the same result if we thin the process by a factor $\alpha > 0$, i.e., if we replace λ by $\alpha \cdot \lambda$. The derivation involves selecting 1 out of N_x points under the condition that $N_x \geq 1$. If $N_x \sim Po(\alpha \cdot \lambda)$ we have

$$\begin{aligned} P(N = 1 \mid N \geq 1) &= \frac{\alpha \cdot \lambda \exp(-\alpha \cdot \lambda)}{1 - \exp(-\alpha \cdot \lambda)} \\ &= \frac{\alpha \cdot \lambda}{\exp(\alpha \cdot \lambda) - 1} \\ &\to 1 \end{aligned} \tag{109}$$

for $\alpha \to 0$. Therefore, with high probability there will only be one point that has second coordinate x if the point process has been thinned with a small value of α. Hence, the step, where we randomly select one of the points, becomes almost obsolete if we thinned the process sufficiently.

If the Poisson point process is a process in time, then there will be a first point for which $X = x$ and the probability of $M = m$ will be the probability that this first point satisfies $M = m$. With a process in time, one can introduce a stopping time that stops the process when $X = x$ has been observed for the first time. This often simplifies the interpretation, but exactly this kind of reasoning goes wrong in many presentations of the double-slit experiment (Example 3). In this experiment, the point process is not a process in time because no arrival times for the photons are recorded. If one had precise records of time, the uncertainty for energy and frequency would destroy the interference pattern.

5.3. Markov Chains

The idea of randomizing the sample size has various consequences, and sometimes this leads to great simplifications. Let Φ denote some Markov operator that generates a Markov chain. The time average is usually defined as

$$\Phi_n = \frac{1}{n} \sum_{i=0}^{n-1} \Phi^i \tag{110}$$

Many properties of such time averages are known, but what complicates the matter is that the composition of Φ_m and Φ_n is not a time average. If we instead define

$$\Psi_t = \sum_{i=0}^{\infty} \frac{t^i}{i!} \exp(-t) \cdot \Phi^i \tag{111}$$

then Ψ_t is a Markov operator and $\Psi_s \circ \Psi_t = \Psi_{s+t}$. The Markov operators Ψ_t generate a Markov process in continuous time, and this Markov process tend to have better properties than the original Markov chain. For instance, all recurrent states under Φ get positive transition probabilities under Ψ. The same idea was also used in [51] to prove that if Φ is an affine map of a convex body into itself, then Ψ_t converges to a retraction of the convex body to the set of fix points of the affine map Φ. Many theorems in probability involving averages should be revisited to see what the consequences are if the usual average is replaced by taking a weighted average with Poisson distributed weights.

 doi:10.20944/preprints202410.1897.v1

5.4. Inequalities for Information Projections

Let $Q = (q_1, q_2, \dots, q_n)$ be a unital measure on a finite set and let C denote the convex set of unital measures such that the mean value of $f(i)$ is ν. Then the information projection P of Q on C can be determined by using Lagrange multipliers.

$$\ln\left(\frac{P(i)}{Q(i)}\right) + \frac{P(i)}{P(i)} - 1 = \beta \cdot f(i) + \gamma \cdot 1, \tag{112}$$

$$\frac{P(i)}{Q(i)} = \exp(\beta \cdot f(i) + \gamma), \tag{113}$$

$$P(i) = \exp(\beta \cdot f(i)) \cdot \exp(\gamma) \cdot Q(i). \tag{114}$$

The moment generating function is defined by $Z(\beta) = \sum_i \exp(\beta \cdot f(i)) \cdot Q(i)$ and in order to get a unital measure we shall choose $\gamma = -\ln(Z(\beta))$. Thus, P is element in the *exponential family*

$$P(i) = \frac{\exp(\beta \cdot f(i))}{Z(\beta)} \cdot Q(i). \tag{115}$$

The mean value is $\frac{Z'(\beta)}{Z(\beta)}$ and β should be chosen so that this equals ν.

If we drop the condition that the projection should be unital, we get simplified expressions. The Lagrange equation becomes

$$\ln\left(\frac{P_\beta(i)}{Q(i)}\right) + \frac{P_\beta(i)}{P_\beta(i)} - 1 = \beta \cdot f(i), \tag{116}$$

$$P_\beta(i) = \exp(\beta \cdot f(i))Q(i), \tag{117}$$

and the mean value of this measure is $Z'(\beta)$.

Theorem 15. *Let Q be a unital measure on a finite set and let X be a random variable such that*

$$\int X \,\mathrm{d}Q = 0, \tag{118}$$

$$\int X^2 \,\mathrm{d}Q = 1, \tag{119}$$

$$\int X^3 \,\mathrm{d}Q < 0. \tag{120}$$

Then there exists $\delta > 0$ such that

$$D(P\|Q) \geq \frac{1}{2}\left(\int X \,\mathrm{d}P\right)^2 \tag{121}$$

for all measures satisfying $\int X \,\mathrm{d}P \in [0, \delta]$.

Proof. For a fixed value of $\int X \,\mathrm{d}P$ the right-hand side is minimized for the distribution P_β determined by (117), so it is sufficient to prove the inequality for $P = P_\beta$. Thus, we have to prove that

$$\beta \cdot Z''(\beta) - Z(\beta) + \beta \geq \frac{1}{2}(Z'(\beta))^2 \tag{122}$$

for $0 \leq \beta \leq \delta$ for some $\delta > 0$. We differentiate both side, and it is sufficient to prove that

$$\beta \cdot Z''(\beta) \geq Z'(\beta) \cdot Z''(\beta), \tag{123}$$

$$\beta \geq Z'(\beta), \tag{124}$$

 doi:10.20944/preprints202410.1897.v1

because $Z''(\beta) > 0$. We differentiate once more, and we have to prove that

$$1 \geq Z''(\beta). \tag{125}$$

Now we differentiate one more time, and we have to prove that

$$0 \geq Z'''(\beta). \tag{126}$$

for $0 \leq \beta \leq \delta$ but this holds by continuity because $Z'''(0) = \int X^3 \, \mathrm{d}Q < 0$. □

Similar results hold for measures on infinite sets, but in these cases one should be careful to choose the parameters so that the information projection exists. In a number of important cases, the inequality above may be extended to all positive (or negative) values rather than positive (or negative) values in a neighborhood of zero. Some of these cases are mentioned below.

Hypergeometric distributions can be approximated by binomial distributions. A lower bound is given by the following inequality [52].

$$D(P\|bin(n,p)) \geq \frac{\left(\sum_{x=0}^{n} \tilde{\mathcal{K}}_2(x;n,p)P(x)\right)^2}{2} \tag{127}$$

where $\tilde{\mathcal{K}}_2(x;n,p)$ is the second normalized Kravchuk polynomial. For hypergeometric distribution $hyp(N,K,n)$ with $\frac{K}{N} = p$ this lower bound can be written as

$$D(hyp(N,K,n)\|bin(n,p)) \geq \frac{n(n-1)}{4(N-1)^2} \tag{128}$$

As demonstrated in [52] this lower bound is tight for fixed values of n and p and large values of the parameters N and K.

Binomial distributions can be approximated by Poisson distributions. For any distribution P we have the following lower bound on information divergence.

$$D(P\|Po(\lambda)) \geq \frac{\left(\sum_{x=0}^{\infty} \tilde{C}_2^{\lambda}(x)P(x)\right)^2}{2} \tag{129}$$

where $\tilde{C}_2^{\lambda}(x)$ denotes the second normalized Poisson-Charlier polynomial [53]. If P is binomial $bin(n,p)$ and $\lambda = np$ then we get the inequality

$$D(bin(n,p)\|Po(\lambda)) \geq \frac{p^2}{4} \tag{130}$$

and this inequality is tight for fixed λ and n tending to infinity [35,54–56].

In the central limit theorem the rate of convergence is primarily determined by the skewness and if the skewness is zero it is primarily determined by the excess kurtosis. If the skewness is non-zero then lower bounds on divergence involve both skewness and kurtosis, so for simplicity we will only present the case where the skewness is zero and the excess kurtosis is negative. For all measures P for which the integral $\int \tilde{H}_4(x) \, \mathrm{d}P \leq 0$ where $\tilde{H}_4$ denotes the fourth normalized Hermite polynomial, we have

$$D(P\|N(o,1)) \geq \frac{\left(\int \tilde{H}_4(x) \, \mathrm{d}Px\right)^2}{2}. \tag{131}$$

If P is a unital measure with variance excess kurtosis κ then according to [[57] Theorem 7] this inequality reduces to

$$D(P\|N(0,1)) \geq \frac{\kappa^2}{48}. \tag{132}$$

 doi:10.20944/preprints202410.1897.v1

Lower bounds for the rate of convergence have been discussed in the papers [57,57–59] where this and similar inequalities were treated in great detail.

For all these inequalities, it gives simplifications if we do not require that the measures are unital. The hard part is to prove that the inequalities do not only hold in a neighborhood of zero, but that they hold for all values of interest. This hard part of the proofs is still hard without the assumption that the measures are unital, but it should be noted that the hard part is not relevant if we are only interested in lower bounds on the rate of convergence.

6. Discussion and Conclusions

Expectation theory may be considered an alternative to Kolmogorov's theory of probability. Still, it may be better to view the two theories as two ways of describing the same situations using measure theory. The language of probability theory focuses on experiments where a single outcome is classified according to mutually exclusive classes. The language of expectation theory focuses on experiments where the results are given as tables of frequencies. There will be no inconsistency if the two languages are mixed. Expectation measures can be understood within probability theory, as is already the case in the theory of point processes. If our understanding of randomness is based on expectation measures, then an expectation measure can be interpreted as the expectation measure of a process in a larger outcome space.

In this paper we have only worked out the basic framework and interpretation of expectation theory. This opens up a lot of new research questions. For instance, it should be possible to justify the use of Haar measures as prior distributions rigorously in the same way as Haar probability measures on compact groups can be justified as being probability measures that maximize the rate-distortion function [60]. In expectation theory, it is natural to consider sampling with a randomized sample size. Since many results in probability theory are formulated for fixed sample sizes there is a lot of work to be done to generalize the results to cases with random sample sizes. Our present results suggest that simpler or stronger results can be obtained, but in many cases new techniques may be needed.

In this paper, we discussed finite expectation measures and valuations on topological spaces and downsets of posets that satisfy the DCC. The valuation approach is more general, but it is an open research question which category is most useful for further development of expectation theory.

Quantum information theory can be based on generalized probabilistic theories. In these theories a measurement is defined as something that maps a preparation into a probability measure. A *state* is defined as an equivalence classes of preparations that cannot be distinguished be any measurement. With a shift from probability measures to expectation measures one should similarly change the focus in quantum theory from states that are represented by density operators (positive operators with trace 1) to positive trace class operators. Thus, the focus should shift from the convex state space to the state cone. For applications in quantum information theory it is still too early to say what the impact will be, but shifting the focus away from mutually exclusive events may circumvent some of the paradoxes that have haunted the foundation of quantum theory for more than a century.

Below there is a list of concepts from probability theory and standard quantum information theory and how they relate to the concepts that have been introduced in the present paper.

 doi:10.20944/preprints202410.1897.v1

Probability theory	**Expectation theory**
Probability	Expected value
Outcome	Instance
Outcome space	Multiset monad
P-value	E-Value
Probability measure	Expectation measure
Binomial distribution	Poisson distribution
Density	Intensity
Bernoulli random variable	Count variable
Empirical distribution	Empirical measure
KL-divergence	Information divergence
Uniform distribution	Poisson point process
State space	State cone

Funding: This research received no external funding.

Acknowledgments: I want to thank Peter Grünwald and Tyron Lardy for stimulating discussions related to this topic.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

bin	binomial distribution
DCC	decending chain condition
E-statistic	Evidence statistic
E-value	Observed value of an E-statistic
hyp	Hypergeometric distribution
IID	Independent identically distributed
KL-divergence	Information divergence restricted to probability measures
MDL	Minimum description length
Mset	Multiset
N	Gaussian distribution
PM	Probability measure
Po	Poisson distribution
mset	Multiset
Poset	Partially ordered set
Pr	Probability

References

1. Kolmogorov, A.N. *Grundbegriffe der Wahrscheinlichkeitsrechnung*; Springer: Berlin, 1933.
2. Lardy, T.; Grünwald, P.; Harremoës, P. Reverse Information Projections and Optimal E-statistics. *IEEE Transactions on Information Theory* **2024**, pp. 1–1. doi:10.1109/TIT.2024.3444458.
3. Perrone, P. Categorical Probability and Stochastic Dominance in Metric Spaces. phdthesis, Max Planck, Institute for Mathematics in the Sciences, 2018.
4. nLab authors. monads of probability, measures, and valuations. https://ncatlab.org/nlab/show/monads+of+probability%2C+measures%2C+and+valuations, 2024. Revision 45.
5. Shiryaev, A.N. *Probability*; Springer: New York, 1996.
6. Whittle, P. *Probability via Expectation*, 3 ed.; Springer texts in statistics, Springer Verlag: New York, 1992.
7. Kallenberg, O. *Random Measures*; Springer: Schwitzerland, 2017. doi:doi:10.1007/978-3-319-41598-7.
8. Lawvere. The Category of Probabilistic Mappings. Lecture notes.
9. Scibior, A.; Z..; Ghahramani.; Gordon, A.D. Practical probabilistic programming with monads. Proceedings of the 2015 ACM SIGPLAN Symposium on Haskell; Association for Computing Machinery: New York, NY, USA, 2015; Haskell '15, p. 165–176. doi:10.1145/2804302.2804317.

10. Giry, M. A categorical approach to probability theory. Categorical Aspects of Topology and Analysis; Banaschewski, B., Ed.; Springer Berlin Heidelberg: Berlin, Heidelberg, 1982; pp. 68–85.
11. Lieshout, M.V., Spatial Point Process Theory. In *Handbook of Spatial Statistics*; Handbooks of Modern Statistical Methods, Chapman and Hall, 2010; chapter 16.
12. Dash, S.; Staton, S. A Monad for Probabilistic Point Processes. ACT, 2021.
13. Jacobs, B. From Multisets over Distributions to Distributions over Multisets. Proceedings of the 36th Annual ACM/IEEE Symposium on Logic in Computer Science; Association for Computing Machinery: New York, NY, USA, 2021; LICS '21, pp. 1–13, [arXiv:2105.06908]. doi:10.1109/LICS52264.2021.9470678.
14. Rényi, A. A characterization of Poisson processes. *Magyar Tud. Akad. Mat. Kutaló Int. Közl.* **1956**, *1*, 519–527.
15. Kallenberg, O. Limits of Compound and Thinned Point Processes. *Journal of Applied Probability*, *12*, 269–278. doi:10.2307/3212440.
16. nLab authors. valuation (measure theory). https://ncatlab.org/nlab/show/valuation+%28measure+theory%29, 2024. Revision 29.
17. Heckmann, R. Spaces of valuations. In *Papers on General Topology and Applications*; New York Academy of Sciences, 1996. doi:10.1111/j.1749-6632.1996.tb49168.x.
18. Blizard, W.D. The development of multiset theory. *Modern Logic* **1991**, *1*, 319 – 352.
19. Monro, G.P. The Concept of Multiset. *Mathematical Logic Quarterly* **1987**, *33*, 171–178. doi:10.1002/malq.19870330212.
20. Isah, A.; Teella, Y. The Concept of Multiset Category. *British Journal of Mathematics and Computer Science*, *9*, 427–247. doi:10.9734/BJMCS/2015/18415.
21. Grätzer, G. *Lattice Theory*; Dover, 1971.
22. Wille, R. Formal Concept Analysis as Mathematical Theory. In *Formal Concept Analysis*; Ganter, B.; Stumme, G.; Wille, R., Eds.; Number 3626 in Lecture Notes in Artificial Intelligence, Springer, 2005; pp. 1–33.
23. Topsøe, F. Compactness in Space of Measures. *Studia Mathematica* **1970**, *36*, 195–212. doi:Studia Mathematica.
24. Alvarez-Manilla, M. Extension of valuations on locally compact sober spaces. *Topology and its Applications* **2002**, *124*, 397–433. doi:https://doi.org/10.1016/S0166-8641(01)00249-8.
25. Harremoës, P. Extendable MDL. Proceedings ISIT 2013; IEEE Information Theory Society, , 2013; pp. 1516–1520. doi:10.1109/ISIT.2013.6620480.
26. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley, 1991.
27. Csiszár, I. The Method of Types. *IEEE Trans. Inform. Theory* **1998**, *44*, 2505 – 2523. doi:10.1109/18.720546.
28. Harremoës, P. Testing Goodness-of-Fit via Rate Distortion. Information Theory Workshop, Volos, Greece, 2009. IEEE Information Theory Society, 2009, pp. 17–21. doi:10.1109/ITWNIT.2009.5158533.
29. Harremoës, P. The Rate Distortion Test of Normality. Proceedings ISIT 2019. IEEE Information Theory Society, 2019, pp. 241–245. doi:10.1109/ISIT.2019.8849707.
30. Harremoës, P. Rate Distortion Theory for Descriptive Statistics. *Entropy*, *25*, 456. doi:https://doi.org/10.3390/e25030456.
31. Rényi, A. On an Extremal Property of the Poisson Process. *Annals Inst. Stat. Math.* **1964**, *16*, 129–133.
32. McFadden, J.A. The Entropy of a Point Process. *J. Soc. Indst. Appl. Math.* **1965**, *13*, 988–994.
33. Harremoës, P. Binomial and Poisson Distributions as Maximum Entropy Distributions. *IEEE Trans. Inform. Theory* **2001**, *47*, 2039–2041. doi:10.1109/18.930936.
34. Harremoës, P.; Johnson, O.; Kontoyiannis, I. Thinning and Information Projection. 2008 IEEE International Symposium on Information Theory. IEEE, 2008, pp. 2644–2648.
35. Harremoës, P.; Johnson, O.; Kontoyiannis, I. Thinning, Entropy and the Law of Thin Numbers. *IEEE Trans. Inform Theory* **2010**, *56*, 4228 – 4244. doi:10.1109/TIT.2010.2053893.
36. Hillion, E.; Johnson, O. A proof oof the Shepp-Olkin entropy concavity conjecture. *Bernoulli* **2017**, *23*, 3638–3649, [arXiv:1503.01570]. doi:10.3150/16-BEJ860.
37. Dawid, A.P. Separoids: A mathematical framework for conditional independence and irrelevance. *Ann. Math. Artif. Intell.* **2001**, *32*, 335–372. doi:https://doi.org/10.1023/A:1016734104787.
38. Harremoës, P. Entropy inequalities for Lattices. *Entropy* **2018**, *20*, 748. doi:10.3390/e20100784.
39. Harremoës, P. Divergence and Sufficiency for Convex Optimization. *Entropy* **2017**, *19*, Article no. 206, [arXiv:1701.01010]. doi:10.3390/e19050206.
40. Csiszár, I. I-Divergence Geometry of Probability Distributions and Minimization Problems. *Ann. Probab.* **1975**, *3*, 146–158.

41. Pfaffelhuber, E. Minimax Information Gain and Minimum Discrimination Principle. Topics in Information Theory; Csiszár, I.; Elias, P., Eds. János Bolyai Mathematical Society and North-Holland, 1977, Vol. 16, *Colloquia Mathematica Societatis János Bolyai*, pp. 493–519.
42. Topsøe, F. Information Theoretical Optimization Techniques. *Kybernetika* **1979**, *15*, 8 – 27.
43. Csiszár, I.; Tusnády, G. Information geometry and alternating minimization procedures. *Statistics and Decisions* **1984**, *Supplementary Issue 1*, 205.237.
44. Li, J.Q. Estimation of Mixture Models. Ph.d. dissertation, Department of Statistics, Yale University, 1999.
45. Li, J.Q.; Barron, A.R. Mixture Density Estimation. Proceedings Conference on Neural Information Processing Systems: Natural and Synthetic; , 1999.
46. Harremoës, P. Bounds on tail probabilities for negative binomial distributions. *Kybernetika* **2016**, *52*, 943–966. doi:10.14736/kyb-2016-6-0943.
47. Harremoës, P.; Tusnády, G. Information Divergence is more χ^2-distributed than the χ^2-statistic. 2012 IEEE International Symposium on Information Theory; IEEE, , 2012; pp. 538–543. doi:10.1109/ISIT.2012.6284247.
48. Györfi, L.; Harremoës, P.; Tusnády, T. Some Refinements of Large Deviation Tail Probabilities. arXiv: 1205.1005.
49. Kass, R.E.; Wasserman, L.A. The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association* **1996**, *91*, 1343–1370.
50. Grünwald, P. *the Minimum Description Length principle*; MIT Press, 2007.
51. Harremoës, P. Entropy on Spin Factors. Information Geometry and Its Applications; Ay, N.; Gibilisco, P.; Matúš, F., Eds. Springer, 2018, Vol. 252, *Springer Proceedings in Mathematics & Statistics*, pp. 247–278, [arXiv:1707.03222].
52. Harremoës, P.; Matúš, F. Bounds on the Information Divergence for Hypergeometric Distributions. *Kybernetika* **2020**, *56*, 1111–1132. doi:10.14736/kyb-2020-6-1111.
53. Harremoës, P.; Johnson, O.; Kontoyiannis, I. Thinning and Information Projections. arXiv:1601.04255.
54. Harremoës, P. Convergence to the Poisson Distribution in Information Divergence. Preprint 2, Mathematical department, University of Copenhagen, 2003.
55. Harremoës, P.; Ruzankin, P. Rate of Convergence to Poisson Law in Terms of Information Divergence. *IEEE Trans. Inform Theory* **2004**, *50*, 2145–2149. doi:10.1109/TIT.2004.833364.
56. Kontoyiannis, I.; Harremoës, P.; Johnson, O. Entropy and the Law of Small Numbers. *IEEE Trans. Inform. Theory* **2005**, *51*, 466–472. doi:10.1109/TIT.2004.840861.
57. Harremoës, P., Lower Bounds for Divergence in the Central Limit Theorem. In *General Theory of Information Transfer and Combinatorics*; Springer Berlin Heidelberg: Berlin, Heidelberg, 2006; pp. 578–594. doi:10.1007/11889342_35.
58. Harremoës, P. Lower bound on rate of convergence in information theoretic Central Limit Theorem. Book of Abstracts for the Seventh International Symposium on Orthogonal Polynomials, Special functions and Applications; , 2003; pp. 53–54.
59. Harremoës, P. Lower Bounds on Divergence in Central Limit Theorem. *Electronic Notes in Discrete Mathematics* **2005**, *21*, 309–313. doi:10.1016/J.ENDM.2005.07.076.
60. Harremoës, P. Maximum Entropy on Compact groups. *Entropy* **2009**, *11*, 222–237. doi:10.3390/e11020222.