Organodynamics: A General Theory of Dynamical Systems based on Chance Organization

Part IV of V: Prediction Dynamics

Release 1.0, April 2014

Grant Holland; Santa Fe Alliance for Science; Santa Fe, New Mexico;
email: grant.holland at organiccomplexsystems.org; April, 2014.

Dynamics of Stochastic Systems

In physics, *dynamics* refers to accounting for the forces of nature and the effects that they have on the time evolution of physical systems. These dynamics are captured in physics by *laws of motion*. In classical physics, these laws are *deterministic*. This means that they uniquely determine the state of a physical process at any given time in the future, as long as the conditions are known at an established starting time.

Physicists use mathematics to model these dynamics. The mathematics specify an equation, or system of equations, whose solution determines unique future states of the system at specified times, having been given the state of the system at an initial time (initial conditions). These equations are called the *equations of motion*. Very often, physicists model a given application with a system of differential equations whose solution constitutes the equations of motion.

Physics provides a number of different theories for modeling system dynamics. If you are working with Newton's version of mechanics - an historically early "approximation", then you use Newton's equations of motion. Lagrange and Hamilton offered refinements on Newton's mechanics, and under those regimes one uses Lagrangian or Hamiltonian equations of motion.

But these ideas aren't limited to physics. The ideas of forces of nature and their resulting motions can be generalized to that of *change* and its *mechanisms of change* in applications beyond physics. Markets aren't driven directly by the laws of the strong force, the weak force, the electromagnetic force nor gravity. And neither is biology, ecology, or social science. Even certain natural phenomena such as the climate, weather patterns, are difficult to model by directly starting with the four physical forces of nature. Nevertheless, these systems can be characterized as "state that changes over time".

Yet, these extra-physics applications are still driven by their own mechanism of change. And, as in physics, those mechanisms can be modeled with mathematics. Thus, it is reasonable to extend the concept of "dynamical systems" to include them as well. Under that generalization, it may be more appropriate to refer to their "equations of motion" as their "equation of change". In both cases, the dynamics of the system being model is represented by some system of mathematical equations whose dependent variables take on values that represent initial conditions, and whose independent variables represent predictions of future state within the time evolution of the system being modeled.

The field of study within which such a generalization has occurred is *systems theory*. And the type of theory that has taken on that generalization is *dynamical systems theory*. Of course, classical mechanics has provided the initial dynamical systems theories, as discussed. But physics has provided newer dynamical systems theories as well. For example, *statistical mechanics* and *quantum mechanics* and are dynamical systems theories that deal with a novel form of "equations of motion". They both introduce *stochastic dynamics*, in which their "equations of motion" do not predict precise solutions, but merely narrow the possible solution set by the use of probability distributions.

An example of a broadly adopted dynamical systems theory that speaks to applications beyond physics is *nonlinear dynamics* and chaos theory*.* This theory permits the application to be practically any system whose time evolution is characterized by, what it calls, "sensitivity to initial conditions". This phenomenon is also called the "butterfly effect". The mathematics that is used by nonlinear dynamics to model the "forces at work" in driving change within its target systems is nonlinear equations.

Organodynamics is also a dynamical systems theory in this more general since. For the complex systems that it is designed to model, the "mechanisms of change" may or may not directly be forces of nature. However, whatever they are, organodynamics uses the mathematics of probability theory, information theory and stochastic processes to model them.

In fact, like for nonlinear dynamics, the class of applications that organodynamics seeks to model can be characterized by a brief phrase: "systems whose change of state is characterized by the manner in which it is organized, and whose change of organization over time is subject to chance variation." This last aspect, "change of state being subject to chance variation", begs for the application of probability theory. This fact means that organodynamics features *stochastic dynamics*. And that the mathematics underlying organodynamics has more in common statistical mechanics and quantum mechanics than it does with classical mechanics and nonlinear dynamics.

Deterministic, Nondeterministic and Stochastic Dynamical Systems

It is the nature of a *deterministic* dynamical systems theory (such as classical mechanics or nonlinear dynamics) that it is able to provide a mathematical mechanism (equations, rules, algorithms, etc.) that *precisely predicts* the outcomes of the future time steps of a system process, having been given an initial condition. In fact, this determinism is so complete for these theories that the appellation "prediction" is seldom used. Rather, we simply say that specific outcome (for each time step) is calculated.

However, organodynamics is not a deterministic dynamical systems theory, and does not intend, nor desire, to be. This is because the highly complex systems for which organodynamics seeks to provide a modeling platform are not, themselves, deterministic in nature. Rather, they are characterized by *chance variation*.

Nevertheless, such a nondeterministic dynamical systems theory still need to provide some kind of "dynamics" mechanism that enables a system modeler to "say something about" the time evolution of non-deterministic system that it is modeling. In other words, the dynamics of a nondeterministic dynamical systems theory should be able to provide a *set of constraints* for each time step that results in limiting the possible outcomes for each time step to some subset of all logical possibilities.

There are many mathematical mechanisms that could be employed by a nondeterministic dynamical systems theory to so constrain the possible answers. For example, one way is to provide a set of inequality relations, whose truth set represents a subset of the entire domains space of the inequalities specified.

A special class of *nondeterministic* dynamical systems theories is that of *stochastic dynamical systems theorie*s. This class uses the concept of a *probability space*, often abbreviated to a *probability distribution*, as its technique of imposing *constraints* on the set of outcomes that are allowed for each time step in a dynamical system process.

Using a discrete probability distribution to impose constraints in the allowed outcomes works in two different ways. First, if a certain outcome is to be eliminated as a possible realized outcome for a particular time step, then it will be assigned a probability of zero.

However, using a probability distribution has an advantage over some other mechanisms because the probability assignments can give different weights to the different possible outcomes. In probability theory, these weights carry the semantics of “likelihood”.

The point being made here is that probability theory can be used to impose a kind of constraint on the possible outcomes of dynamical process. This operates as a special case of *nondeterministic dynamics* called *stochastic dynamics*.

Organodynamics is this type of systems theory. That is, organodynamics is a *stochastic dynamical systems theory*.

To put more formality to these ideas, we shall define some terminology to describe some types of dynamical systems theories here: *deterministic*, *nondeterministic* and *stochastic*. Because of the particular types of dynamical systems that we intend to model, we shall narrow our interests to *discrete* processes whose domain spaces are *finite*. This assumption will simplify how we define the three types of systems theories.

> *1. Deterministic* discrete dynamical theories provide precise predictions of the outcomes of future time steps of the dynamical systems that they model. They do this by predicting that the outcomes of each time step will precise specified elements of the state space.
>
> *2. Nondeterministic* discrete dynamical theories provide imprecise predictions of the outcomes of future time steps of the dynamical systems that they model. They do this by predicting that the outcomes of each time step will members of some specified subset of the state space. This is typically accomplished through the express of constraints on the state space.
>
> *3. Stochastic* discrete dynamical theories are a special case of nondeterministic stochastic systems. They operate by imposing constraints on the state space. This is accomplished by specifying a probability distribution for each time step. This approach treats the state space also as a sample space of the distribution. And the probabilities can eliminate some of the states by assigning a probability of zero to it. In addition, the probability assignments are capable of giving more likelihood weight to some of the state than to others.

The Mathematical Foundation of Organodynamics

Of course, organodynamics falls into the definition that we just gave for a *stochastic discrete dynamical systems* theory. This article presents the mathematical mechanisms that organodynamics uses to provide its *dynamics*.

As we shall shortly see, the principle mathematics that underlies the dynamics of organodynamics is that of *entropic functionals*. It turns out, as we shall show, that *entropic functionals* are uniquely suited to the dynamics of organodynamics.

Entropic functionals are a set of mathematical constructs that are developed and studied in a branch of probability theory named *information theory*. However, as we shall see, these constructs are not studied, or mentioned, in mathematical statistics, another offshoot of probability theory.

For that reason, we shall see that *information theory* – defined as the study of entropic functionals – provides the principle mathematical foundation for the dynamics of organodynamics.

We shall develop these ideas more fully in this article. First, though, we shall look at some of the traditional approaches to stochastic mathematics, and ascertain their limitations that prevent them for forming the foundation of organodynamics.

Perspectives

The literature reveals many differing perspectives on the theory of probability, mathematical statistics, the theory of stochastic processes and on information theory. We believe that these various and widely adopted perspectives are all reasonable and “valid” in their own rights – even when they disagree. Each of them derives from differing histories and from reasonable ways to apply these disciplines to a myriad of very different applications.

Organodynamics does not take issue with any of these viewpoints. However, like the other theories and applications, in order to comprise a consistent and comprehensive foundation, organodynamics must take its own stand and select it own perspectives on these issues.

Therefore, what follows represents particular choices of perspective that organodynamics takes on the disciplines of statistics, stochastic processes and information theory. While not rejecting other perspectives, the treatments presented here arise from the applications at hand and will remain consistent with the particular perspective on these disciplines herein adopted by organodynamics.

Traditional Stochastic Treatments

In this section, we take a look at the probability models that are offered by the dominant literature to see if we can use any of it as the probabilistic foundation of organodynamics.

Probability Spaces Recap

In Part II of this series of articles, we formalized the ideas of chance variation and probability theory by defining the formal structure called a *discrete probability space*. This idea placed all of the necessary mathematical equipment to support these ideas into a single construct. We defined *discrete probability space* [Ash and Doleans-Dade 2000] as consisting of three required parts:

1. **Ω**: A set of elements, x_i, treated as sample points
2. **F**: A sigma-algebra on those sample points, elements of which are called “events”. For discrete spaces, we designate that F = 2^{Ω}, the set of all subsets of Ω.

3. ρ: Probability assignments on Ω.

(Note: a *continuous probability space* differs from the above discrete probability space in that the probability assignments ρ are applied to the sigma-algebra F rather than to the set of elements Ω.)

Two noteworthy, simple and textbook examples of probabilistic applications that are supported by this formalization are 1) a flipping a single coin and 2) a throwing a pair of dice.

In the coin flipping experiment, Ω is the set { H, T} where H represents "heads" and T represents "tails". **F** is the set of all subsets of Ω: { {H,T}, {H}, {T}, and Φ}. And if we want to model a fair coin, then $\rho(H) = ½$, and $\rho(T) = ½$. Notice that in this experiment, the elements of the sample space, H and T, do not suggest any particular order. Nor they suggest any particular numeric value.

Note that there is nothing intrinsically numeric about either outcome, and they are not inherently ordered. Of course, an ordering can be imposed on them, and numbers can be assigned. However, the coin flipping application possesses no semantics that suggests of requires that any traits of numbers. There is no notion of amount that is carried by either "heads" or "tails"; and the two do not even possess any natural order semantics. To force some arbitrary assignment of numbers to "heads" and to "tails" would be a semantically useless activity. Not only that, but the results could be misleading.

As for the dice throwing experiment, lets assume that the faces of both die are decorated with symbols that represent in integers from 1 to 6. Admittedly, some dice have pictures for faces – pictures that do not suggest numbers. Such dice would be represented by an equivalent probability space. A reasonable probability space model of this experiment would say that Ω is the set consisting of 36 possible pairs – given that we distinguish the two die. Example sample points in this space would be (2,5) and (4,4). Again, **F** is the set of all possible subsets of Ω. Thus, it would contain 2^{36} such subsets. As for the probabilities, ρ, we want to model two fair dice. Thus the probabilities assigned to the sample points are all the same (1/36).

While we are at it, we may as well recap the organodynamics probability model. An organodynamic model begins with an underlying set of elements S. You may expect that S will be the Ω, the sample space, of the probability space for organodynamics. But it is not. Rather, we first use create another space whose elements are complex entities, each of which represents a possible way that S can be organized – according to a rather complex organizational construct name an *organization of S*. We symbolized this construct as O_{TRS}. Recall that each of these O_{TRS} constructs is mathematically rich, containing a topological space, each of whose open sets has a correspond relations defined. We also symbolized the entire set of all of these O_{TRS} topological constructs. And we symbolized it as $\mathbf{O_S}$. $\mathbf{O_S}$, then, is the set of all possible *organizations* of S. Typically; this is a very large, though finite, set.

Thus, our sample space Ω in organodynamics is $\mathbf{O_S}$. One who is modeling a complex system using organodynamics is left with the task of assigning probabilities to all of these O_{TRS} constructs in $\mathbf{O_S}$ – obviously a formidable task; not unlike the one faced by statistical mechanics. We shall leave the discussion of how this assignment can be made until the next article in this series. For now, though, let us emphasize that it is the individual instances of the form O_{TRS} that constitute the sample points of the application being modeled, and it is these that must be assigned probabilities. We still must comment about the sigma-algebra **F** of an organodynamics application. As before, this will be the set of all subsets of our Ω - which, in this case, is the set of all subsets of $\mathbf{O_S}$. Obviously, the **F** of a typical organodynamics application is gargantuan, but finite.

Thus, all three of these experiments start out as a finite *probability space*. Moreover, the probability space defines a complete set of mathematical machinery necessary for "doing probability theory" on the application mentioned.

What's Missing

Usually when we engage in a probability model of an application, one of the first things we think about doing is "taking the average". In fact, we at least expect probability theory to provide the necessary mathematical equipment to be able to take averages using statistics such as *expected value* and the *mean*. These are both *functionals* that map an entire probability distribution (which is an abbreviation for a probability space) to a single real number. The semantics of these particular functionals is to specify the *centrality* of the probability distribution that it is characterizing.

In fact, it is reasonable to suggest that mathematical statistics is a discipline that has a central purpose of characterizing a probability space (distribution) by providing a set of functionals that measure it.

In addition to the mean, we also hear quite a bit about *variance* and *standard* deviation. Together with the mean, these functionals collectively are used to characterize certain probability distributions.

The question that immediately arises, then, is "Are these functionals (mean and variance) defined for the three example probability spaces that we just looked at?"

The answer is No! There is not enough "equipment" in any of these three probability spaces even to define a *mean*. Here's why. The formula for the mean is:

$$\mu(X) = \sum p(x_i)*x_i \text{ for all } x_i \text{ in } \Omega.$$

However, for all three of these example probability spaces, "$p(x_i)*x_i$" is undefined, since "*" is multiplication and "x_i" is not a number in any of the above three example applications. In the coin toss example, x_i is either H or T. In the dice example, x_i is a pair of die faces. In the organodynamics example, x_i is an embellished topological space of the form O_{TRS}.

Therefore, the notion of *mean*, or *expected value*, or even *average*, does not apply to, is semantically undefined for, these three probability spaces. It is undefined. The principle reason for this is that none these three example probability spaces are algebraic *fields*; which means that addition and multiplication are not defined for their members. For the same reason, variance and standard deviation cannot be defined on their sample spaces either. Specifically, their sample spaces are also not metric spaces.

Of course, we could try to force the situation by 1) mapping each sample point to a real number, and then 2) using the resulting real numbers as a revised sample space. However, for these three applications, doing so would not be semantically meaningful. Thus, consequently, the forced notion of a mean in these cases would be semantically meaningless. So there is no reason, or justification for making such a mapping to the reals.

Of course, if the *mean* is a meaningless concept for these three applications, then so is *variance*, and so is *standard deviation.* In fact, the statistical concepts of *moments* and *central moments* are also meaningless for these three probability spaces.

These three examples are complete probability spaces – even without some notion of mean, or *variance* or *standard deviation* defined for them. They are perfectly proper probability applications, even without a *mean*.

So, it is perfectly viable for some probability spaces to not have *means* or *variances*. There is no need to redefine *probability space* to force all of them to have these particular functionals.

Random Variables

Probability applications are frequently referred to as *random variables*. This term is often thrown around somewhat loosely in the literature and in textbooks. And this fact causes a difficulty for the theory of organodynamics that I shall address in this section.

There is some ambiguity in the usage of the phrase *random variable* that is responsible for much of this difficulty. So I must first sort this out. In addition, an attitude has developed that assumes that all "interesting" probability spaces qualify as *random variables*, and that every topic in probability theory should be defined a some kind of random variable.

However, I will show that there are some very interesting and highly complex probability spaces that, in fact, do not qualify as random variables – as the term is formally defined in probability theory. It happens that organodynamic probability spaces are one class of these.

First, I will address two distinct usages of the term *random variable* in probability theory and its applications.

Random Variable: the Loose Meaning

Deterministic mathematics uses variables in expressions and "open sentences" (equations, inequalities, etc.) to operate as placeholders for members of a specified set, called the domain set. This idea of using "variables" in deterministic mathematics is, or course, very useful and ubiquitous. So, why not apply a similar idea to probability spaces?

In the probability space case, such a variable would be called a *random* variable to distinguish it from the *deterministic* variables we just discussed. ([Lemons 2002] prefers the phrase *sure variable* instead of *deterministic variable*. So, I shall use that here.) Of course, we would have to determine "What would be the "domain set" for these *random variables*? An obvious answer is the *sample space* of the probability space.

Of course there are some differences between using *sure variables* versus *random variables*. With sure variables, the focus is on which members of the domain make the open sentence in question a true statement. This may be a one or more member of the domain. For *random variables*, on the other hand, the interest is in which (exactly one) of the sample space will be realized by a trial of the probability space in question.

Nevertheless, the concept of *variable* used in deterministic mathematics does seem applicable to the world of probability spaces, as described. This is the “loose meaning” of *random variable*.

One further comment concerning this loose meaning of random variable: It seems to mean exactly what it says: a variable that applies to randomness. Certainly, this is an appealing trait.

One final comment concerning this loose meaning: It applies equally well to all probability spaces.

Random Variable: the Strict Meaning

The phrase *random variable* has undergone an evolution in which its meaning has been narrowed, specialized. This specialization involves the requirement that every sample point in the probability space be associated with a real number. Once this has been done, then the probability space itself, as well as the variable symbol used to represent it, qualifies as a random variable.

In his book *Discrete Stochastic Processes* [Gallager 2011], R. G. Gallager describes this strict concept of *random variable* as follows:

> The outcome of a probabilistic experiment often specifies a collection of numerical values such as temperatures, voltages, numbers of arrivals or departures in various time intervals, etc. Each such numerical value varies, depending on the particular outcome of the experiment, and thus can be viewed as a mapping from the set **Ω** of sample points to the set **R** of real numbers…. These mappings from sample points to real numbers are called random variables.

Gallager also provides a definition that formally specifies these ideas, which I omit here, at [Gallager 2011, p. 11].

For example, suppose we define a probability space in such a way that each “trial” involves throwing six coins into the air. Such a sample space has *configurations of six coins* as its sample points. In fact, the sample space has 2^6 = 64 such member configurations. Each of these 64 configurations has some probability, and we can create a sigma-algebra of events. Typically, the probabilities of the sample points are equal.

However, this <u>probability space is *not* a *random variable*</u> according to the strict definition of the term. The reason is that its sample space has not been mapped to real numbers.

However, suppose we define a mapping that takes each or these 64 samples points and maps it to some meaningful real number. This mapping itself is formally called a *random variable.* As well, the probability space is also called a random variable as well as the symbol used to represent it. In this strict meaning of “random variable”, all three of these uses are tolerated. However, the formal definition has the mapping itself as being the “random variable”.

For example, this mapping could associate each configuration to its number of heads. Clearly the codomain of this consists of the set of numbers from 0 to 6.

Moreover, when a probability space involves such a random variable mapping, the codomain becomes the “new sample space” while inheriting the probabilities. Generally, however, the random variable mapping is many-to-one. For example, there are six distinct configurations that map into the same codomain element of “1” – all of the configurations with exactly one head. While at the same time, there is only one configuration with zero heads, and one configuration with 6 heads. Consequently, while the probability distribution associated with the “configuration” sample space was uniform, the new probability distribution associated with the inherited codomain of real numbers is not uniform.

[Ross 1996] formalizes what I exemplified in the previous paragraph when he says that

> Consider a random experiment having sample space S. A random variable X is a function that assigns a real value to each outcome in S. For any set of real numbers *A*, the probability that X will assume the value that is contained in the set *A* is equal to the probability that the outcome of the experiment is contained in $X^{-1}(A)$.

Advantages of the Strict Meaning

One advantage of a probability space conforming to the strict meaning of *random variable* is that the concepts of a *mean*, or *expected value*, can be defined. In fact, if the probability space is not (strictly) a random variable, then the mean is not definable!

This can be clearly seen by inspecting the definition of mean:

$$\mu(X) = \sum p(x_i)^* x_i \text{ for all } x_i \text{ in } \Omega.$$

Notice that the expression on the right sums the product of two numbers: $p(x_i)$ and x_i. Both of these must be numbers (or at least members of a field) in order to be multiplied. If the probability space is not a *random variable* in the strict sense, the this product does not exist.

Of course, if means don’t exist for a probability space, then neither do variances, standard deviations, skewness, kurtosis nor any of the moments and central moments defined by mathematical statistics. But, of course, these moments and central moments form the central toolkit of mathematical statistics.

And, of course, mathematical statistics is the principle discipline of probability theory that is use to characterize chance variation in probability spaces. Most practitioners are unaware of any other discipline in sight that can play this role of characterizing chance variation in probability spaces.

Dominance of the Strict Meaning

Thus, it is no wonder that mathematicians, without ever really saying so, pretty much require a probability space to also be a random variable. This is exemplified by the fact the definitions and developments of more advanced fields of probability require that the probability spaces that they deal with are (strict) random variables. For example, look at this definition of *stochastic processes* by the widely acclaimed textbook on the subject by Sheldon Ross:

> A *stochastic process* $X = \{X(t), t \in T\}$ is a collection of random variables. [Ross 1996 p 41].

So, the strict meaning of *random variable* that requires a probability space to map its sample points to real numbers and then to attribute probabilities to those real numbers has come to dominate the usage of the term in probability and applications literature.

Such a definition perfectly well accommodates an entire range of stochastic processes that make use of probability spaces that *are* random variables. Such processes include Brownian motion, Poisson processes, birth-death processes and countless other stochastic processes used in applied mathematics and mathematical physics. However, it fails to include others, such as certain Markov processes, random walk processes, and others.

Unfortunately, the implication is made that a probability space must also be a random variable (have real number assignments) in order be deemed worthy of treatment by any serious theory.

The Impact of the Dominance of the Strict Meaning on Organodynamics

The probability space of organodynamics defines a very rich sample space $\mathbf{O_S}$ whose topological sample points O_{TRS} are each very complex. In fact, each of these topology sample points are some rich that there is no meaningful way to associate each one with a real number. Moreover, to do so would represent a loss of all of the complex information that is inherent in each of these topologies.

Thus, this probability space is NOT a random variable in the strict sense. Yet it remains exceedingly large and complex. In addition, the theory that we developed on top of this probability space leverages more advanced probability constructs such as stochastic processes. This is evident in our definitions of the OSP and the ODSP.

But, if we followed the definition of [Ross 1996, p 41] of *stochastic process*, then these definitions could not be possible.

The Resolution

Fortunately, all that is necessary to construct a discrete stochastic process is to define it as a sequence of probability spaces over time – as we did in Part II. These probability spaces need not be (strictly) random variables – as we have seen by our development of the OSP and the ODSP in Part III.

However, we must address the problem that we will not be able to leverage the defining framework used by mathematical statistics for characterizing chance variation in a probability space – the use of *moments* and *central moments*. It can be argued that the moments and central moments framework is the preeminent mechanism for characterizing chance variation in mathematical statistics. If we can't use this framework in organodynamics because organodynamics has probability spaces that are not random variables, then what do we use?

The answer is information theory – defined as the study of *entropic functionals*. As we shall see later, the entropic functionals defined by information theory provide a comprehensive alternative probabilistic framework to mathematical statistics for characterizing any probability space – whether or not it qualifies as a random variable.

While both frameworks have their advantages and disadvantages, we shall find that the entropic functionals of information theory provide the power that we need to characterized chance variation in organodynamics – as well as to describe various types of time evolutionary behavior in OSPs and especially in ODSPs.

Additionally, to break the ambiguous "loose" and "strict" uses of the phrase "random variable", I shall use a new term in these articles for the "loose" meaning. In those cases, I shall substitute the phrase *chance variable* in place of the phrase *random variable* whenever the loose meaning is needed.

For example, I shall say that organodynamics uses *chance variables*, but not *random variables*.

Information Theory and Predictability

Previously in this article we have surveyed the body of work in probability theory and its offshoots looking for a mathematical foundation for organodynamics. What we found was that the predominant work in the field focuses on stochastic systems that can be described as probability spaces that satisfy the condition of *random variables* – probability spaces whose sample points are, or can be represented by, real numbers.

Unfortunately, as we showed in the previous section, the probability spaces of organodynamics cannot generally be assumed to be random variables. While these organodynamic probability spaces are indeed very rich and interesting, they nevertheless fail in general to satisfy the conditions of being random variables.

This fact means that none of the traditional disciplines of probability – with their powerful arsenal of moments, central moments, and moment generating functions - can provide a satisfactory mathematical foundation for organodynamics.

Fortunately, however there is a viable alternative – another offshoot of probability theory that also advances a robust and arsenal of mathematical structures in its own right. In fact, this arsenal provides the essentially equivalent powers of moments and moment-generating functions for the purpose of characterizing probability spaces and their distributions. Like moments and moment generating functions, this toolkit is designed to qualify the stochastic constraints provided by stochastic dynamics.

This arsenal is called *entropic functionals*, and the sub-discipline of probability theory whose focus is to develop these functionals is named *information theory*.

Information Theory

There is some confusion in the literature as to what information theory is, what it studies and what is its scope and boundaries, and just what general area of investigation it belongs to.

For the purposes of developing organodynamics, it has been most prudent for me to establish a position on these questions so that I can proceed in consistent manner. In the first part of this section, I shall briefly summarize how I shall define information theory for the purposes of these articles. After that, I shall recount and interpret some history around the development of information theory, and use that interpretation as an argument to justify the position I have arrived at as to how to define the subject.

Admittedly, my position is biased toward my needs to shape the definition of information theory so that it provides the kind of mathematical foundation for organodynamics that I desire. On the other hand, I believe that my characterization of the field is not at significant variance from some other investigators, and is a reasonable one in any event.

Defining Information Theory

At an intuitive level, I understand *information theory* to be *the study of uncertainty* as represented via a mathematical concept known as *statistical entropy* – which we shall henceforth refer to simply as *entropy*.

I initially arrived at this understanding by studying the mathematical definition of entropy. I was further encouraged to adopt this understanding by reading the works of [Shannon 1948], [Tolman 1938], [Khinchin 1957]], [Jaynes 1957] and [Kleeman 2012].

My sympathies are with Kleeman when he says in the first lecture of his graduate seminar on stochastic processes and predictability:

> The central idea of information theory is to measure the uncertainty associated with random variables. [Kleeman 2012, Lecture 1, page 1, paragraph 1, first sentence.]

Kleeman's use of the term *random variable*, here, is in the "loose" sense of the term as I defined it in the previous section. In other words, he does not require that a value function (mapping to the reals) be defined. In the terminology that I introduced in the previous section, he could have substituted my term *chance variable*.

Information theory defines *entropy* to be a function that maps a probability space to a non-negative real number, and that provides a measure of the *degree of uncertainty inherent in that probability space*. Since a probability distribution is an abbreviation for a probability space, then *entropy* can be considered as a measure of the degree of uncertainty inherent in a probability distribution.

This makes entropy a *functional* from a set of probability spaces into the non-negative real numbers. Later in this section we shall attempt to show the intuition behind this claim.

The "invention" of information theory is generally credited to Claude Shannon, who introduced it in his 1948 paper entitled "The Mathematical Theory of Communication" [Shannon 1948].

In that paper, Shannon suggests that entropy measures the amount of information that is produced. At the same time, however he suggests that entropy is also a measure of how much uncertainty is present. This point is amplified by [Khinchin 1957, p. 1]. Thus, it seems that entropy is simultaneously a measure of the amount of information present and the amount of uncertainty present. Thus Shannon equates information and uncertainty.

This seems counter-intuitive to many. However, notice that if one already knows something (i.e. there is zero amount of uncertainty about it), then the value of obtaining that information again is also zero. On the other hand, if one has zero information regarding some phenomenon (maximum uncertainty and therefore maximum entropy), then the value of obtaining that information is at its maximum. In this sense, entropy is simultaneously a measure of uncertainty and degree of information.

Brief Interpretation of the History of Information Theory

In the field of thermodynamics, developed in the middle of the 19th century, *entropy* is a measure of the amount of energy that is unavailable for work in a closed system. It is defined as the integral of the inverse of the temperature with respect to the change in heat. Note that this definition is deterministic. That is, for specific inputs (temperature and change in heat), the amount of entropy is always the same.

However, in the later half of the 19th century, the atomic theory of matter began to gain considerable adoption. This theory implied that matter was not really solid, but consisted of vast numbers of particles moving around randomly. But if the constitution of a solid, gas or liquid is actually "moving around randomly", then how can the deterministic science of thermodynamics accurately account for entropy and other thermodynamic quantities.

To resolve these questions, the science of statistical mechanics was developed. This physics attempted to provide statistical analogs for all of the major thermodynamic quantities – including entropy. Now, these analogs could not be precisely equivalent to their thermodynamics counterparts. This is especially true because one was deterministic (always produce the same answer for the same initial conditions), while the other was subject to random fluctuations (chance variation). Nevertheless, the two should produce the same results over time, in some sense.

In fact, we can exemplify how different the two systems are by using *entropy* as an example. Lets call the entropy of thermodynamics "thermodynamic entropy"; and the entropy of statistical mechanics "statistical entropy". Thermodynamics entropy is defined by a formula, as we showed above, that has only temperature and change of heat as its inputs. Moreover, it is a deterministic calculation in the since that the same inputs always produce the same output. On the other hand, the calculation for statistical entropy, due to Gibbs, takes only probabilities as inputs. It is nondeterministic, or stochastic, by nature. The two are analogous, but different – not equivalent mathematically or physically.

In 1948, Claude Shannon, a mathematician who worked for Bell Telephone Laboratories, was tasked with developing a mathematical theory of communications. He decided that he needed a mathematical foundation for this theory. He had noticed that the sending and receiving of messages was subject to chance variation and uncertainty; and he wanted to represent this aspect with probabilities. So he needed a mathematical theory based on probabilities that also defined some metric for the degree of uncertainty of a probability space (distribution).

He recalled his studies of statistical mechanics, and realized that the concept of statistical entropy defined there was exactly what he needed – even though statistical mechanic was invented for the purpose of "doing physics", while Shannon's intention was to "do communications theory".

What Shannon realized was that what mattered was that Gibbs' definition of entropy only required probabilities as inputs. This is what the statistical mechanics situation had in common with the communications theory situation. This meant that Gibbs' entropy was amenable to any application that has probabilities! It didn't have to be physics. It could be any subject as long as it had a probability distribution. In fact, Shannon realized that Gibbs definition of entropy is a measure of the degree of uncertainty inherent in a probability distribution.

Of course, Shannon was trying to develop a mathematical theory of communications. But he had already realized that the field of communications "has probabilities". Therefore, Gibbs' formula for entropy would work for it – as well as for statistical mechanics.

In other words, Shannon realized that the mathematics that Gibbs developed for doing statistical mechanics was good for many applications other than statistical mechanics. In fact, it was also good for providing a mathematical foundation for Shannon's nascent "theory of communications". Shannon realized it was bigger than that: this mathematics would be good for any situation that "has probabilities".

Resolving the Confusion Around Information Theory

Shannon presents these ideas in section 6 of his 1948 paper [Shannon 1948, pp 10-12], where he attributes them to Gibbs. He also references [Tolman 1938]. As well, he refers to these ideas using the term "information theory". After that, beginning in section 7, he applies these "information theory" ideas of entropy to communications theory – which was his initial intention.

My interpretation of this is that Shannon leveraged the idea of entropy that he got from Gibbs by reading [Tolman 1938]. And that he considers these ideas to belong to "information theory". I further interpret that Shannon applied these ideas as a mathematical foundation for developing his "mathematical theory of communications" throughout the remainder if his 1948 paper "A Mathematical Theory of Communication".

In other words, I prefer to distinguish "information theory" from "communications theory". I regard communications theory as an application of information theory. Further, I regard information theory as a branch of mathematics – specifically of probability theory. On the other hand, I regard communications theory as a branch of electrical engineering and computer science.

In fact, I would argue that failing to distinguish information theory from communications theory is like confusing calculus with celestial mechanics. Just because Newton invented the calculus in order to be able to "do celestial mechanics" does not mean that the calculus "is celestial mechanics", and does not apply to other subjects.

It is important to understand that information theory stands alone as a branch of probability theory, and is applicable to any applications that have probabilities. It is not limited to communications theory or to computer science. And it is not be confused with "information technology".

Unfortunately, the phrase "information theory" is very often linked to communications theory. For example, John R. [Pierce 1980], in his widely referenced book entitled "information theory", interchanges the phrases "information theory" and "communications theory" throughout. It is telling that his subtitle is "Symbols, Signals and Noise". He has obviously written a book about communications theory, but has named it *Information Theory*.

We already have one name – "communication theory" – for the study of messages, senders, receivers, noise, etc. We do not need a second term for it ("information theory). And we also need a term to name the "study of entropy and related functionals" as an offshoot of probability theory. My position is that the term "information theory" should be reserved for that endeavor.

For this series of articles on organodynamics, I shall use *information theory* to mean:

> Information theory: a branch of probability theory that is concerned with characterizing the chance variation inherent in a probability space (i.e. *distribution*) through the development and application of *statistical entropy* and a collection of related probabilistic constructs named *entropic functionals*.

The Information Theory Repertoire

We have been very interested in predictability of dynamical systems in these articles. We have pointed out that notion of “dynamics” implies some kind of mechanism that describes the chance evolution of a system over time. Classical dynamical systems are completely predictable (have entropy 0); which is a trait that makes them very useful to many practical applications.

We have been making the case that nondeterministic dynamical systems theories can be useful also whenever they can *narrow* the prediction of future outcomes, even if they cannot precisely determine them. Our argument has been that in highly complex systems, we may not be able to narrow the precision or our predictions completely, but that we can use nondeterministic dynamical systems theories to narrow, or constrain, them significantly.

In organodynamics, we shall use entropy and entropic functionals to provide this narrowing of our predictions. Information theory provides these mathematical constructs that provides a characterization of probability spaces by using only their probabilities.

Essentially, the probabilities of a space alone are sufficient to collectively characterize to degree that *chance variation* is at work within the space. And, the unique approach of information theory is to attempt such a characterization by working only with the probabilities of the space. Hence, we shall find that the only input values of functions and formulas defined by information theory are, in fact, the probabilities of the space. These functions and formulas generally require the probabilities associated with all of the sample points in the space being characterized.

The repertoire of mathematical constructs that constitute information theory is a set of functionals that map a probability space into the non-negative real numbers. The basal functional in this set is named *entropy*, sometimes called *statistical entropy*, or *stochastic entropy*.

Probability theory defines a number of *derived distributions* that develop from an initial one, and that form the body of theory of probability. Examples of these are *conditional distributions*, *joint distributions*, stochastic processes and dependent stochastic processes.

In order to characterize these extensions to an initial distribution, information theory defines a number of other *entropic functionals* beyond basic entropy. Some of these are: *joint* entropy, conditional *entropy*, *relative entropy*, *mutual information*, and *entropy rate*.

These *entropic functionals* form a repertoire that constitutes the body of mathematics that is information theory. We shall next introduce and define these entropic functionals, discuss their application, and show an example of how the might be applied by referring to the fRMI model of human cognition exemplar system that we presented in the previous article.

Entropy

Entropy is a measure of the degree of uncertainty in a probability space (or distributions). It is often symbolized by the function ”H”, for historical purposes. Entropy can be understood as a functional that maps a probability distribution X into the non-negative real numbers.

Entropy is defined for all finite probability spaces, and also for many continuous ones. However, organodynamics confines (for now) its attention to probability spaces with finite sample spaces.

> Entropy of probability distribution S: Let X be a probability space (S, F, P) with finite or countable sample space S. Then the entropy of X, H(X) is
>
> $$H(X) = -\sum_{i=1}^{n} p(x_i) log(p(x_i))$$

Note that the inputs to H(X) are probabilities, and only probabilities. Furthermore, all probability assignments in X are included in H(X).

Notice that the value of H(X) will be non-negative. Since the log of any probability is non-positive, then the summation will be non-positive. And the negation operator renders H(X) as non-negative.

Note further that if X also happens to define a value function on its sample points, then the values of that function are ignored. That is, only its probabilities are used – not any real values associated with the distribution. In other words, if the distribution happens also to be a random variables (defines a real valued function on S), then such values are ignored by H(X).

This, of course, means that X must be a probability distribution (space), but it need not be a random variable. It can be merely a chance variable, as I have defined the term. X is allowed to be a random variable, but is not required to be.

Therefore, entropy can be understood as a function that "measures" a probability distribution X to determine the degree of chance variation, or uncertainty, that is inherent in it.

We have mentioned that [Shannon 1948] also characterized entropy as a measure of the "amount of choice" as well as the "amount of information" in a probability space – as well its degree of uncertainty. Of course, all such descriptions are merely *interpretations* the semantics of entropy. It true meaning is found strictly in the mathematics of its definition.

On the other hand, reasonable interpretations of the meaning of *entropy* are appropriate on an application-by-application basis. And, either "high entropy" or "low entropy" or "intermediate degrees of entropy" may be deemed "better" or "worse" on an application basis. Some other interpretations of the meaning of entropy that have been observed in real applications include: "instability/stability", "degree of opportunity", "amount of freedom", "unpredictability/predictability", and many others.

It can be shown that entropy for a given probability space has a range that is bounded below at zero. For a finite space, entropy is bounded above at log(N), where N is the cardinality of the sample space.

The entropy of a probability space (or probability distribution) can be understood as a measure of "how random" the distribution is. If entropy is zero, then the space is essentially deterministic, totally predictable or deterministic. In other words, "stochastic" or "probabilistic" includes "deterministic" as a special case. For such a distribution, the probability of exactly one sample point is 1.0; while the probability of all other sample points is zero. The probability distribution is at the maximum for the uniform distribution on a sample space. This fits with intuition, since all probabilities are equally likely for this distribution, and therefore give the least information regarding which sample point to predict.

But, every value between 0 and log(N) is an entropy value for some distribution of probabilities over any sample space (the codomain of entropy is an continuum). This means that the "amount of randomness" that is possible for a distribution ranges across the continuum from 0 to log(N). Thus, the concept of "partial randomness" is fully realized in information theory. Yes, you can be "a little bit random".

In this way, entropy is the first of these entropic functionals to assist in narrowing the prediction of the outcome of trail in an experiment for a probability space. The role of entropy in this regard is to give a measure of how predictable the probability space overall is.

The lower the entropy the more predictable is the distribution. The predicted sample point, then is the sample point whose uncertainty value, according to the measure u(x) = -log(p(x)). Of course, this is simply the sample point with the highest probability. It may or may not be unique.

Intuitive Approach to Entropy

Students of information theory often spend some amount of time grasping for an intuitive derivation for the definition of entropy just presented. Of course, many sources let the above definition stand.

However, there is a more intuitive approach, which we shall present here.

Lets first note that, since the time of Aristotle, people have been trying to provide a mathematization of the concept of "uncertainty" in terms of probabilities. It has often been noted that – intuitively – that "likelihood" and "uncertainty" seem inversely proportional, according to common usage. That is, the less likely something is, the more uncertain it is – and vice versa.

Of course, we already have a way of measuring likelihood – with probabilities. So, let's create an intuitive mathematical definition based on this for a measure of *uncertainty*. Lets use the symbolism "u(x)" to identify this measure of the uncertainty of an event x. That is, if x is an event, and p(x) is the probability of that event, and uncertainty and probability are inversely related, the our initial formula for the uncertainty of x, u(x) would be:

$$u(x) = 1/p(x).$$

This first approximation is fine, but there are obviously many more possible formulas that also represent and inverse relationship between uncertainty and likelihood. Before we settle on this one, lets see if there is some other principle involved that would suggest a different choice of a definition of u(x).

It happens that there is: It would be very handy if our definition of u(x) insured that statistically independent events are additive. For example, if I construct an experiment that involves flipping a coin followed by rolling a game die, then it seems intuitive that the "degree of uncertainty" inherent in that combined experiment would be the same the sum of the die experiment and the coin experiment by themselves. That is, a property that we would desire for our measuring function u(x) is:

$$u(x \text{ and } y) = u(x) + u(y).$$

However, this relationship is not supported by our initial attempt at defining u(x) = 1/p(x).

However, it turns out that the following redefinition of u(x) is additive for statistically independent events:

$$u(x) = \log(1/p(x)) = -\log(x).$$

Notice that this definition of uncertainty also proscribes an inverse relationship between u(x) and p(x) – a relationship between uncertainty and probability that we desire. Moreover, it is additive for statistically independent events, since:

$$u(x\^y) = \log(1/p(x\^y)) = -\log(p(x\^y)) = -\log(p(x)*p(y)) = -\log(p(x))-\log(p(y)) = u(x)+u(y)$$

Since this revised u(x) exhibits both relationships ('inverseness" and additivity) that we want, we shall adopt it as our definition of "the uncertainty of a sample point".

Moreover, as Shannon demonstrates in his appendix 2, [Shannon 1948, p. 28], the above definition of u(x) is unique in being additive. Thus, it is the correct choice for a measure of the uncertainty of a probability space.

Since most sources that define entropy begin with the unmotivated definition that we presented in the previous section, there is little attempt to provide a name for this function that we have developed to measure the uncertainty of a single sample point, u(x). However, Shannon does mention the concept of the "entropy of a single sample point" at the top of page [Shannon 1948, p. 13], after he has define *entropy* – a measure of an entire sample space.

One sometimes encounters an attempt to apply the term "surprisal" to this measure of the uncertainty of a single sample point – that we a calling "u(x)". But the term "surprisal" has not been widely adopted.

Nevertheless, if u(x) can be understood as the "degree of uncertainty inherent in a single sample point", then it follows that the expected value of the u(x)'s of all of the sample point would be a reasonable measure of the degree of uncertainty inherent in the probability distribution as a whole!

In fact, the definition of entropy that we presented in the previous section can be derived by calculating the expected value of the u(x)'s over all sample points in the probability space. Such an expected value would be:

$$E(u(x)) = \Sigma_{x\in S}\, p(x)*u(x) = \Sigma_{x\in S}\, p(x)*\log(1/p(x)) = -\Sigma_{x\in S}\, p(x)*\log(p(x)) = H(X)$$

Joint Entropy

So far, we have discussed the role of probability spaces to model chance variation; and we have further entertained the idea that the entropic functional named *entropy* can provide a useful measure of just how random and unpredictable, or conversely how deterministic and predictable, such a probability space can be. This is an important way to characterize the behavior of chance variation.

But so far we have only discussed one chance variable at a time. It would be very interesting, and useful, to be able to characterize the interrelationships between two distinct chance variables at once to see if either has some kind of influence – chance-related influence – on the other. We may be tempted to call this “causality”. Probably a better appellation is *statistical dependence* or *stochastic dependence*.

All of the next several entropic functionals that we shall discuss make use of statistical dependence (if any) between or among two or more probability spaces, as they interrelate with each other, in order to “narrow the predictability” of the evolution of a dynamical system over time.

The first step in this issue is to discuss the mechanism by which multiple probability spaces (distributions) are combined into a single joint space. The second step is to be able to discern a) whether or not there is any *stochastic dependence* at work among the particular distributions that are being jointed, and b) if so, then how much stochastic dependence? The latter question is of interests, because it indicates just how much predictability has been introduced into the system as a result of this stochastic dependence.

We shall discuss the “a)” part of the above question in the present section. The mechanism used in probability theory to combine two distributions into one is the *joint distribution*. We assume that the reader is familiar with this concept; and shall only add that the sample space of the joint distribution is the set of all ordered pairs of the two initial distributions[1].

What is of interest in such a joint distribution is their probabilities – the probabilities of all of the pairs – which constitutes the joint distribution. Of course, for any two initial distributions, there is only one joint sample space; but there are an infinity of joint distributions, because probabilities can be assigned to their joint sample space in an uncountable number of ways – ways that possible have no bearing on the probabilities of the two initial spaces.

For a given joint distribution, it is these probability assignments that characterize the degree of stochastic dependence or independence of the two chance variables of that distribution. Therefore, it is the degree of stochastic dependence that determines how much the predictability of a stochastic process that involves this joint distribution can be narrowed.

The next four entropic functionals that we shall discuss are all based upon this idea – that the degree of stochastic dependence between two or more chance variables provides the degree of predictability of a dynamical system.

For this section, however, we want to point out that we can also apply *entropy*, as we have so far defined it, to a joint distribution as well as too a “standalone” distribution. That is, from a simple perspective, a joint distribution is simply a distribution that has sample points, and those sample points have probabilities. Thus we can apply the above definition of entropy to it also. This is called *joint entropy*.

So, given that the sample points if a joint distribution are pairs (x, y) of sample points of the two constituent distributions X and Y, then the definition of *joint entropy* for joint distribution (X, Y) is:

$$H(X, Y) = -\Sigma_{x \in (X,Y)}\, p(x,y)*\log(p(x,y))$$

[1] Joint distributions are not limited to joining only two chance variables. In fact, any countable number of chance variables can be combined to form a joint distribution. We shall need these multi-dimensional joint distributions later. But for now we shall work with joint distributions of two dimensions.

Joint probability spaces, distributions (chance variables) and entropy can be generalized to any number spaces. Of course, in these articles, we are restricting our interests to discrete numbers. This means that, our general definition of joint entropy is:

$$H(X_1, X_2,\dots X_n) = -\Sigma_{(x1,\ x2,\dots\ xn)\in(X1,\ X2,\dots\ Xn)}\ p(x_1, x_2,\dots x_n)*\log(p(x_1, x_2,\dots x_n))$$

Conditional Entropy

Conditional entropy is the first of four entropic functionals that makes an accounting for the degree of dependency between or among two or more distributions as they interrelate in a joint distribution. Of course, conditional entropy is meaningful only in the cases that we are working with a joint distribution.

For simplicity, let's treat the case of a joint distribution of exactly two initial distributions. This is immediately generalizable to the case of any finite number of initial distributions.

Conditional entropy is simply the entropy of the *conditional distribution* of a specified joint distribution. The conditional distribution is obtained from the joint distribution p(X, Y) by dividing each entry by its row sum p(X=x). This operation effectively normalizes each row of the joint probability matrix so that it sums to one and becomes a probability distribution in its own right. The row then represents the probabilities of Y given that X=x for the associated row.

This new probability matrix, obtained by dividing all entries of each row by its row sum, is referred to as the *conditional distribution* of (X, Y), written p(Y|X) and named the probability of Y given X.

One can informally assess the degree to which the two chance variables X and Y of joint distribution (X, Y) are *stochastically dependent* by inspecting the conditional distribution p(Y|X). If the rows (or row distributions) are all the same, then X and Y are stochastically independent. Otherwise, X and Y are stochastically dependent. Notice that there is only one way that the rows of p(Y|X) are stochastically independent, but uncountably many ways that they are stochastically independent.

Much applied mathematics and mathematical physics literature leaves one with the impression that only stochastic independence is interesting and that stochastic dependence is of little interest. This is likely because the determination of joint probabilities of p(X, Y) is completely calculable from the initial distributions if and only if the joint chance variables are assumed to be stochastically independent. If independence cannot be assumed, then the joint probabilities must be obtained by some other – usually laborious – manner. Unfortunately, science is ripe with cases of making the unsupportable assumption that statistical independence is warranted, simply for the sake of calculability, when nothing could be further from the case – to disastrous consequences [Hoyle and Wickramasinghe 2001].

However, in information theory, the opposite is the case. In the first place, degree of stochastic dependence is a continuous scale, with only one point on that continuum representing statistical independence. For another, the more mutually stochastically dependent two (or more) variables are, the more predictable (less random, nondeterministic) they are. Stochastic dependence is more or less the basis for predictability in information theory.

Of course, one can also calculate the related conditional matrix p(X|Y) - or the probability of X given Y – from the same joint distribution p(X, Y). This is done by dividing each entry of p(X, Y) by its column sum. The resulting p(X|Y) matrix has columns that are probability distributions. Both of these matrices are important and figure prominently into the equalities of information theory. Further discussion of these topics is beyond the scope of these papers. We refer the reader to [Cover and Thomas 1991; pp. 2-16].

Strictly speaking, a conditional probability matrix is not really a "probability distribution", because its entries do not sum to one. In fact, each of its rows (or columns) sums to one. Nevertheless, it figures prominently into the definitions of several of the entropic functionals discussed here, so it is not inappropriate that it be referred to as one.

We have briefly introduced *conditional distribution* of the two variables X, Y of a joint distribution p(X, Y). But we have yet to define *conditional entropy*. In the same way that, intuitively, *joint entropy* is the *entropy* of the joint distribution p(X, Y), the *conditional entropy* H(Y|X) is simply the *entropy* as applied to each entry of the matrix p(Y|X). Thus, joint entropy – for two variables X and Y of joint distribution p(X, Y) is defined as:

$$H(Y|X) = \sum_{x\in X}\sum_{y\in Y} p(x,y)*\log(1/p(y|x))$$

Or, equivalently,

$$H(Y|X) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)*\log(p(y|x))$$

The values of H(X|Y) and H(Y|X) pertain to the degree of interdependence between X and Y. Intuitively, conditional entropy H(X|Y) can be understood as the amount of uncertainty of X that is not shared with Y. Likewise, conditional entropy H(X|Y) can be understood as the amount of uncertainty of Y that is not shared with X. In fact, the more dependent are X and Y, the smaller H(X|Y) and H(Y|X) become.

Whereas, *entropy* H(X) and joint entropy H(X, Y) decreases whenever uncertainty decreases, H(X|Y) and H(Y|X) decrease whenever stochastic independence decreases. Just like entropy, stochastic independence reduces predictability. So the lower are the measures of both, the higher is the level of predictability.

Relative Entropy

Relative entropy is one way to compare two probability spaces (distributions, chance variables) for "how far apart they are" in what they suggest about the uncertainty inherent in the space. Like all entropic functionals, relative entropy defines a calculation for each sample point in the sample space, and then multiplies that each such term by a probability before summing all the products.

The probability that it uses as a multiplier for each term is taken from one of the two distributions that it is comparing. Thus, relative entropy is always "relative" to the distribution whose probabilities are used as the multiplier.

The symbolism for relative entropy is D(p || q). The "D" is sometimes interpreted to mean "distance"; and p and q are the two probability distributions whose "distance" is being compared. Bu convention, the first-named distribution, p, is the one whose probabilities are being used as the multiplier.

From what we have said, we can write the definition of relative entropy between probability distributions p and q on a specified sample space as

$$D(p \,||\, q) = \sum p(x) \,[\, u(p(x)) - u(q(x)) \,]$$

However, this is equivalent to

$$D(p \,||\, q) = \sum p(x) \log (p(x)/q(x))$$

The latter expression provides for more efficient calculation, and is the typically provided as the definition or relative entropy in most information theory texts [Cover and Thomas 1991]

Relative entropy provides a way to measure the “entropic distance” of one probability distribution from another (on the same sample space). However, this entropic functional is not actually a metric, because it is not symmetric and it does not actually satisfy the triangle equality. Nevertheless, relative entropy often functions as though it were a metric, largely because as its value approaches zero, it simultaneously approaches symmetry as well as the triangle equality.

Thus, in many applications relative entropy functions as a metric. For example, in predictability applications, if a sequence of probability distribution exhibits the fact that the relative entropies of consecutive distributions approaches zero, then the sequence approaches a stationary (or equilibrium) distribution. So, relative entropy is important to applications of information theory to predictability.

Mutual Information

Mutual information is an entropic functional that measures the degree to which two specified chance variables in a joint distribution relationship are stochastically dependent.

Recall, given a joint sample space (X, Y) with established constituent probability distributions for X and Y, that there is exactly one joint probability distribution on (X, Y) – call it $p_0(X, Y)$ – that is stochastically independent.

Thus, given another joint probability distribution – call it $p_K(X, Y)$ – with the same constituent distributions on X and Y, it will have to be, to some degree, stochastically dependent. The question is, “How stochastically dependent is $p_K(X, Y)$?”

Mutual information provides a measure of how stochastically dependent $p_K(X, Y)$ is.

It accomplishes this by using *relative entropy* to compare $p_K(X, Y)$ to $p_0(X, Y)$ – from the perspective of $p_K(X, Y)$. In other words, mutual information measures “how far away” p_0 is from p_K using relative entropy to perform the measurement.

The symbol for mutual information of joint distribution (X, Y) is I(X; Y). However, as we have discussed, there are uncountably many possible joint distributions whose symbol is (X, Y). They all have the same two sample spaces X and Y, and those sample spaces all have the same constituent (“marginal”) probability distributions p(X) and p(Y). However, all of these joint distributions have unique joint probabilities. Therefore, we shall distinguish then via the subscript K. As well, we must specify which of these distributions we are using mutual information to measure its degree of stochastic dependence (or “distance” from stochastic independence”). We shall make this distinction with the subscript K.

So, the mutual information $I_K(X; Y)$ is:

$$I_K(X; Y) = D(p_k(X, Y) \;||\; p_0(X, Y)\;) = \sum p_k(x,y) * \log (p_k(x,y)/ p_0(x,y))$$

Degree of Dependence and Relationships among Entropic Functionals

In the previous sections, we have discussed the idea that statistical dependence, in some sense, reduces entropy. This sense can be depicted by the fact that the joint entropy p(X, Y) can never be more that the sum of the constituent entropies p(X)+p(Y). And, it can only be the same if X and Y are statistically independent. But if X and Y are statistically dependent – as they are for most interesting applications [Kleeman 20012, lecture 3, p. 3.] – then p(X, Y) < p(X)+p(Y). It is in this sense that statistical dependence reduces entropy.

With the previous three entropic functionals – all defined in terms of the concept of *joint distribution*, we emphasized their leveraging the concept of entropy to provide some kind of measure of statistical dependence or independence between two chance variables (or probability distributions or probability spaces. These entropic functionals included conditional entropy, relative entropy and mutual information.

We mentioned, but did no emphasize, that all of these can be generalized to consider any finite number of chance variables – and are not limited to only two. For example, the mutual information of 5 chance variables – I(P;Q:R;S;T) is a measure of the degree of mutual stochastic dependence among those five probability spaces (or probability distributions) or chance variables.

Entropy Rate

These entropic functionals and their reduction of entropy under statistical dependence, has an immediate application to the prediction of the long-term behavior of discrete dependent stochastic processes. This includes Markov processes.

Discrete stochastic processes are probability theory's representation of dynamical systems and their time evolutions. Whenever those processes are *mutually dependent*, we can bring to bear the entropic functionals of this section to measure the degree to which they are predictable.

Here's how. Recall that a discrete dependent stochastic process can be understood as a sequence of conditional probability distributions, each conditioned on one or more of the previous time steps in the sequence. It turns out that under a variety of conditions, the "average of the joint entropy" converges to a limit – a limit that can be calculated.

Such as average joint entropy is called *entropy rate*. Entropy rate, when it exists, implies a degree of predictability of the dynamical system.

Admittedly, as time steps are added to a stochastic process, then its joint entropy an never get smaller. And as long as the entropy of the added time step is non-zero, then the entropy of the stochastic process will increase over time.

However, the entropy rate is essentially an entropic measure of the growth of the entropy of the process. And the value of the entropy rate limits that growth. Lets now discuss the conditions under which the limit named entropy rate exists, and therefore limits the growth of the entropy of the dynamical system.

First, entropy rate is defined as:

$$H(\chi) = \lim_{n\to\infty} 1/N * H(X_1, X_2,\ldots X_n)$$

In other words, *entropy rate* is the limit of the average entropy of the first n steps of a discrete stochastic process as n increases without bound. What is of interest is the conditions under which this limit exists, and thus entropy rate. In this section, we shall visit some of the conditions that entropy rate does exists, and therefore bounds the growth of entropy of a stochastic process, and of a dynamical system that it models.

We shall discuss three special conditions that ensure that entropy rate exists, and therefore bound the growth of entropy in a process. Refer to [Kleeman 2012, lecture 3] for more discussion on these issues.

The first case is a stochastic process in which the time steps are independent and identically distributed (i.i.d.) chance variables. Thus, in this case, we don't even have mutual statistical dependency, and yet the *entropy rate* exists anyway. What "tames" the entropy of such a stochastic process is simply time-homogeneity.

In other words, the fact that the distributions for all of the time steps are the same is enough for the limit (involved in entropy rate) to exist. In fact, in this case, we can say that the *entropy rate* of the process is equal to the entropy of any of its individual time step distributions.

But, we can get even more "narrowing" of our prediction regarding the time evolution of a discrete stochastic process if there is, in fact, mutual dependency. Lets look at two special cases of that.

First, lets define a concept that is related to entropy rate for the case that mutual dependence is involved. This concept is called *modified entropy rate*, but I prefer to call it *conditional growth*, because it calculates the amount of entropy added to the joint entropy of a stochastic process by a single new time step (variable), given that the new time step may be statistically dependent on some of the earlier time steps. Of course, if the newly added time step is mutually independent with all of the previous time steps, then the amount of entropy that it adds to the joint entropy is the same as its individual entropy $H(X_n)$. Otherwise the amount of entropy added to the joint entropy is less than $H(X_n)$. This will be the case for the "interesting" stochastic processes that occur in most applications.

"Modified entropy" calculates the conditional entropy of the nth time step, given all of the previous time steps. Then, modified entropy takes the limit of that as n increases without bound. In other words, modified entropy is defined as:

$$H'(\chi) = \lim_{n\to\infty} H(X_n \mid X_1, X_2,\dots X_{n-1})$$

Clearly, as indicated above, modified entropy is the amount of entropy added to joint entropy by the addition of a new term X_n. It can be shown that if X_n is mutually independent of the previous time steps of the process, then the amount of entropy added to the conditional entropy is $H(X_n)$.

But if there is statistical dependence involved between X_n and any of the previous tie steps, then the amount of entropy added by including X_n is less than $H(X_n)$. Thus, it is easy to see why I prefer the name *conditional* growth.

This additional amount of extra entropy can be as little as zero, if the conditional entropy $H(X_n \mid X_1, X_2,\dots X_{n-1})$ is zero. Therefore, the amount that the joint entropy grows is reduce by the amount of statistical dependence between X_n and the previous time steps. This "growth" can be no more than (X_n), but as little as zero – even though $H(X_n)$ is positive.

The first special case involving mutual dependence is when we the stochastic process involved is stationary after some time step. That is, for some time t, $p_{t-1}(x,y) = p_t(x,y)$ for all times t and sample points (x,). In this case we are interested in dependent stochastic processes, so each sample point will be pairs.

For these stationary stochastic processes, it can be shown that *entropy rate* is equal to *modified entropy rate*, or as I call it, *conditional growth*. That is, for a stationary dependent stochastic process,

$$H(\chi) = H'(\chi).$$

This equality is useful because in the case of stationary dependent stochastic processes, one can focus on the conditional entropy added to the process as a whole. An as long as there is there is mutual dependence, then this growth rate gets smaller, or can be used to narrow predictions.

Another special case involves stationary, time-homogeneous Markov processes. Since we are dealing, in this case, with stationarity, we can calculate entropy rate $H(\chi)$ by using the definition of modified entropy $H'(\chi)$, which takes the limit of the conditional probability of the last term only – and thus simplifies both the calculation and permits to think in terms of "growth". Recall that the definition of modified entropy is

$$H'(\chi) = \lim_{n \to \propto} H(X_n \mid X_1, X_2, \ldots X_{n-1})$$

However, since we are dealing in this case with a Markov process, then we know that the conditional probability of X_n given the outcomes of all prior steps is the same as the conditional probability of X_n given the previous step only – since that is the Markov property.

This means that, in this case, our definition of modified entropy reduces to:

$$H'(\chi) = \lim_{n \to \propto} H(X_n \mid X_{n-1})$$

which, of course, is much easier to calculate.

Bayesian Analysis

Suppose, by some mechanism, one has a probability distribution that one suspects is a reasonable representation of some phenomenon. However, it would be desirable to make use of additional observations to refine this distribution to an updated on. Bayesian statistics is methodology for doing this. Based upon certain relationships involved in joint probability distributions and their related conditional distributions, this methodology specifies a technique for updating the probabilities of an initial estimated distribution with a refined version. The initial distribution is called the *prior distribution*, and the refinement is called the *posterior distribution*.

This technique can be used iteratively to continue to utilize new data to develop more and more refinements. That is, in the first iteration, the initial prior distribution is refined to become the first posterior distribution. In the second iteration, the posterior distribution of the first iteration becomes the prior distribution of the second iteration and a new posterior distribution results for the second iteration.

This iterative process can be continued until the prior and the posterior distributions of an iteration become "very close" to each other. This occurrence can be interpreted as an approximate convergence of sequence of posterior distributions to the final one.

Once can interject one of the entropic functionals of information theory at this point to assist in assessing whether a convergence has effectively occurred. The entropic functional in question is *relative entropy*. This functional works here because it takes two probability distributions as parameters and calculates a quasi-measure of their "distance". While relative entropy is not a true metric, it approaches one whenever the value of relative entropy nears zero.

The result is that Bayesian analysis can also be brought to bear to develop a converging sequence of probability distributions whose approximate limit is a probability distribution that can approximate a stationary limit for the dynamical system being modeled.

Also, because Bayesian analysis is based on conditional probability, the resulting limiting distribution of the analysis can be reconstructed as a conditional distribution. In other words, we have a conditional, or dependent, stochastic process, or (ODSP) to work with as the result of the Bayesian analysis.

A point of commonality of Bayesian analysis and information theory is that neither requires that the distributions involved are random variables in the strict sense that we are using in these articles (that the distributions in question have a "value function" defined.) This fact makes both information theory and Bayesian analysis amenable to use by organodynamics.

Concerning the Stability of Some Stochastic Processes

In this article, we have tried to show that not all stochastic processes must "wander off into chaos"; but that, instead, there is a class of stochastic processes that asymptotically approach stability.

To make this case, we relied on a family of entropic functionals to show that conditional entropy and related functionals can establish some conditions under which a certain kind of entropy approaches a finite limit over time. This kind of entropic functional is named *entropy rate*.

We looked at some cases in which entropy rate does converge, and is therefore defined. We also looked at some case where the entropy rate can be simplified.

Entropic Functionals and Prediction Methodology

We have discussed above a number of entropic functionals from information theory and how the leverage any mutual statistical dependency among probability spaces in order to reduce the uncertainty of a stochastic process.

However, we have not yet constructed a method for applying these tools to prediction in the face of uncertainty. Prediction methodology is a growing field that has produced somewhat elaborate and often vague methodological results. However, we shall sketch an approach to prediction, using information theory and entropic functionals, in this section that we have collected from various sources.

The general idea [Kleeman 2012, lecture 8] is to define a *prediction distribution* and to determine an *equilibrium distribution*. The prediction distribution drives the stochastic process that should converge asymptotically to the second distribution involved: the *equilibrium distribution*.

The equilibrium distribution is a stationary distribution to which the prediction distribution approaches asymptotically. Thus, the *equilibrium distribution* represents the prediction. The modeling problem, then, is to determine what to use for the prediction distribution. Generally, the prediction distribution is a family of distributions that are defined by one or more parameters. A dynamical model of some kind is then used to evolve the model forward, where the parameters may be refined at each time step [Kleeman 2012, lecture 8, p 4].

For dependent stochastic processes, a transition matrix can be used to evolve the process forward. And the entropy rate can be used to ascertain that the process converges.

In addition, the relative entropy can be used between the evolution of the *prediction distribution* and the *equilibrium distribution*, with the *equilibrium distribution* serving as the prior. One can check this relative entropy between the current time step of the prediction distribution with the equilibrium distribution. When the relative entropy falls to a value close to zero, then the model of the equilibrium distribution is validated.

Also, [Majda, et. al. 2014] present a framework in which the predictability of a dynamical system, modeled as a stochastic process, can be quantified using a combination of relative entropy and mutual information.

All of this research indicates that a methodology for developing predictive stochastic models of complex dynamical systems is possible. In other words, using only probabilities – as relative entropy and mutual information do, predictable models of complex dynamical systems can be developed, under the right mathematical conditions.

The material presented here is not sufficient to form a modeling methodology for applying entropic functionals to stochastic prediction. However, some work has been done in this area, and there is the expectation that this approach is promising. However, considerable further research is needed in order to turn it into a viable modeling paradigm.

What remains to be accomplished is the development of a methodology that takes an equilibrium distribution and a stochastic process consisting of a family of prediction distributions and determine how to use these entropic functionals to satisfactorily demonstrate that the prediction distribution asymptotically approaches the equilibrium distribution arbitrarily close.

This initiative is should be pursued as the next major effort in this research.

Recap: Information Theory and Stochastic Dynamics

Entropic functionals work by reducing the degree of uncertainty of unknown (e.g. “future”) outcomes of probabilistic phenomenon. Given the nature of the process being modeled, this degree of the “narrowing” of uncertainty can range from none at all to exact determination.

The entropic functionals, then, provide a specific type of nondeterministic dynamics by using the probabilities of a probability space (distribution) model to “narrow” the prediction of future outcomes to a small subset of the whole state space.

Whereas, deterministic theories succeed in narrowing their predictions to a single member of the state space, nondeterministic theories can't guarantee that much accuracy. They can however, significantly reduce the size of their prediction space. Entropic functionals, as a stochastic theory, uses probabilities to implement this reduction.

Entropic functions accomplish this by measuring the distribution of uncertainty that is inherent in the probabilities of the application at hand. The less uncertainty is found inherent in the distribution of probabilities, the more the prediction is narrowed.

Implicit in this finding is that: All that the entropic functionals need as inputs is the probabilities of an application. In other words, once an application is modeled as a probability space (with a probability distribution), then entropic functionals can be used to predict its time evolution, to some extent. The entropic functionals describe the *degree of narrowing* that can be expected by the probability distribution of the application, while the probability distribution itself represents the "weighted" collection of outcomes to which the prediction is narrowed. This weighting is provided by the probability measures themselves.

These "versions of entropy" (entropic functionals) provide a flexible set of views of this predictability. Some can account for the effects of new information; others can predict several time periods into the future; while others work one step at a time.

An advantage of entropic functionals is that they can be used to compare the stability, degrees of opportunity and other interpretations of uncertainty between two diverse applications that have very different state spaces. However, a disadvantage is that they can't make very good use of non-probabilistic parameters associated with an application. These assertions suggest a comparison of the functionals and techniques of mathematical statistics and those of information theory, which we shall entertain next.

Stochastic Dependence: The Key to Constrained Behavior

An undercurrent to this article is this: the key to constrained behavior in stochastic dynamics is dependent conditional probability. In a joint distribution between two or more chance variables, and in particular when those two variables represent distinct time steps within a stochastic process, the degree of stochastic dependency is the degree of constraint on the unpredictability, the uncertainty, of behavior of the process over time.

We have said that the "dynamics" of a dynamical systems theory is some mechanism by which the time evolution of a target system can be constrained. In organodynamics, such mechanisms are based on conditional probability and conditional entropy (in any of its forms, including mutual information and entropy rate). These mean that for any specific application, some rational must be demonstrated that shows that consecutive outcomes in a process of the target application are not always stochastically independent. For example, if for all events x_i at time t and x_j at time t+n, it can be shown that there is some reason within the domain of the application so that it is not always the case that $p(x_i, x_j) = p(x_i)p(x_j)$, then the outcomes of the process are not mutually stochastically independent. In such a case, then a stochastic constraint exists on the process.

Of all of the possible joint distributions of two chance variables, only one of them is stochastically independent; whereas all of the rest (infinitely many) are, to some degree, stochastically dependent. Thus, stochastic dependence is, in some sense, the ordinary case; where is stochastic independence is very special – and not to be casually assumed.

The measure of the absence of stochastic independence, and thus some degree of stochastic constraint, is *mutual information*, *entropy rate* and certain others of the entropic functionals defined by information theory. And, these are not binary measures. Rather, the measure the *degree to which* stochastic dependence is at work within the process.

The higher degree of stochastic dependence (the lower the degree of stochastic independence) over time in the process, then the more constrained is the uncertainty of behavior in the process.

And, most of the entropic functionals that we have been discussing are, in one way or another, an entropy-like measure on the degree of to which these chance processes are either "tamed" or "wild" in this regard. In fact, we have shown that *determinism* is a *special case of stochastic dependence* where stochastic dependence is at its maximum.

For example, consider the MCD exemplar system of atoms in a closed space. Recall that in this process, the outcome of the system at any point in time is some molecular configuration of all the atoms in the space. Suppose that we know what that configuration is at time t_1. If knowing what the configuration is provides any helpful information in guessing which configurations are more likely at time t_2, then we have to conclude that there is stochastic dependence going on between times t_1 and 2. (It would also mean, in the case on this example, that some of the forces of nature – especially the electromagnetic force - were at work within this example; thus adding some degree of determinism.)

That is to say, there is something about the nature of the application at hand that gives reason to argue for stochastic dependence. It is this kind of condition that provides a rationale for concluding that stochastic dependence is involved in the organodynamic process that models the application. Notice that the rationale for arguing for stochastic dependence comes from the peculiarities of the application. However, when translated to pure organodynamic theory, only probability and information semantics are needed.

This essential stochastic dependency, then, is the key to constrained behavior, including predictability in chance dynamics. We have seen that the following entropic functionals, in one way or the other, are measures of this: *conditional entropy*, *mutual information* and *entropy rate*.

Unfortunately, many scientists fail to take the constraining affects of stochastic dependence into account when they are making various arguments concerning the nature of chance variation. This failure frequently leads to their rejection of chance phenomena of describing well-behaved phenomena.

One of the most notorious cases of this misunderstanding – in my view - is that of astronomer Fred Hoyle – inventor of the steady state theory of the origin of the universe – along with his collaborator N. Chandra Wickramasinghe. In their article [Hoyle and Wickramasinghe 2001] they argue that it is unreasonably improbable that life could have begun as a random event in a "warm little pond" – which is of course a reference to Darwin. I suggest that the reader inspect their argument below to see if you can spot their erroneous assumption (according to my above argument):

> The hypothesis questions the viability of chemical processes in a warm little pond. Would these processes yield the molecular arrangements of such observed biological structures as DNA and RNA, or at the enzymes for which such structures code? A typical enzyme is a chain with about 300 links; each link being an amino acid of which there are 20 different types used in biology. Detailed work on a number of particular enzymes has shown that about a third of the links must have an explicit amino acid from the 20 possibilities, while the remaining 200 links can have any amino acid taken from a subset of about four possibilities from the bag of 20. This means that with a supply of all the amino acids supposedly given, the probability of a random linking of 300 of them yielding a particular enzyme is as little as
>
> $$\frac{1}{(20)^{100}.4^{200}} \approx 10^{-250}$$
>
> The bacteria present on Earth in its early days required about 2000 such enzymes, and the chance that a random shuffling of already-available amino acids happens to combine so as to yield all the required 2000 enzymes is
>
> $2000!\ [10^{-250}]^{2000}$
>
> which works out at odds of one part in about $10^{500,000}$, with the factorial hardly making any difference, large as it might seem.
>
> A probability as small as this cannot be contemplated. So to a believer in the paradigm of the warm little pond there has to be a mistake in the argument. [Hoyle and Wickramasinghe 2001]

Of course, their intention with this line of thinking is to present a "toy" thought experiment, rather than a complete argument. Nevertheless, their rhetoric exposes their assumption of *stochastic independence*. This is clear by virtue of the fact that their calculation for the joint event of the assembly of 300 amino acids into an enzyme chain is the *product* of those 300 events individually. Such multiplication is proper for calculating the probability of a joint event if, and only if, those events are stochastically independent!

In other words, they arrived at their probability estimate for a chance-based joint event by assuming stochastic independence. Of course, they did not state this assumption. Rather, they depended on a readership that, unfortunately, also has the propensity to make this assumption (without realizing it) whenever joint probabilities are used. But I claim that an assumption of stochastic independence is unwarranted here.

Hoyle and Wickramasinghe are making simplified assumptions because they need some theoretical way to arrive at joint probabilities, rather than empirically observing them – simply because in this case taking an empirical approach is too massive an undertaking, and not available to them in any event.

However, they present no evidence that stochastic independence is warranted. Moreover, I suspect strongly that electromagnetic forces involved in linking amino acids into polypeptide chains has the effect of biasing (conditioning) the chemistry toward some degree of stochastic dependence.

Another unfortunate assumption on the part of many thinkers is that, when it comes to the issue of randomness versus deterministic, that the available choices are perceived to be two and only two: either total determinism or total randomness. In other words, there is a broad failure to understand that *randomness comes in* degrees. This failure leads to the false assumptions we are discussing.

Also false is the thinking that “Random joint events necessarily implies stochastic independence.” Of course, information theory – with its entropic functions and their continuum of values between zero and some maximum real number – demonstrates a formal understanding of the continuous nature of randomness and uncertainty.

My point is that saying that “randomness is involved” need not mean either that 1) the distribution is uniform, or 2) a joint distribution is stochastically independent. In other words, saying that “randomness is involved” need not mean that entropy is at a maximum.

Rather, knowing that “randomness is involved” says nothing at all about the degree of uncertainty. It could be anything from zero to the maximum entropy defined for the situation. What one needs to characterize the degree of uncertainty of the situation is to apply the pertinent entropic functional, which will provide a measure of the degree of uncertainty of the situation.

Other Applications of Entropic Functionals

The semantics of entropy is that it is the measure of the *degree of uncertainty* inherent in a probability space. As with most mathematical constructs, this use of the word “uncertainty” is merely one of many linguistic interpretations of the mathematics. Shannon provided three more such interpretations. He suggested that entropy is also a measure of the amount of choice inherent in a process, the amount of information produced by a process, and the amount of change of information produced by a process. [Shannon 1948, p. 10.]

In truth, just about any characterization of a probability application that relates to *chance variation* is subject to measurement by entropy. Some of these include degrees of: stability, robustness, opportunity, resilience, reliability and adaptability. The interpretation of the “meaning” of applying entropy to probability applications is limited only to the number of applications that practitioners find to model using information theory. In fact, it is the context and circumstances of those applications that often give meaning to the application of entropy and entropic functionals.

As stated, information theory is applicable to any situation that has probabilities and that can be modeled as a probability space. As with any mathematical system, the semantics of applying that mathematics is often determined by the application itself.

Application of Entropic Functionals to Organodynamics

Lets now return to our model of human cognition using fMRI data. We shall now move this example forward to include the application of information theory to exemplify how the entropic functionals can be used to model our expanded notion of the dynamics of a dynamical systems model.

First, let’s recap where we were when we last visited this example. Our subject was a human brain that is understood to consist of billions of neuron cells. In addition, the brain is organized into some number *brain* regions, which form a portioning of all of the brain’s neurons.

Our fMRI data consists of two kinds of observations: 1) *neuronal activity* and 2) *neuronal interactions* (between a source neuron and other neurons). Neuronal activity is evidence of the participation of an observed neuron in an observed mental activity (cognition). This set of brain neurons constitutes the *underlying set* of an organodynamic system. We shall subsequently be interested in how many ways this *underlying set* can be organized over time. Each region of the brain can be further organized into compartments by observing the cooperation of groups of neurons in several functional mental activities.

Together, the division of the brain into "regions" and the functional grouping of neurons within those regions suggest a finer-grained set of neuronal organization than the brain region approach alone. This organization will consist of a number of functional groups that, taken together, include all observed neurons. However, this set of functional groupings will not, in general, be mutually exclusive or non-overlapping.

Nevertheless, this set of functional groups forms a *cover* of the modeled neurons. And this cover induces a topology on the underlying set of neurons. Moreover, each open set in that topology is associated with a relation (set of ordered pairs) of its interacting neurons. Together, this topology and its associated relations represent the way that the cognition of the brain is organized at the moment of the fMRI reading. This topology and its associated neuronal relations constitute the structure ("extended topology") that organodynamics uses to represent the (momentary) organization structure of the cognition of the brain. We call this structure an *organization* of the underlying set (of neurons), and symbolize it with O_{TRS}. Thus, O_{TRS} represents system state in organodynamics. And the set of all possible such system states (extended topologies) is a state space in organodynamics. This state space is an *organodynamics state space* (OSS) because its elements are organizations. We symbolize it as $\mathbf{O_S}$.

However, as cognitive function changes over an fMRI session, the set of neurons that are active changes, as well as the active functional groups and thus the extended topology will change over time. This change of organization over time constitutes a *trajectory*.

However, this change in organization of the system over time is subject to chance variation. We therefore represent this change from one point in time to another with a conditional (dependent) probability distribution p(Y|X). That is, we want to represent this particular dynamical system as an *organodynamic dependent stochastic process* (ODSP).

In this case the sample space X is $\mathbf{O_S}$, and so is the sample space of Y. (For now, lets drop the "S" subscript on O.)

Thus, we are dealing with the conditional probability distribution $p(\mathbf{O} \mid \mathbf{O})$. But we have to take into account that the dependence of the outcome of step n may be influenced by any of the previous n-1 steps. This is particularly true of cognition, because distant past memories can affect present outcome. This means that we should deal with a more general conditional probability of $p(\mathbf{O_n} \mid \mathbf{O_{n-1}}, \mathbf{O_{n-2}}, \ldots , \mathbf{O_1})$. Or course, as a first approximation, we could make the simplifying assumption that the Markov property holds for cognition (one-step memory only). In that case, we can use $p(\mathbf{O_n} \mid \mathbf{O_{n-1}})$.

If we make the assumption that the underlying set of neurons does not change over time, then the possible topologies on that underlying set does not change over time. It is reasonable to also assume that the constituent probabilities of $\mathbf{O_S}$ do not change over time. Of course, $p(\mathbf{O_n} \mid \mathbf{O_{n-1}}, \mathbf{O_{n-2}}, \ldots , \mathbf{O_1})$, if for no other reason than each successive time step is dependent on additional previous steps.

Nevertheless, we are dealing here with an *organodynamic dependent probability space* (ODPS) that can be abbreviated to an *organodynamic dependent probability distribution* (ODPD). While this space is a *chance variable*, it is not a *random variable* as we have defined it, since there is no reasonable or useful mapping from each O_{TRS} sample point to the real numbers. This fact leaves us powerless to define any moments (mean, etc.) or central moments (variance, skewness, kurtosis, etc.) for these spaces – mathematically they are indefinable.

However, the entropic functionals of information theory use only probabilities as inputs. Consequently, the provide a powerful set of construct with which to measure and characterize a OPD and, moreover, *an organodynamics dependent stochastic process* (ODSP).

But the ODSP is a dependent stochastic process. This means that we can apply a number of the entropic functionals to it. For example, we can calculate the entropy rate of the ODSP to see if it is asymptotic to a limiting OSP. If so, then we know that it approaches a particular behavior in the long run. Another approach can be used if one suspects a particular limiting OSP distribution. In this case, the relative entropy between the suspected limiting OSP can be calculated for each time step in the ODSP. This results in a sequence of relative entropy real values. If this sequence approaches a limiting value, then the ODSP can be understood a being well behaved in the long run.

If one can further assume that the stochastic process in question exhibits the Markov property (it is a Markov process), then each time step need only be dependent on it previous time step – making the calculation of conditional probability simpler.

For organodynamics, we have said that the stochastic processes involved are dependent, piecewise-homogeneous processes. This piecewise homogeneity means that *segments* (finite contiguous time steps) will be time homogeneous – that is, they will be Markov chains. Moreover, from Markov chain theory, we also know that if they are aperiodic ergodic chains (one can go from any state to any other in any single time step), then they are guaranteed to be stationary.

As mentioned earlier, the development of methodologies for characterizing the long-run behavior of stochastic processes in general using information theory and entropic functionals is nascent. Nevertheless, some work has been done [Majda, et. al. 2014]. However, much research is needed in this area.

Markov Transition Matrices in ODSPs

In general, any time step of dependent stochastic process in probability theory may depend upon or be influenced by the outcomes of any of the time steps that precedes it in the process. This means that the conditional probability at time t_n may be influenced by the outcomes at any or all of the time steps at t_1 , t_2 ,…, t_{n-1}.

We have discussed that sometimes it happens that the outcome at time t_n is conditioned on the outcome at time t_{n-1} only, and not by the outcomes of any previous times. When this happens, we say that the Markov condition is satisfied. And, all of our conditional probabilities only have to go back one time step. If we can make the assumption that the Markov principle holds, our expression for all of the probabilities in our ODSP will be enormously simplified.

At least for a first approximation, I recommend that we first do this in organodynamics. Refinements can be made later. But our first approximation for any modeling venture will generally begin by making the Markov assumption.

That being said, then our first approximation ODSP will always be a Markov process. Therefore, we can use a Markov matrix to represent each time step. The reader is referred to [Kemeny and Snell 1976] for the use of Markov transition matrices in Markov processes.

Recall that a typical ODSP in organodynamics is piecewise-homogeneous. This means the following for representing an ODSP by using Markov transition matrices: The ODSP can be understood as a sequence of finite-length subsequences, each of which expressed the same Markov transition matric for each of its time steps.

Homogenization of the Piecewise Homogeneous OSP

We have seen how it is the nature of complex dynamical systems observed in the field are usually piecewise homogeneous; because, life is messy and probability distributions change over time within the same dynamical system. In fact, the underlying sample spaces also change.

For this reason, we have had to complexify the stochastic processes used in organodynamics, the ODSP, so that it allows the probability distribution to change over time. This had led me to characterize ODSPs generally as *piecewise homogeneous*.

Unfortunately, *piecewise homogeneity* complexifies any calculations that we might want to perform on the probability distributions. And, piecewise homogeneity is much more complicated to deal with than the complete homogeneity.

The good news is that there is a way that we can “homogenize” these piecewise homogeneous ODSPs so that they become a completely homogeneous – in which case a single transition matrix can represent every time step in the ODSP.

The bad news, however, is that, in order to do this we have to make the single new transition matrix very much larger than any of the distinct transition matrices involved in any of the time steps of the initial *piecewise homogeneous* OSP.

This single new transition matrix is very much larger because we essentially have to combine all of the sample points and all of the probabilities into a single joint ODPD.

Philosophically, this amounts to understanding, when the modeling process begins, what the entire “universe” of sample points and joint probabilities across all time will be. Then, a single probability matrix will be defined that includes all of those possibilities.

Within this new, more comprehensive, Markov matrix, all of the Markov matrices that showed up in the earlier ODPD will be “scattered” around this new comprehensive Markov matrix.

The behavior of the ODSP with the new, comprehensive, Markov matrix is that, over time, it will “wander around” this new comprehensive Markov matrix whenever the “homogeneity behavior” changes.

Preview of Part V

The long-range behavior of certain very complex dynamical systems, such as biological life, must be adaptive. Such behavior is of necessity not predictable. Nevertheless, it is constrained behavior in the sense that it does not “wander off into chaos”.

In Part V, we shall look at what kind of mathematical dynamics – based once again on chance variation, probability theory, information theory and stochastic processes – can be developed to characterize persistent long-term behavior in this class of complex adaptive dynamical systems.

References

[Ash and Doleans-Dade 2000] Ash, Robert B., Catherine A. Doleans-Dade; *Probability and Measure Theory*; Second Edition; Academic Press; San Diego, CA; 2000.

[Cover and Thomas 1991] Cover, Thomas M. and Joy A. Thomas; Elements of Information Theory; John Wiley & Sons, Inc.; New York; 1991.

[Doob 1996] Doob, J. L.; *Stochastic Processes*; John Wiley and Sons, Inc.; New York, Chichester, Brisbane, Toronto; 1953.

[Gallager 2011] Gallager, R. G.; *Discrete Stochastic Processes*; Draft of 2nd Edition; course syllabus; MIT Open Courseware, Electrical Engineering and computer Science graduate seminar; at http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-262-discrete-stochastic-processes-spring-2011, January 31, 2011.

[Hoyle and Wickramasinghe 2001] Hoyle, Fred and Wickramasinghe, N. Chanda; *Evolution of Life: a Cosmic Perspective*; On ActionBioscience.org at http://www.actionbioscience.org/original/wick_hoyle.html; May 2001.

[Jaynes 1957] Jaynes, E. T.; *Information Theory and Statistical* Mechanics; *The* Physical Review, Vol 106, No 4, 620-630, May 15, 1957.

[Kemeny and Snell 1976] Kemeny, John G., J. Laurie Snell; *Finite Markov Chains*; Springer-Verlag; New York, Berlin, Heidelberg, Tokyo; 1976.

[Khinchin 1957] Khinchin, A. I; *The Mathematical Foundations of Information Theory*; Dover Publications, Inc.; 1957; New York.

[Kleeman 2012] Kleeman, Richard; *Information Theory and Predictability*; Courant Institute of Mathematical Sciences, NYU; course syllabus, MATH-GA3011.001, *Information Theory and Predictability*; at http://www.math.nyu.edu/faculty/kleeman/syllabusinfo.htmlSpring, 2012.

[Majda, et. al. 2014] Majda, Andrew; Richard Kleeman; David Cai; *A Mathematical Framework for Quantifying Predictability through Relative Entropy*; Courant Institute of Mathematical Sciences, NYU;2014.

[Pierce 1980] Pierce, John R; *Information Theory: Symbols, Signals and Noise*; Second Revised Edition; Dover Publications, Inc.; New York; 1980.

[Ross 1996] Ross, Sheldon M.; *Stochastic Processes*; Second Edition; John Wiley and Sons, Inc.; Hoboken New, Jersey; 1996.

[Shannon 1948] Shannon, Claude E.; *The Mathematical Theory of Communication*; At http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf; 1948.

[Tolman 1938] Tolman, Richard C.; *The Principles of Statistical Mechanics*; Oxford University Press; 1938; unabridged Dover Edition; Dover Publications, Inc. New York; 1979.

[Lagrangian 2014] Wikipedia article on Lagranginan Multiplier; at http://en.wikipedia.org/wiki/Lagrange_multiplier#Example_3:_Entropy. Proof that the uniform distribution of a finite probability space has maximum entropy.

[Lemons 2002] Lemons, Don. S.; An Introduction to Stochastic Processes in Physics; The John Hopkins University Press, 2002.