2011

Towards an Information Theoretic Framework for Evolutionary Learning

Stuart William Card
Syracuse University

Abstract

The vital essence of evolutionary learning consists of information flows between the environment and the entities differentially surviving and reproducing therein. Gain or loss of information in individuals and populations due to evolutionary steps should be considered in evolutionary algorithm theory and practice. Information theory has rarely been applied to evolutionary computation – a lacuna that this dissertation addresses, with an emphasis on objectively and explicitly evaluating the ensemble models implicit in evolutionary learning. Information theoretic functionals can provide objective, justifiable, general, computable, commensurate measures of fitness and diversity.

We identify information transmission channels implicit in evolutionary learning. We define information distance metrics and indices for ensembles. We extend Price’s Theorem to non-random mating, give it an effective fitness interpretation and decompose it to show the key factors influencing heritability and evolvability. We argue that heritability and evolvability of our information theoretic indicators are high. We illustrate use of our indices for reproductive and survival selection. We develop algorithms to estimate information theoretic quantities on mixed continuous and discrete data via the empirical copula and information dimension. We extend statistical resampling. We present experimental and real world application results: chaotic time series prediction; parity; complex continuous functions; industrial process control; and small sample social science data. We formalize conjectures regarding evolutionary learning and information geometry.

TOWARDS AN INFORMATION THEORETIC FRAMEWORK FOR EVOLUTIONARY LEARNING

Stuart William Card

M.S. Syracuse University, 1990

B.S. Utica College, 1985

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer and Information Science in the Graduate School of Syracuse University

August 2011

Table of Contents

List of Illustrative materials

Acknowledgements

The research upon which this dissertation is based has occupied many years and consumed a significant share not only of my life but also of the lives of several other individuals who have very patiently guided and supported me. First are my parents, Philip and Marilyn Card, who provided my earliest education, imbued me with a love of learning, and by example invested me with a desire to serve my fellow creatures by contributing to the store of human knowledge and its practical applications. Second collectively are the many teachers in schools both formal and informal who labored to inform me in the full sense of the word. Third are my coworkers at Critical Technologies Inc., who have understood my need to pursue this research, despite the time it has taken me away from work having been greatly to the detriment of our business, and who have struggled to fill the gaps my frequent absences have left. Fourth is my advisor Professor Chilukuri Mohan, who has understood my need to try to fulfill my obligations to my coworkers and others outside academia, despite the time these have taken me away from this research having greatly delayed this dissertation, and who never failed in any of our meetings to encourage me and lead me to some new insight that I would not have found without his guidance. Fifth and always last because she makes that sacrifice for me but never least because I love her for that and more, is my wife Brenda Card, who has worked outside our home to support us and done all the tedious chores of our life together while I have done my work. While I have enjoyed the thrill of learning, they have borne my burdens for many years, and I can thank them only by dedicating my work to free minds.

1 Introduction

Pluralitas non est ponenda sine necessitate.

William of Ockham, circa 1320

We consider the source with the maximum entropy subject to the statistical conditions we wish to retain. The entropy of this source determines the channel capacity which is necessary and sufficient.

Claude Shannon, 1948 [92]

Information can tell us everything. It has all the answers. But they are answers to questions we have not asked, and which doubtless don't even arise.

Jean Baudrillard, 1987 [9]

1.1 Thesis

By “evolutionary learning”, we mean both population learning by evolution (or evolutionary computation) and the evolution of individuals that learn[1]. The former is the overall scope of our inquiry and the latter is our primary focus. We leave the definition of “learning” to others, and say only that clearly it involves processing information, generally sensor signals and/or communication symbols, the domain of information theory.

Information theory enables rigorous definition of metrics quantifying information flows, gains and losses, as well as other notions such as epistasis. It has been applied to machine learning (e.g. [80]), but rarely to evolutionary computation – a lacuna that this dissertation addresses, with an emphasis on explicitly evaluating and exploiting the ensemble models that are implicit in evolutionary learning. It is useful in all phases of evolutionary computation, from terminal and non-terminal set selection, through survival and reproductive selection, to objective evaluation of ensemble models as such at end-of-run.

[1] Most evolutionary algorithms are oriented towards optimization. Genetic Programming (GP) is oriented towards learning, and through the use of indirect encodings, can evolve individuals that develop and learn. GP is thus the canonical example of evolutionary learning used in our analytic and experimental work.

The vital essence of evolutionary learning consists of information flows between the natural or artificial environment and the entities differentially surviving and reproducing therein. Gain or loss of information in individuals and populations due to evolutionary steps should be explicitly and objectively considered in evolutionary algorithm theory and practice.

From our perspective, there are three fundamental desiderata – as evolution proceeds, the mutual information between the target and models should:

- ► not decrease substantially in the population;
- ► increase (concentrate) in the best individuals; and
- ► distill from the inputs, leaving behind their excess entropies.

We assert that biological evolution processes ensembles implicitly (e.g. populations of many individuals, each with a genotype of many chromosomes, each with many genes) but evolutionary computation practice typically flattens this to only a single level (populations). Ensemble processing may be important to the evolutionary process –

> There is no theoretical reason to expect evolutionary lineages to increase in complexity with time, and no empirical evidence that they do so. Nevertheless, eukaryotic cells are more complex than prokaryotic ones, animals and plants are more complex than protists, and so on. This increase in complexity may have been achieved as a result of a series of major evolutionary transitions. These involved changes in the way information is stored and transmitted.
>
> Eors Szathmary and John Maynard Smith, 1995 [101]

It also may be useful from a pragmatic engineering point of view –

> This paper proposes an evolutionary approach for the composition of solutions in an incremental way. The approach is based on the metaphor of transitions in complexity discussed in the context of evolutionary biology. Partially defined solutions interact and evolve into aggregations until a full solution for the problem at hand is found… In this model, we consider partially defined solutions that specify just an incomplete solution to the problem at hand. As a consequence, these partially defined solutions cannot hope to solve the complete problem unless they collaborate with each other. When beneficial collaborations are identified, a transition can occur, i.e., they are aggregated into a new (more complex) solution.
>
> Defaweux, Lenaerts, van Hemert and Parent, 2005 [34]

We develop and illustrate application of information theoretic methods for objectively evaluating ensembles as such. This enables more sophisticated survival and reproductive selection. It also provides a new instrument for observing the dynamics of evolution[2].

The remainder of Chapter 1 provides motivation and other preliminaries. Chapter 2 identifies information transmission channels implicit in evolutionary learning. Chapter 3 defines information theoretic measures for evaluating ensemble models. Chapter 4 develops algorithms for estimating these quantities and the confidence associated with those estimates. Chapter 5 applies these measures to evolutionary algorithm analysis, Chapter 6 applies them in computational experiments and Chapter 7 applies them to real data. Chapter 8 formalizes conjectures fundamental to evolutionary learning. Chapter 9 summarizes contributions of this work.

[2] This processing of ensembles has about it a feeling of *compression*, in both the computer science and thermodynamics senses, involving degrees of freedom, entropy and energy: this remains to be explored; this thesis may help provide a framework for such exploration. An early foray was [86].

1.2 Motivation

An important class of machine learning problems involves automatic construction of models of unknown (“black box”) processes with observed inputs and outputs. Our information theoretic approach was initially motivated by practical difficulties encountered in model construction by Genetic Programming (GP). GP symbolic regression applied to stochastic chaotic time series prediction should perform system identification with minimal preconceptions as to model form, producing not only predictions, but also parsimonious meaningful descriptions, capturing local and global characteristics of stochastic attractors, yielding insight into the hidden dynamics (see Appendices A: Background and B: Related Work).

However, we encountered difficulties such as premature convergence (early loss of diversity) when attempting to learn the defining equation for the simplest known chaotic flow [98][97]:

$$x''' = -ax'' + (x')^2 - x \tag{1}$$

Figure 1 shows the time domain behavior of the observable, which appears to be roughly sinusoidal, but with jittering amplitude, frequency and phase. Figure 2 shows the time domain behavior of the observable (displacement *x*) and its first 3 derivatives (velocity *x’*, acceleration *x”* and jerk *x’”*) over a shorter period of time. Figure 3 shows the frequency domain spectral footprint of the observable as a decibel plot of a Fast Fourier Transform (FFT). Figure 4 shows the self mutual information versus lag; the total span is approximately twice the period.

Figure 5 unfolds the attractor in phase space; it is a 2-D perspective projection of a 3-D space in which the coordinates are successive elements of a lag vector. Figure 6 again unfolds the attractor in phase space; it is a 2-D perspective projection of a 3-D space in which the coordinates are the observable and its first and second differences (estimates of its first and second derivatives). The attractor is neither as simple as it appears on a first look nor as complex as it appears on a second look: it is a non-linearly warped Möbius strip; but some trajectories make the Möbius half-twist and others do not, and these different trajectories are interleaved at arbitrarily close spacings (exponential divergence from initially nearby points, a.k.a. sensitive dependence upon initial conditions).

Due to deception and poor scaling, relatively high fitness individuals in early generations did not contain the building blocks needed to evolve correct solutions in later generations. A search for the underlying causes of these difficulties motivated application of information theory and reformulation of evolutionary objectives. For instance, commonly used fitness measures, such as root mean squared error (RMSE), often fail to reward individuals whose presence in the population is needed to explain substantial portions of the data variance. Fitness indicators must be developed that reward individuals for their potential incremental contributions to solution of the overall problem, perhaps by explicitly identifying building blocks suitable for recombination.

This motivates use of Mutual Information (MI) as a fitness indicator, conveying how much one item (a model in the GP population) reveals about another (the target data set). MI is invariant under transformations that can otherwise conceal an individual's potential contribution to solutions (Maarten Keijzer, private correspondence). MI is invariant to invertible transformations[3] and degraded at most by a calculable and typically small amount by non-invertible transformations, so it is a promising candidate for identifying high fitness simple forms in early generations [56] that are likely to be good building blocks for later generations [31]. For an extended series of motivating examples see Chapter 5.

Another failing of traditional indicators is an inability to objectively evaluate ensembles. From a biological (or economic) perspective, highly fit teams of symbionts are likely to stick together; less fit teams are likely to break up and their former members disperse, possibly joining more effective teams. Individuals that do not become members of highly fit teams are unlikely to survive*. Thus survival of biological individuals is promoted not only by individual fitness, but also by participation in a highly fit team;* this in turn requires not only that the team have high fitness, but also that the individual be permitted to remain on the team, presumably due to a perception by other team members that the individual is making a significant contribution to team fitness, and not imposing an excessive energy or economic cost upon the team.

[3] This is easy to show by contradiction of the Data Processing Inequality. For a detailed treatment of this see [60].

These biological notions have engineering utility: the value of an individual in an evolutionary learning context depends on its own quality as well as its contribution to ensemble quality. Traditional evaluations of ensembles have been *ad hoc*, e.g. using voting, summing or weighted averaging of evaluations of the constituent individuals to characterize the ensemble. Information theoretic functionals, by contrast, can objectively and quantitatively evaluate ensemble models as such, without requiring knowledge of how the constituents might best be assembled into a single more complex model.

Information theory is also applicable to other aspects of evolutionary algorithms. For instance, extant diversity indicators are often arbitrary, may reflect diversity irrelevant to solving the problem, and are typically incommensurate with fitness measures. By contrast –

- Information theoretic functionals can provide objective, justifiable, general, computable, commensurate measures of fitness and diversity.
- They can be used to guide input (terminal and/or non-terminal set), reproductive and survival selection.
- They should be heritable across mutation and recombination, and can remain so as parental fitness increases.
- They can be used to estimate potential improvement of the upper tail of a brood versus the more fit of their parents: evolvability, the *sine qua non* of evolutionary algorithm success.

1.3 Preliminaries

This section presents the GP model learning context, our notation, and the standard definitions of some information theoretic functionals including entropy, mutual information, redundancy and synergy. To clarify the approach, we focus on our initial test problem, the easiest we could identify that presents fundamental difficulties for automated model generators – Sprott’s simplest chaotic flow, Equation (1) above. This is a “jerk equation” of which the general form is:

$$x''' = f(x'', x', x) \qquad \textbf{(2)}$$

Many systems cannot be described by such a single, higher order, Ordinary Differential Equation (ODE), but must be written as a system of coupled first order ODEs; systems of order *o* may also be rewritten in that same form:

$$\begin{aligned} x_1' &= f_1(x_1, x_2 \ldots x_o) \\ x_2' &= f_2(x_1, x_2 \ldots x_o) \\ &\ddots \\ x_o' &= f_o(x_1, x_2 \ldots x_o) \end{aligned} \qquad \textbf{(3)}$$

Writing

$$\mathbf{x} = (x_1, x_2 \ldots x_o) \qquad \textbf{(4)}$$

and setting **x′** = **y,** we can rewrite in the general form

$$\mathbf{y} = \mathbf{f}(\mathbf{x}) \qquad \textbf{(5)}$$

into which problems of many types may be cast, so our approach is widely applicable.

Figure 7 illustrates a general jerk system. Figure 8 shows it in an alternate form, with direct feedback of higher order terms replaced by feedback of only the observable, passed through a chain of differentiators. Figure 9 generalizes to an arbitrary feedback system, where there may be multiple observables and/or state variables, as in Equation 3 above. Finally Figure 10 simplifies still further: to an open loop system, or to consideration of only the feedforward path of a feedback system; these correspond to Equation 5 above. Note that difficulties will arise in modeling a system such as that in Figure 8 due to the estimates of the derivatives being computed after the injection of observation noise.

Solving problems of this form by GP involves a population of *m* functions $\mathbf{f}_i$ regressed on *n* data points $\mathbf{x}_j$ to approximate $\mathbf{y}_j$ by $\mathbf{z}_{i,j}$:

$$\forall i \in [1..m] : \forall j \in [1..n] : \mathbf{z}_{ij} = \mathbf{f}_i(\mathbf{x}_j) \qquad \textbf{(6)}$$

The input **X**, target output **Y** and estimated output $\mathbf{Z}_i$ data vectors comprise sample distributions:

$$\begin{aligned} \mathbf{X} &= \{\mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_n\} \\ \mathbf{Y} &= \{\mathbf{y}_1, \mathbf{y}_2 \ldots \mathbf{y}_n\} \\ \forall i \in [1..m] : \mathbf{Z}_i &= \{\mathbf{z}_{i,1}, \mathbf{z}_{i,2} \ldots \mathbf{z}_{i,n}\} \end{aligned} \qquad (7)$$

We consider functions $\mathbf{f}_i$ to be model genotypes and output data sets $\mathbf{Z}_i$ (given the input data set **X**) to be model phenotypes; the genotype–phenotype correspondence is in general many-to-one.[4]

[4] We defer consideration of generative and developmental systems, indirect encodings, morphogenesis, ontogenesis, etc., where different environmental influences on multiple instances of the same genotype can yield different phenotypes.

Assuming discrete distributions with *v* distinct values, Shannon's entropy is defined[5] by[6]

$$\mathrm{H}(\mathbf{Y}) = -\sum_{i=1}^{v} \mathrm{p}(\mathbf{y}_i)\,\mathrm{lg}(\mathrm{p}(\mathbf{y}_i)) \qquad \mathbf{(8)}$$

Joint entropies of arbitrary numbers of terms are similarly defined. Entropy quantifies the *a priori* uncertainty (information deficit) in a distribution[7]. It is used to define the mutual information (MI) between distributions

$$\mathrm{I}(\mathbf{Y};\mathbf{Z}_j) = \mathrm{H}(\mathbf{Y}) + \mathrm{H}(\mathbf{Z}_j) - \mathrm{H}(\mathbf{Y},\mathbf{Z}_j) \qquad \mathbf{(9)}$$

where the distributions used to illustrate the definition are those of the output data sets of the target function and one of the candidate models, and H(**Y**,**Z**$_j$) is their *joint entropy*, i.e., the entropy of their joint distribution. The usual interpretation of mutual information is that it quantifies how much, on average, each of two random variables tells us about the other; here, how much the uncertainty in the target output is reduced by observing the model output, and *vice versa*. MI is not defined for functions, only for distributions: we seek better indicators of the fitness of functions; we attribute fitness to a function, based on the MI of its output data set with that of the target function, applying both to the same input data set.

[5] For continuous distributions, differential entropy is defined instead. It can be estimated as the limit of the sum of the discrete entropy of ever-finer quantizations of the continuous distribution and the log of the bin size. It does not have all the properties we desire. With care, the definition given for discrete data may be applied to discretizations of a continuous distribution.

[6] We adopt the notational convention for logarithms in equations of log() for base 10 (common), ln() for base e (natural) and lg() for base 2; in the text, when we use the word "log", either the base doesn't matter, or we mean 2.

[7] We calculate here the entropy of a sample distribution, possibly biased versus that of the hidden distribution from which our observed sample was drawn. In future work we may correct for bias, but it is unimportant here; see Section 4.1 for a comment.

Considering subpopulations (including potential reproductive pairings in selection tournaments), we quantify the information about **Y** provided jointly by outputs $\mathbf{Z}_j$ and $\mathbf{Z}_k$ of two candidate models by their incremental mutual information or marginal redundancy:

$$\mathrm{I}(\mathbf{Y};\mathbf{Z}_j,\mathbf{Z}_k) = \mathrm{H}(\mathbf{Y}) + \mathrm{H}(\mathbf{Z}_j,\mathbf{Z}_k) - \mathrm{H}(\mathbf{Y},\mathbf{Z}_j,\mathbf{Z}_k) \qquad \textbf{(10)}$$

We quantify the redundancy of populations by the “total correlation”:

$$\mathrm{C}(\mathbf{Z}_1,\mathbf{Z}_2 \ldots \mathbf{Z}_m) = \sum_{i=1}^{m} \mathrm{H}(\mathbf{Z}_i) - \mathrm{H}(\mathbf{Z}_1,\mathbf{Z}_2 \ldots \mathbf{Z}_m) \qquad \textbf{(11)}$$

The synergy-redundancy measure [26] or interaction information [69] is:

$$\begin{aligned} \mathrm{S}(\mathbf{Y};\mathbf{Z}_j;\mathbf{Z}_k) &= \mathrm{I}(\mathbf{Y};\mathbf{Z}_j,\mathbf{Z}_k) - ((\mathrm{I}(\mathbf{Y};\mathbf{Z}_j) + \mathrm{I}(\mathbf{Y};\mathbf{Z}_k)) \\ &= -\mathrm{I}(\mathbf{Y};\mathbf{Z}_j;\mathbf{Z}_k) \end{aligned} \qquad \textbf{(12)}$$

which in the 3 variable case agrees with our preferred one of the proposed definitions of the multivariate mutual information[8], except for the change in sign with an odd number of variables, as in the second line of Equation 12.

Note that H(), I() and R() must all be positive or zero; but S() can be negative (redundancy exceeding synergy), positive (synergy exceeding redundancy) or zero (synergy and redundancy equal, of which a special case is independence when they are both zero). In some cases, the mutual information between the target and each model may be zero, but the two models may jointly explain the target completely due to their synergy.

[8] http://en.wikipedia.org/wiki/Multivariate_mutual_information

Example 1: The simplest example of maximal synergy XORs fair coins x_1 and x_2:

$$y = \mathrm{f}(x_1, x_2) = x_1 \oplus x_2 \qquad \textbf{(13)}$$

Two simplistic candidate models are observations of each of the coins:

$$\begin{aligned} z_1 &= \mathrm{f}_1(x_1, x_2) = x_1 \\ z_2 &= \mathrm{f}_2(x_1, x_2) = x_2 \end{aligned} \qquad \textbf{(14)}$$

The entropies of the target and the models are:

$$\mathrm{H}(Y) = \mathrm{H}(Z_1) = \mathrm{H}(Z_2) = 1 \qquad \textbf{(15)}$$

The pairwise MI values are:

$$\mathrm{I}(Y;Z_1) = \mathrm{I}(Y;Z_2) = \mathrm{I}(Z_1;Z_2) = 0 \qquad \textbf{(16)}$$

Yet synergy of the two models enables them jointly to fully explain the target:

$$\mathrm{I}(Y;Z_1,Z_2) = 1 \qquad \textbf{(17)}$$

An “information diagram” is a Venn diagram of entropies and mutual information, possibly including higher order MI (interaction information), where *areas of overlap of 3 or more terms can be negative*[9].

Figure 11 is an information diagram showing 2 models of an input/output relationship. In Figure 12, the corresponding areas are labeled not symbolically but rather with their numerical values from Example 1. The interaction information or synergy-redundancy measure S(Y;Z_1;Z_2) is positive (one bit), so the multivariate mutual information I(Y;Z_1;Z_2) is negative.

[9] http://en.wikipedia.org/wiki/Information_diagram

Intuitively we regard Venn diagram areas as positive, so information diagrams must be interpreted with great care. In some cases, there are both positive and negative contributions to an area corresponding to a multivariate mutual information, which may be net negative, net positive or net zero.

Example 2. Each bit is an independent, unbiased, binary random variable, except that 3 of the bits are in a mutual exclusive OR relationship. These 3 bits introduce one bit of synergy in modeling any one of the 3 vectors containing one of these 3 bits by the other 2 vectors. There is also redundancy among the 3 vectors as well as genuinely independent information. The equal quantities of synergetic and redundant information cancel (in this example exactly, but in real data perhaps approximately) in the calculation of the synergy-redundancy measure, giving the appearance of independence.

$$
\begin{aligned}
&\mathbf{x} = (b_{10}, b_9, b_8, b_7, b_6, b_5, b_3, b_2, b_1, b_0) \\
&\mathrm{H}(\mathbf{X}) = 10 \\
&\mathbf{y} = (b_{11}, b_8, b_7, b_6, b_5, b_4) \\
&\mathrm{H}(\mathbf{Y}) = 6 \\
&\mathbf{z}_j = (b_{10}, b_7, b_5, b_3, b_1) \\
&\mathbf{z}_k = (b_9, b_7, b_6, b_3, b_2) \\
&b_{11} = b_{10} \oplus b_9 \\
&\mathrm{H}(\mathbf{Z}_j) = \mathrm{H}(\mathbf{Z}_k) = 5 \\
&\mathrm{I}(\mathbf{Y};\mathbf{X}) = 5 \\
&\mathrm{I}(\mathbf{Y};\mathbf{Z}_j) = \mathrm{I}(\mathbf{Y};\mathbf{Z}_k) = 2 \\
&\mathrm{I}(\mathbf{Y};\mathbf{Z}_j,\mathbf{Z}_k) = 4 \\
&\mathrm{S}(\mathbf{Y};\mathbf{Z}_j;\mathbf{Z}_k) = 0
\end{aligned}
$$

Example 2 is illustrated in Figure 13.

In Figure 13, the tilde notation indicates that the bit is *not* contained in the vector corresponding to the surrounding H() region, but that a bit of entropy measure is assigned to this region, so the aggregated regions containing it that correspond to 2 and 3 term MI will have the correct total measure. Likewise the star notation on the bit with the negative sign in the central region of multivariate MI indicates that this one bit of negative entropy measure is needed to cancel each of the tilde marked bits in summing the 2 term MI in which each participates.

Difficulties using information diagrams for multivariate mutual information motivated introduction [52] of “information graphs” (used later in this dissertation). The zero valued synergy-redundancy measure, seeming to indicate independence in cases that actually contain both synergy and redundancy, is troubling: it may necessitate, in future work, introduction of more sophisticated indicators of multivariate dependence; e.g. “total absolute correlation” as the sum of the magnitudes of all multivariate mutual information terms of the power set.

The possibility of synergy is both a curse and a blessing. Building blocks cannot be identified if their fitness only becomes apparent when they are combined. However, when synergy between individuals in a population can be detected, it can be exploited to guide evolutionary steps: computing joint MI between the target and the constituents of an ensemble model enables evaluating that ensemble and estimating the potential of recombining its constituents.

Equipped with the information theoretic tools laid out in Chapter 1, we now proceed, in Chapter 2, to identify the objects to which these tools can be applied.

2 Channels Implicit in Evolutionary Learning

Information theory, in its primary application to communications engineering, generally considers transmission channels. Evolutionary computation and machine learning algorithms and the processes they are used to model are generally not considered in this context. However, it is possible to identify, in evolutionary learning, information sources, sinks and flows; the paths taken by flows to sinks from sources can be regarded as channels.

The main question that is typically asked about a channel is, what is its capacity (the maximum rate at which information can be transmitted over it with arbitrarily low probability of error)? In the context of evolutionary computation, this question has been addressed by only two researchers ([107] and [108][109]). The former has identified many issues involved in estimating evolutionary channel capacity but has not published any results. The latter has published estimates of the channel capacity of evolution in two special cases. There are several different channels that one might consider. Here we identify and briefly discuss each of the major channels implicit in evolutionary learning.

In this chapter, we identify channels implicit in evolutionary learning in 3 areas: Section 2.1, the environment; Section 2.2, the genome; and Section 2.3, from the environment to the genome.

2.1 Environment

A central problem in evolutionary learning is the automatic generation, maintenance and exploitation of probably approximately correct, hybrid symbolic/numeric models of the world, the self and other agents, for prediction, what-if analysis and control. We focus on developing theory and techniques for evolving models of a single observed process in the learning agent's environment. The observed process is the composition of a hidden process and a measurement process; each can be analyzed as a channel as follows.

2.1.1 Hidden Process

We are often interested in modeling a real-world process where we cannot directly observe the internals of the process, nor all its inputs and outputs. This hidden process may be represented conceptually as:

$$\mathbf{y}_{hidden} = \mathbf{f}_{hidden}(\mathbf{x}_{hidden}) \qquad \textbf{(18)}$$

This is purely conceptual, however: we can never know any of these three *noumena*, only their *phenomena*. Undeterred by this, however, we may reason about the corresponding information theoretic quantities H($\mathbf{Y}_{hidden}$), H($\mathbf{X}_{hidden}$) and I($\mathbf{Y}_{hidden}$;$\mathbf{X}_{hidden}$) and their relationships with those we can actually measure. These all apply to a more general case, where the relationship is not functional, but described by joint, marginal and conditional probabilities p(**x**,**y**), p(**x**), p(**y**), p(**x**|**y**) and p(**y**|**x**). The role of function $\mathbf{f}_{hidden}$ is in this case played by copula C_{hidden}() over the marginals to yield the joint distribution per Skar's Theorem [93] (the 'hidden' subscript dropped in the equation below for typographical clarity):

$$F_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y}) = C(F_{\mathbf{x}}(\mathbf{x}), F_{\mathbf{y}}(\mathbf{y})) \qquad \textbf{(19)}$$

2.1.2 Measurement Process

Our knowledge of the hidden process is limited to what can be inferred from measurements of its inputs and outputs. The measurement of these quantities can be expressed as

$$(\mathbf{x}_{measured}, \mathbf{y}_{measured}) = \mathbf{f}_{measure}(\mathbf{x}_{hidden}, \mathbf{y}_{hidden}) \qquad \textbf{(20)}$$

$\mathbf{f}_{hidden}$ and $\mathbf{f}_{measure}$ are unrelated: the former is a function between hidden input and output vectors, whereas the latter is a relationship between these hidden quantities and their measured values; we merely highlight the fact that both are processes through which information flows. Like $\mathbf{f}_{hidden}$, $\mathbf{f}_{measure}$ may really not be a function but rather a copula, but we retain the functional notation for clarity in illustrating concepts.

The measurement process typically will separate, at least approximately, as:

$$\begin{aligned} \mathbf{x} &= \mathbf{f}_{measureX}(\mathbf{x}_{hidden}) \\ \mathbf{y} &= \mathbf{f}_{measureY}(\mathbf{y}_{hidden}) \end{aligned} \qquad \textbf{(21)}$$

where we have dropped the subscripts on the measured quantities, as these will be our principal objects of study henceforth. If measurement is also deterministic, the Data Processing Inequality guarantees that:

$$\begin{aligned} &\mathrm{H}(\mathbf{Y}) \le \mathrm{H}(\mathbf{Y}_{hidden}) \\ &\mathrm{H}(\mathbf{X}) \le \mathrm{H}(\mathbf{X}_{hidden}) \\ &\mathrm{I}(\mathbf{Y};\mathbf{X}) \le \mathrm{I}(\mathbf{Y}_{hidden};\mathbf{X}_{hidden}) \end{aligned} \qquad \textbf{(22)}$$

However, if there is measurement noise, as will typically be the case, the first two inequalities may not hold. If measurement noise is correlated (e.g. when multiple sensors' readings are transmitted by parallel analog electrical transmission lines), the third inequality may not hold, and spurious relationships may be inferred. If there is “leakage” of information between $\mathbf{X}_{hidden}$ and $\mathbf{Y}$ (directly, not via $\mathbf{X}$) or between $\mathbf{Y}_{hidden}$ and $\mathbf{X}$ (directly, not via $\mathbf{Y}$), other inference problems may arise. Such difficulties can be quantified by generalizing our representation of the measurement relationships.

As in the case of the hidden process, the relationships may not be functional; additionally, in the case of the measurement process, they also may not be separable; so we may instead represent them with the joint probability $p(\mathbf{x}_{hidden},\mathbf{y}_{hidden},\mathbf{x},\mathbf{y})$, its various marginals and conditionals.

This leads us to consider various joint, marginal and conditional entropy and MI terms, as depicted in the information diagram of Figure 14 and detailed below.

- $H(\mathbf{X}_{hidden})$ the entropy of the hidden inputs
- $H(\mathbf{Y}_{hidden})$ the entropy of the hidden outputs
- $H(\mathbf{X}_{hidden}|\mathbf{Y}_{hidden})$ the conditional entropy of the hidden inputs given the hidden outputs
- $H(\mathbf{Y}_{hidden}|\mathbf{X}_{hidden})$ the conditional entropy of the hidden outputs given the hidden inputs

  These last two quantify the extent to which the hidden inputs and outputs do not fix each other: even with perfect measurement, it may not be possible to infer fully deterministic pre- or post-dictive rules, as the hidden relationship may be at least partially stochastic.
- $I(\mathbf{X}_{hidden};\mathbf{Y}_{hidden})$ the mutual information between the hidden inputs and hidden outputs

  This quantifies the strength of the hidden relationship and thus upper bounds the strength of any relationship that may be correctly inferred from measurements.
- $H(\mathbf{X})$ the entropy of the measured inputs
- $H(\mathbf{Y})$ the entropy of the measured outputs
- $H(\mathbf{X}_{hidden}|\mathbf{X})$ the conditional entropy of the hidden inputs given the measured inputs
- $H(\mathbf{Y}_{hidden}|\mathbf{Y})$ the conditional entropy of the hidden outputs given the measured outputs

  These last two quantify the extent to which the measurements do not fully reflect the variation in the hidden variables – [unintentional] filtering by the measurement process.
- $H(\mathbf{X}|\mathbf{X}_{hidden})$ the conditional entropy of the measured inputs given the hidden inputs
- $H(\mathbf{Y}|\mathbf{Y}_{hidden})$ the conditional entropy of the measured outputs given the hidden outputs

  These last two quantify [unintentional] noise injected by the measurement process
- $I(\mathbf{X}_{hidden};\mathbf{X})$ the mutual information between the hidden inputs and measured inputs
- $I(\mathbf{Y}_{hidden};\mathbf{Y})$ the mutual information between the hidden outputs and measured outputs

  These last two quantify the strengths of the relationships between the hidden variables and their measurements and thus upper bound any justifiably inferred entropies.
- $H(\mathbf{X}|\mathbf{Y})$ the conditional entropy of the measured inputs given the measured outputs
- $H(\mathbf{Y}|\mathbf{X})$ the conditional entropy of the measured outputs given the measured inputs

  These last two quantify the extent to which the measured inputs and outputs do not fix each other; they represent unavoidable error in estimates of each based on the other.
- $I(\mathbf{X};\mathbf{Y})$ the mutual information between the measured inputs and measured outputs

  This quantifies the strength of the measured relationship and thus upper bounds the strength of any relationship that may be justifiably inferred from the measurements.
- $I(\mathbf{X}_{hidden};\mathbf{Y}|\mathbf{X})$ conditional MI: hidden inputs & measured outputs given measured inputs
- $I(\mathbf{Y}_{hidden};\mathbf{X}|\mathbf{Y})$ conditional MI: hidden outputs & measured inputs given measured outputs

  These last two are examples of the many subtle conditionals, here information 'leakage'.

In some cases, the evolving learning agents may at least partially control the sensor design, transmission path between sensors and sensor signal processing system, and the plan for acquiring sensor readings (including “active learning”, e.g. [12]). In such cases, both the modeling process and the measurement process may be improved by evolution. Beyond merely listing the various conditional entropies etc., and pointing out this opportunity, we do not address the possibility of evolving the measurement processs in this dissertation.

2.1.3 Observed Process

We emphasize that the modeling process is concerned entirely with **x**, **y**, H(**X**), H(**Y**), I(**Y**;**X**) and the implicit function **f**: **X**->**Y** (or the more general relationship represented by the copula C()). While we hope to infer relationships among the hidden variables, the success of this enterprise is entirely dependent upon the fidelity of the measurement process; the modeling process can only infer relationships based upon the *observed* variables. The hidden function may be simpler or more complex than the observed one. Loss of information (i.e. intentional or unintentional filtering) may preclude inference of relationships that exist among the hidden variables. Correlated measurement noise may lead to inference of relationships that exist among the observed, but not the hidden, variables; if the correlated noise is large relative to the signals, these spurious relationships may dominate those we seek. To address these issues, we use the entropies of the previous section.

The MI between inputs and outputs may be regarded as the entropy of yet another set of Random Variables (RVs): those common to input and output sets.

Example 3: Each bit is equally likely zero or one, independent of the others:

$$\begin{aligned}
&\mathbf{x} = (b_7, b_6, b_5, b_4, b_3, b_2) \\
&\mathrm{H}(\mathbf{X}) = 6 \\
&\mathbf{y} = (b_5, b_4, b_3, b_2, b_1, b_0) \\
&\mathrm{H}(\mathbf{Y}) = 6 \\
&\mathbf{z} = (b_5, b_4, b_3, b_2) \\
&\mathrm{H}(\mathbf{Z}) = 4 = \mathrm{I}(\mathbf{Y};\mathbf{X})
\end{aligned} \qquad \mathbf{(23)}$$

This example cleanly partitions information into distinct sets of independent bits, some of which are exclusively in **X**, some exclusively in **Y** and some entirely in **Z**.

Example 4: Couple 2 unbiased binary RVs so as to agree ~89% of the time: each RV has 1 bit of entropy; they share half a bit of MI; but that shared half bit cannot be isolated from either of the RVs: in this case it is impossible, even in principle, to partition the entropy as in the preceding example.

Consider I(**Y***hidden*;**X***hidden*) between hidden inputs and outputs to be H(**Z***hidden*) of a nominal intersection of the hidden input and output variable sets. Likewise consider I(**Y**;**X**) between observed inputs and outputs to be H(**Z**) of a nominal intersection of the observed input and output variable sets.[10] We may use these new, synthetic entropies to identify yet another MI, that between the hidden and the observed input-output variable intersections, I(**Z***hidden*;**Z**). As those variable intersections contain all the information that exists regarding the relationships between the corresponding inputs and outputs, this MI quantifies how well either of those relationships models the other. As we cannot know any of the quantities pertaining to the hidden variables, we cannot actually compute this; it just

[10] This is not as arbitrary as it appears: the MI is the negative of the differential entropy of the copula density. The first direct published statement of this that we find is in [32]. An explicit treatment is given in [67]. See also Section 4.2 of this thesis.

introduces a way of quantifying the similarity of two relationships and demonstrates the existence of bounds on how much we can know about the hidden relationship. Having thus warned the reader, we will henceforth concern ourselves only with the observed process represented as:

$$\mathbf{y} = \widetilde{\mathbf{f}}(\mathbf{x}) \tag{24}$$

The tilde above the function name highlights: the hidden process may not be deterministic; and even for a deterministic hidden process, there is usually measurement noise in observed outputs, precluding their deterministic computation from observed inputs. Thus we content ourselves with a deterministic model that is ‘ideal’ in two important senses:

$$\begin{aligned} &\mathbf{z} = \mathbf{f}(\mathbf{x}) : \mathrm{H}(\mathbf{Z}) = \mathrm{I}(\mathbf{Z};\mathbf{X}) = \mathrm{I}(\mathbf{Z};\mathbf{Y}) = \max_{j}(\mathrm{I}(\mathbf{Z}_{j};\mathbf{Y})) = \mathrm{I}(\mathbf{X};\mathbf{Y}) \\ &p(\mathbf{y} \mid \mathbf{z}) = p(\mathbf{y} \mid \mathbf{x}) \\ &p(\mathbf{z} \mid \mathbf{y}) = \textstyle\sum_{k:\mathbf{z}=f(\mathbf{x}_k)} p(\mathbf{x}_k \mid \mathbf{y}) \end{aligned} \tag{25}$$

First, the MI between the ideal model outputs and the observed outputs is a maximum over all deterministic models based only on the observed inputs, equal to that between the observed outputs and inputs, which is the best that can be achieved due to the Data Processing Inequality. The ideal model leaves no residual information unexploited (we will later say that it is fully *sufficient*).

Second, among all models with the foregoing property, the ideal model entropy is a minimum, equal to the MI between the model outputs and the inputs, again equal to that between the observed outputs and inputs. The ideal model retains no excess input entropy; i.e. all its entropy is justified by its explanatory information (we will later say it is maximally *efficient*).

Z serves in the explanation of **Y** as a proxy for **X**, telling us all that **X** could. Likewise **Z** serves in the explanation of **X** as a proxy for **Y**, telling us all that **Y** could. **Z** is the ideal concise summary, not of **Y** or **X**, but of their relationship: the “information bottleneck” [103]; in the context of prediction the “causal states” [90].

- $\tilde{\mathbf{f}}$ (not really a function) denotes: filtering, from H(**X**), of H(**X**|**Y**) that tells us nothing about the *observed* outputs; a possibly nonlinear but nonetheless invertible coordinate transformation; and injection of H(**Y**|**X**) the entropy in observed outputs about which we can learn nothing from observed inputs.

- Its conceptual inverse $\tilde{\mathbf{f}}^{-1}$ (also not a function) denotes: filtering, from H(**Y**), of H(**Y**|**X**) that tells us nothing about the *observed* inputs; an invertible transformation; and injection of H(**X**|**Y**) the entropy in observed inputs about which we can learn nothing from observed outputs.

- **f** (a deterministic function) denotes: filtering, from H(**X**), of H(**X**|**Z**) that tells us nothing about the *model* outputs; an invertible transformation; and injection of H(**Z**|**X**) the entropy in model outputs about which we can learn nothing from observed inputs (a null operation, as **f** is deterministic).

- Its conceptual inverse $\mathbf{f}^{-1}$ (not really a function) denotes: filtering, from H(**Z**), of H(**Z**|**X**) that tells us nothing about the observed inputs (a null operation, as **f** is deterministic); an invertible transformation; and injection of H(**X**|**Z**) the entropy in observed inputs about which we can learn nothing from model outputs.

These ideal modeling information flows are illustrated in Figure 15. Alas, the ideal model **f** may not be realizable. It is easy to find examples where no deterministic function of **X** will preserve, in its outputs **Z**, all of the information I(**X**;**Y**), while filtering all of the surplus input entropy H(**X**|**Y**). Likewise $\mathbf{f}^{-1}$ may not be realizable: preserving all the information I(**X**;**Y**) may not be compatible with filtering all the excess entropy H(**Y**|**X**). Example 4 above is one simple example of this problem; another follows.

Example 5:

$$p(x=0,y=0)=p(x=1,y=0)=p(x=1,y=1)=\frac{1}{3}$$
$$\mathrm{H}(\mathbf{X})=\mathrm{H}(\mathbf{Y})\approx 0.92$$
$$\mathrm{H}(\mathbf{X},\mathbf{Y})\approx 1.59$$
$$\mathrm{I}(\mathbf{X};\mathbf{Y})\approx 0.25$$
$$\mathrm{H}(\mathbf{X}\,|\,\mathbf{Y})=\mathrm{H}(\mathbf{Y}\,|\,\mathbf{X})=\frac{2}{3}$$

In this simple case, no deterministic function of **X** will yield a **Z** with less entropy than **X** itself, while preserving all the information I(**Y**;**X**); neither will any deterministic function of **Y** yield a **Z** with less entropy than **Y** itself, again while preserving all the information I(**Y**;**X**).

Later, we will introduce various distance metrics and fitness indicators. We would like to be able to objectively measure how close a candidate model is to the best possible *realizable* model. Unfortunately, it is not apparent *a priori* how to assess how good a model can be realized; the indicators that we have developed measure instead how close a candidate model is to the *ideal* model, which may not be realizable.

Restricting ourselves to deterministic true functions, we can at best perform 2 of the 3 steps described for each of the "functions" described above: the filtering of the undesired conditional entropy; and the coordinate transformation. Injection of the desired conditional entropy is not computable: certainly we could generate [pseudo] random numbers with the same average entropy value, or even the same distribution; however, while this would yield the right statistical distribution, it would actually degrade the mutual information, compared with what we had after the 2 feasible steps.[11]

Illustrating this with Example 3: starting with **X**, **f** can filter out b_7 and b_6, but it cannot inject b_1 and b_0; at best, it can inject bits with the same conditional distribution given b_5, b_4, b_3 and b_2 (although in the given example, the distribution of b_1 and b_0 is not conditional upon them). Likewise, in the reverse direction, $\mathbf{f}^{-1}$ can filter out b_0 and b_1 but cannot inject b_6 and b_7. Thus these functions, although they play inverse roles, are not functional inverses.

So we introduce **g** to map from **Y** *towards* **X**, not via **Z** but rather via a newly introduced set of variables (or labels) **W**. Thus **f** and **g** are deterministically computable but incomplete transformations in the forward and reverse directions, yielding proxies **Z** for **X** in estimating **Y**, and **W** for **Y** in estimating **X**, respectively. In some cases, **Z** and **W** may prove to be identical (as in Example 3), or equivalent (the possibly nonlinear but nonetheless invertible transformation step),

[11] When modeling a process that generates a time series, one may be interested primarily in making accurate short term point predictions, or in learning a predictor that when iterated many times yields an attractor with long term characteristics similar to those of the attractor reconstructed by lag space embedding the observed time series. The latter objective may require injecting random variation with conditional entropy $H(\mathbf{Y}|\mathbf{X})$ according to conditional probability $p(\mathbf{y}|\mathbf{x})$, as in [87]; but doing so will degrade performance against the former (incompatible) objective.

but in general they will be completely different: each may be regarded as a set of labels of conditional probability distributions as in Equation 25. Figure 16 illustrates this scenario. Summarizing the important concepts from the above:

- $\mathbf{f}$ may have little to do with $\tilde{\mathbf{f}}$;
- $\mathbf{g}$ will almost certainly be unlike $\tilde{\mathbf{f}}^{-1}$;
- $\mathbf{f}$ may have even less to do with $\mathbf{f}_{hidden}$.
- nonetheless, it is an upper bound (not tight) on what a permissible (deterministically computable) model can achieve

Thus it will serve as our “ideal model” henceforth: a goal, better than which no model can do, and towards which we hope to direct an evolutionary algorithm.

2.2 Genome

Evolving models are evaluated on the basis of the behaviors (input-output mappings) of their phenotypes; however, they encode those behaviors in their genotypes. Individual genotypes and their aggregates (genomes of [sub]populations) thus contain information of critical interest to us. The flow of this information across generations can be measured using these same tools. The genetic operators of replication (without variation), mutation and recombination (crossover) can be analyzed as channels. In all cases, it is important to define carefully what information being transmitted through the channel is of interest: for instance, is our concern a precise sequence of base pairs (including introns) on a chromosome, or only the encoded protein that would be produced by gene expression (of the exons only)?

2.2.1 Replication

With respect to an individual genotype that is replicated without error, replication is a perfect transmission channel: it carries all the information from the sender (source, parental genotype) to the receiver (sink, offspring genotype); also it introduces no new entropy. If all genotypes in a population are thus replicated, then the average MI between symbol sequences (loosely, chromosomes) at the source and sink is exactly equal to their average entropy. In such a case, the replication channel capacity is exactly equal to the source information rate. If some genotypes are copied more or fewer times than others (have different fitness in the population biology sense), then the replication channel capacity is less than the source rate: some information about the relative frequency of different genotypes is lost between the parental and the offspring generations. However, this is usually an effect, not of reproduction itself (replication), but rather of reproductive selection, which will be considered in the context of the environment -> genome channel.

2.2.2 Mutation

A genotypic copy error is analogous to a communications error. The error incidence frequency is indicative of the extent to which the source information rate exceeds the channel capacity: a mutation rate of zero maximizes the mutation channel capacity; the more mutations occur, the lower is the mutation channel capacity relative to the total size of the population genome communicated through it. The improbability of beneficial mutations is clear in this light: we instinctively recognize the very low likelihood that a bit error in the channel will yield a received message of greater utility to the recipient than the transmitted message.

Genetic repair operators correspond to Forward Error Correction (FEC) in communications systems. Like FEC, they depend upon redundant encodings: explicit, with legal and illegal code-words and minimum distance decoding; or implicit, with valid and invalid string syntax, and syntactic repairs of minimum edit (e.g. Damerau–Levenshtein) distance.

Considering (rather than average distance between strings drawn from distributions) the distance between two specific strings, moves from Shannon entropy (H) based average Mutual Information to Kolmogorov complexity (K) based Mutual Algorithmic Information (MAI). The notion is conceptually similar, but with two inversely related key differences: the MAI is more universal and 'exact' than average MI, in that it characterizes the strength of the relationship between any two specific strings (rather than the average strength of the relationship between strings drawn from their known respective distributions); but MI is computable and MAI is not (it often can be bounded but the bounds are typically loose). MAI is defined as:

$$\mathrm{I}_{\mathrm{K}}(s_1;s_2)=\mathrm{K}(s_1)+\mathrm{K}(s_2)-\mathrm{K}(s_1,s_2) \qquad \textbf{(26)}$$

MAI measures the absolute proximity of two strings. It is complementary with their absolute distance (divergence, disparity), for which competing definitions exist [10][64]:

$$\begin{aligned} \mathrm{d}_{\mathrm{K}}(s_1,s_2) &= \mathrm{K}(s_1 \mid s_2) + \mathrm{K}(s_2 \mid s_1) \\ \mathrm{d}'_{\mathrm{K}}(s_1,s_2) &= \max(\mathrm{K}(s_1 \mid s_2), \mathrm{K}(s_2 \mid s_1)) \end{aligned} \qquad \textbf{(27)}$$

None of these absolute measures of proximity or distance are very useful in the present context, but normalized, they are potentially useful indices of similarity or dissimilarity. Perhaps surprisingly, the distance can be normalized such that it becomes a useful measure of relative dissimilarity, while retaining all the properties required of a distance metric, including in particular the triangle inequality (which is broken by several otherwise useful normalizations). Again there are competing definitions for Kolmogorov complexity based normalized information distance [10][64]:

$$\begin{aligned} \mathrm{D}_{\mathrm{K}}(s_1,s_2) &= \frac{\mathrm{K}(s_1 \mid s_2) + \mathrm{K}(s_2 \mid s_1)}{\mathrm{K}(s_1,s_2)} \\ \mathrm{D}'_{\mathrm{K}}(s_1,s_2) &= \frac{\max(\mathrm{K}(s_1 \mid s_2), \mathrm{K}(s_2 \mid s_1))}{\max(\mathrm{K}(s_1), \mathrm{K}(s_2))} \end{aligned} \qquad \textbf{(28)}$$

While the cited authors prefer the latter (primed notation) definitions of raw distance (Equation 27) and normalized distance (dissimilarity, Equation 28), we prefer the former (unprimed notation) definitions (in both equations):

- they directly correspond to Shannon entropy based raw and normalized phenotypic information distances that we will define in Chapter 3;
- the Shannon divergences are true distances that satisfy the triangle inequality *only* if we use the definitions parallel to the former ones; and
- reported empirical results (including [59]) using either definition show no significant difference in performance.

Here we have strayed from the communications engineering notion of a channel, where only Shannon entropy is used, not Kolmogorov complexity; but this departure is necessary, as we are interested not only in the distribution of different genotypes in a population considered as different messages (symbol sequences) but also in genetic operations on specific individuals (e.g., for Markov chain analysis). Also, distances between input and output strings (discussed here in a Kolmogorov sense) occur in communications theory (Shannon sense): e.g. the Hamming distance gives the number of bit errors in a message[12].

2.2.3 Recombination

We limit discussion here to Genetic Algorithm (GA) crossover with fixed length, two parents and two complementary offspring. Several pairwise dissimilarities may be compared (denoting parental strings s_{p1}, s_{p2} and offspring s_{o1}, s_{o2}):

$$\begin{gathered} \mathrm{D_K}(s_{p1}, s_{p2}) : \mathrm{D_K}(s_{o1}, s_{o2}) \\ \mathrm{D_K}(s_{p1}, s_{o1}) : \mathrm{D_K}(s_{p1}, s_{o2}) \\ \mathrm{D_K}(s_{p1}, s_{o1}) : \mathrm{D_K}(s_{p2}, s_{o1}) \end{gathered} \qquad \textbf{(29)}$$

These measures enable comparison of the inter-offspring and inter-parental similarities, which offspring is more similar to which parent, etc. If the recombination is geometric [71], as is standard bit-wise GA crossover, the distance between complementary offspring will be exactly equal to that between their parents. By merely concatenating strings, the distance between the parents taken jointly and the offspring taken jointly may be represented as

$$\mathrm{d_K}(s_{p1}s_{p2}, s_{o1}s_{o2}) \qquad \textbf{(30)}$$

[12] Dividing this by the message length equals the empirical error rate, which is an estimate of the error probability, on which a lower bound is given by a function of the excess of the source rate over the channel capacity.

This distance is maximized by a crossover under which each offspring inherits half its genetic material from each of its parents (half uniform crossover HUX): this corresponds to any of the possible 90 degree rotations in the search space of the vector connecting the parents around its midpoint to yield the vector connecting the offspring; such a crossover is maximally explorative of the subspace spanned by the parents.

From a Shannon perspective, an interesting question of interpretation arises: does maximizing such a distance minimize the recombination channel capacity (as the offspring are maximally different from their parents); or does it have no effect (as all the information contained in the parents was transmitted to the offspring without loss)?

The answer depends upon the precise definitions and the objectives of the analysis. If the genotypes are regarded as mere bit strings with no semantics, then such a maximally recombinant channel has capacity equal to the source rate, as no bits have been lost; they merely have been re-arranged, as by a network of switches or multiplexors. On the other hand, if semantics are considered, then such maximal crossover is almost certain to be highly disruptive of substrings (including schemata), in which case we would consider it to have introduced many communications symbol errors and reduced channel capacity well below the source rate. Note the replacement of the phrase “bit error” with “symbol error”: a meaningful substring corresponds to a multi-bit symbol.

Very briefly we consider communications network coding. If each node in a network coding flow graph represents a single individual (or sub-population) existing at any time during the evolution of the population, and information flow proceeds from parents (individuals or sub-populations) to offspring (likewise) over time, then clearly it is possible that individuals existing later in time may receive information that has been passed down to them in a distributed fashion, with some bits coming down one ancestral line and other bits down other lines, to be assembled in the descendants in ways that may have non-linear effects (e.g., 2 bits XORed in a communications net, or 2 recessive alleles appearing at the same locus in a diploid organism). This is the whole point of recombinative evolutionary algorithms. Whether the as yet immature mathematics of network coding can be applied usefully to analysis and design of evolutionary algorithms is an intriguing but so far unexplored possibility.

2.3 From Environment to Genome

In the canonical model of evolutionary algorithms, all information flow from the environment to the genome is transmitted through the mechanism of selection. This may not be true in all natural biological cases. Relative concentrations of various chemicals in the environment may favor the occurrence of particular mutations at the molecular level. Radiation may be more likely to break certain DNA bonds than others. Certain bacteria have higher mutation rates under environmental stress: whether this constitutes an information flow or merely an entropy flow is debatable. However, for our purposes, replication, mutation and recombination are purely genomic operators, not directly influenced by the environment; whereas reproductive and survival selection may be influenced strongly, and are influenced strictly, by the interaction of the phenotype with its environment.

2.3.1 Reproductive Selection

Selection, from the current population, of that subset to serve as parents of the next generation, is used in evolutionary algorithms more often than is survival selection; unfortunately it is the more difficult to analyze as a channel, as it is always immediately followed by the confounding effects of the reproductive channel. We will address survival selection first, then extrapolate what we may to reproductive selection.

2.3.2 Survival Selection

Survival selection directly transfers information from the environment to the genome of the population. “Fitness proportionate selection”[13] drives the frequency of incidence of a particular genotype to correspond roughly to the degree of success in that environment of phenotypes pursuing the particular survival and reproductive strategies encoded by that genotype. This can be viewed in terms of the amount of exergy (available energy) in the environment that is extracted from that environment and effectively utilized by those strategies, accounting for all reinforcement learning, survival and reproductive successes (e.g. finding food) and failures (e.g. being driven away from a potential mate by a competitor) in terms of increases and decreases in stored exergy

[13] This term is redundant under the population biologist’s definition of fitness, as pointed out by Altenberg: “fitness proportionate selection sounds like velocity proportionate speed, temperature proportionate heat” etc.; “fitness is the measure of selection”; see http://dynamics.org/~altenber/TALKS/CEC2000/

(embedded energy, emergy). Likewise it can be viewed in terms of the amount of information in the environment relevant to survival and reproduction that is extracted and effectively utilized by the species.[14] Note that selection does *not* transfer information to an individual genotype, as (absent Lamarckian learning or somatic line mutation) an individual's genotype is fixed for its lifetime and thus unable to encode any information received through the selection channel.

Example 6: Consider a Genetic Algorithm (GA) tackling a 5 bit, fully deceptive[15], symmetric trap function based on the OneMax problem, with both the usual "fitness function"[16] and a clairvoyant fitness function that counts the number of correct bits in an individual genotype. In the case of a canonical population, with roughly equal proportions of each of the 32 distinct possible genotypes, selection based on the clairvoyant fitness function removes the less fit members of the population: this increases the *average* fitness (proximity to correctness) of individuals in the population, but not the *total* fitness of the entire population (as entirely correct genotypes are already present);[17] neither does it increase the fitness of any individual members of the population; survival selection only deletes from what is already present.

[14] See especially [39][40]. Note that organisms do not run on energy, but rather *free* energy (availability, exergy, negentropy); see also various writings of Boltzmann, Lotka, Odum and Schrodinger.

[15] See Section 5.2 for more on an information theoretic view of deception and [43]a treatment of deception as one of the core sources of problem difficulty in evolutionary algorithms.

[16] See Section 5.2 for a more complete discussion of fitness in both the population biology sense (reproductive sampling rate) and the engineering sense (solution quality), between which we carefully distinguish there. Here we use the term "fitness" loosely, with the meaning generally intended when the term is used in evolutionary computation.

[17] Assuming that individuals are evaluated separately, as is usually done in Genetic Algorithms; of course, if ensembles are evaluated, using for instance voting (in this case corresponding to majority logic decoding), then the total fitness of the population *is* increased by selection based on our clairvoyant fitness function.

In this sense, selection alone introduces no new information into such a population (where all genotypes are already represented)[18], and neither can the genetic operators (crossover, mutation, etc.).

In a very small population, where *not* all possible genotypes are represented, selection alone still cannot introduce new information, but the genetic operators can. Selection accelerates the rate at which the genetic operators can do so by refining the source material from which they start. Selection then causes that new information to increase its relative incidence in the population.

With the usual fitness function (the count of bits set to one, except for the all ones string which has fitness zero and the all zeroes string which has the global maximum fitness of 5), the problem is fully deceptive: any hyperplane that divides the space of possible solutions in half, will yield an average fitness in the half containing the global maximum that is lower than the average fitness of the other half (which contains the global minimum). Thus selection based on the usual fitness function (constructed to demonstrate deception) will cause our clairvoyant fitness function to decrease. Our clairvoyant function is an information theoretic measure[19], which decreases because the usual fitness function used for selection did not agree with the amount of correct information in an individual genotype. This illustrates that information has always been implicit in the notion of deception, and encourages us to search for fitness functions that do reflect the amount of correct information in genotypes.

[18] In another sense, it does: the relative frequencies of different genotypes in the population are shaped to reflect their correspondence with the correct solution.

[19] Admittedly, one that cannot be computed unless we already know the solution to the problem.

Example 7: Consider a Genetic Programming (GP) approach to identifying a hidden Boolean circuit with 2 inputs *(x_1,x_2)* and 2 outputs *(y_1,y_2)*, with observables (other than the labeled outputs) including not only clean observations of the inputs but also noisy observations of the outputs (which our modeling system sees as merely additional unlabeled observables potentially useful in estimating outputs), and we try each observable as a model of each of the outputs.

$$\begin{aligned} y_1 &= x_1 \wedge x_2 \\ y_2 &= x_1 \vee x_2 \end{aligned} \qquad \textbf{(31)}$$

Simplistic candidate models are the observables (where the *w* variables are noise, not directly observable alone, corrupting the redundant observations of the outputs by flipping bits approximately 10% of the time):

$$\begin{aligned} z_1 &= x_1 \\ z_2 &= x_2 \\ z_3 &= x_3 = y_1 \otimes w_3 \\ z_4 &= x_4 = y_2 \otimes w_4 \end{aligned} \qquad \textbf{(32)}$$

Information theoretic fitness functions have certain significant advantages (as we shall see later), but they are still subject to a form of deception:

$$\begin{aligned} &\mathrm{I}(\mathrm{Y}_1;\mathrm{Z}_1) = \mathrm{I}(\mathrm{Y}_1;\mathrm{Z}_2) = \mathrm{I}(\mathrm{Y}_2;\mathrm{Z}_1) = \mathrm{I}(\mathrm{Y}_2;\mathrm{Z}_2) \approx .3 \\ &\mathrm{I}(\mathrm{Y}_1;\mathrm{Z}_4) = \mathrm{I}(\mathrm{Y}_2;\mathrm{Z}_3) \approx .05 \\ &\mathrm{I}(\mathrm{Y}_1;\mathrm{Z}_3) = \mathrm{I}(\mathrm{Y}_2;\mathrm{Z}_4) \approx .4 \end{aligned} \qquad \textbf{(33)}$$

Here each noise corrupted output observation has greater MI with its corresponding output than does either clean input observation; yet those noise corrupted output observations can never be processed in any manner to yield 100% correct outputs estimates, whereas the clean input observations can (if the hidden AND and OR functions are discovered by the GP system).

In this case, the deception could have been avoided by measuring the fitness not of individual expressions but rather of ensembles (a unique capability of information theoretic fitness functions that is one of their key advantages), but that is not always the case (as we also shall see later).

Chapter 2 having identified information transmission channels implicit in evolutionary learning, Chapter 3 defines information theoretic measures for evaluating ensemble models.

3 Information Theoretic Evaluations of Ensembles

This dissertation proposes a framework within which gains and losses of information can be addressed explicitly and objectively in evolutionary algorithm theory and practice, with three fundamental *desiderata*.

- The first *desideratum* is that the aggregate information in the population that “explains” the target should not decrease as evolution progresses. Information theoretic functionals can measure the population’s total useful information during an evolutionary algorithm run, to determine if indeed it is progressing rather than regressing.

- The second *desideratum* is that this useful information should concentrate progressively in fewer individuals, evolving smaller ensemble (perhaps ultimately individual) models that capture all or most of the target behavior. Information theoretic functionals can measure this concentration.

- The third and least obvious *desideratum* is that the information (in the population as a whole and in its best individuals) that does *not* contribute to modeling the target should decrease. Such excess entropy is not merely inefficient: it contributes to error; it exactly corresponds to all error remaining in the models, once all the target variance has been explained.[20] Information theoretic functionals can measure [reduction of] this excess entropy.

[20] Excess model entropy contributes just as does residual target entropy to error: in Figure 17 it does not matter whether the more complex waveform is a target with residual entropy (an unmodeled signal feature) or a model with excess entropy (ripple or ringing).

These heuristics can be applied analytically to the equations describing the population dynamics of evolutionary algorithms and empirically with numerical measurements during runs.

The first application of the framework is to analysis of evolutionary algorithms, which we began in Chapter 2.

The second application is to the identification and selection, from potentially large sets of observables, of those to be made available as genetic programming terminals. This often is overlooked, but it is not trivial: we must pick the right set, from the power set of the observables, which grows exponentially with their number. Information theoretic functionals can be used to quantify how much each observable, alone and in combination with others, reveals about the target. One potential heuristic would be to select for our terminal set the fewest observables that jointly tell us all about the target; another would be to select those that jointly tell us all about the target, with minimum joint entropy of the terminal data set. Application specific heuristics can also be defined, for instance to facilitate human interpretation of evolved models. Input selection is the focus of Section 6.3 and Chapter 7.

The third application is to the "main loop" of evolutionary computation: information theoretic functionals can provide justifiable, general, computable, commensurate indicators of fitness and diversity. This is described in Chapter 5 and Section 6.1. These indicators can used to apply selection pressure to survival, reproduction or both, of individuals, pairs or larger groups. The ability to measure the *joint* fitness of pairs and groups is novel: previously, the best that could be done was to compute statistics on the *individual* fitnesses of the members of pairs or groups. This relates directly to the next application.

The fourth application of the framework is to ensemble modeling. Recognizing that a population is an oversized ensemble model, we wish to reduce the size of the smallest ensemble required (to adequately model the target) that can be extracted from the population. Again information theoretic functionals can be used: they can measure how much an ensemble tells us about the target, without requiring that we know how to compose the constituents of the ensemble into a single more complex model. Whether the goal is a single or an ensemble model, the size of the smallest ensemble that explains a specified fraction of the target entropy, with no more than a specified limit on excess (non-explanatory) model entropy, tells us how far we are from that goal. This is the focus of Section 6.2.

Information theoretic indicators of fitness and diversity have been developed. These indicators have been implemented in *Mathematica* as described in Chapter 4 and partially validated with numerical experiments.

Why use mutual information to evaluate individual models?

Commonly used fitness measures, such as mean squared error, often fail to reward individuals whose presence in the population is necessary to explain substantial portions of the data variance. Diversity indicators are often arbitrary, may reflect diversity irrelevant to solving the problem, and are incommensurate with fitness measures.

By contract, information theoretic functionals are computable general indicators of fitness and diversity without these typical failings. The most obvious, and most commonly used, measures of individual fitness are mean (or sum) squared (or absolute) error. Without loss of generality, we will consider MSE.

Example 8. Given a target function

$$y = \mathrm{f}(x) = x \tag{34}$$

and a candidate model

$$z_j = \mathrm{f}_j(x) = x \tag{35}$$

MSE will be minimized (indeed, it will be zero).

Example 9. Keeping the same model, but estimating target function

$$y = \mathrm{f}(x) = -x \tag{36}$$

we now have a problem: MSE is enormous; yet our model is ‘close’ structurally to the correct answer (“edit distance” [74] is short). How can we reward this individual for its several desirable attributes? It contains the correct terminal *x*, raised to the correct power 1, and contains no spurious elements that fail to contribute to the solution or that contribute to error.

The next obvious choice is correlation. In Example 8 the correlation coefficient *r* is +1; in Example 9, it is -1, signifying that we have exactly the wrong answer, which guides us immediately to exactly the right answer (by considering not the signed correlation coefficient *r* but rather its absolute magnitude |*r*| or square r^2).

Example 10. Keeping the same candidate model, but now estimating target

$$y = \mathrm{f}(x) = x^2 \tag{37}$$

we again have a problem: correlation is exactly zero, seeming to signify that our model tells us nothing about the target. Yet it does: it still contains the correct terminal *x* (albeit raised to the wrong power) and appears to contain no spurious elements that fail to contribute to the solution or that contribute to error.

Using Mutual Information (MI) as our fitness indicator yields the desired behavior, as it generalizes correlation to arbitrary non-linear relationships. MI is invariant to invertible transformations[21] and degraded at most by a calculable and typically small amount for most reasonable non-invertible transformations[22], so it is effective at identifying high fitness simple forms in early generations [56] likely to be good building blocks for later generations [31].

Assuming 8 bit quantization[23] and a uniform distribution of test cases in *x*,

$$\begin{gathered} \mathrm{H}(\mathbf{Y}) = 7 \\ \mathrm{H}(\mathbf{Z}_j) = 8 \\ \mathrm{H}(\mathbf{Y},\mathbf{Z}_j) = 8 \\ \mathrm{I}(\mathbf{Y};\mathbf{Z}_j) = 7 \end{gathered} \tag{38}$$

which shows that our model captures all the entropy in the target. However, from this we can see that our model has excess entropy: the vestigial sign bit, which is information in the model that does not contribute to the solution, indeed it contributes to error.

[21] For discrete data, permutations; for continuous data, diffeomorphisms (embeddings) certainly and non-diffeomorphic homeomorphisms (which exist in dimensions exceeding 3), possibly; and in the terminology of Crutchfield (Section 3.1), recodings.

[22] The conditional entropy of the input to the transformation, given its output: H(X|X'); the log (base 2) of the number of pre-images.

[23] To work this example exactly, use either: 7 bit magnitude and 1 bit sign, adding 1/2 the value of the Least Significant Bit (LSB) to the absolute magnitude; or 8 bit twos complement with a +0.5 LSB signed offset.

This generalizes to more complex non-linear relationships. An alternative approach is non-linear correlation, but that begs the question, to what order must we go to capture the dependencies? MI avoids having to answer that question by leaping over the infinite regress to capture dependencies at *all* orders.

Why use mutual information to evaluate ensemble models?

The lesser reason is that the benefits MI brings to individual models accrue also to ensembles. The greater reason is that MI can be calculated, not only between 2 scalar Random Variables (RVs), but also between 2 vector valued RVs. Indeed it can be calculated among arbitrary numbers of arbitrary ensembles. This brings a novel and very important benefit: the ability to evaluate ensembles *as such*, without foreknowledge of how to assemble the multiple constituent sub-models into a single more complex model; and to do so objectively, without heuristics. When calculating MI of multiple items against a target, the items are regarded as joint observations, that can fix a more refined conditional probability density of the dependent variable than can any one of their constituent observations alone.

Various generalizations and normalizations of MI enable consistent evaluations of several important characteristics of individual and ensemble models.

In this Chapter, we introduce our selections and definitions for these: Section 3.1, proximity and distance; Section 3.2, dissimilarity, similarity and equivalence; Section 3.3, “sufficiency” and “efficiency” of models; Section 3.4, “necessity” and the Synergy-Redundancy Index (SRI); Section 3.5, model qualities considering inputs that may not fully determine outputs.

3.1 Proximity & Distance

Shannon's entropy H is a measure[24] of uncertainty in a Random Variable (RV) and his mutual information I is a measure of the *proximity* (closeness, nearness) of 2 RVs. This can be generalized if we consider them to be signed measures.[25] Proximity is a natural concept but its formalizations seem unnatural. A familiar complementary concept for which the mathematical formalizations are satisfying is *distance*. Given a set X, a function d: X×X→R satisfying

$$\mathrm{d}(x,y) \geq 0 \qquad \textbf{(39)}$$

$$\mathrm{d}(x,y) = \mathrm{d}(y,x) \qquad \textbf{(40)}$$

$$\mathrm{d}(x,x) = 0 \qquad \textbf{(41)}$$

is a distance. For it to be a *metric* we must also have

$$\mathrm{d}(x,y) = 0 \Leftrightarrow x = y \qquad \textbf{(42)}$$

$$\mathrm{d}(x,z) \leq \mathrm{d}(x,y) + \mathrm{d}(y,z) \qquad \textbf{(43)}$$

The set X and distance d comprise a *metric space* (X,d). Many partial, pseudo, quasi, semi, etc. metrics and distances have been defined in various fields and a substantial body of mathematics deals with metric spaces and their generalizations [36]. To exploit these results we must identify or define distances appropriate for analysis of evolutionary computation. As we have chosen the information theoretic path, we follow Crutchfield [30]: X is a space in which are located "recoding-equivalence classes" of information sources[26], between which the Shannon distance[27] $\mathrm{d_S}$ is the sum of the conditional entropies

$$\mathrm{d_S}(\mathbf{X},\mathbf{Y}) = \mathrm{H}(\mathbf{X} \mid \mathbf{Y}) + \mathrm{H}(\mathbf{Y} \mid \mathbf{X}) = \mathrm{H}(\mathbf{X},\mathbf{Y}) - \mathrm{I}(\mathbf{X};\mathbf{Y}) \qquad \textbf{(44)}$$

[24] In the formal mathematical sense of the word "measure", as shown in [30].

[25] http://en.wikipedia.org/wiki/Information_theory_and_measure_theory appears to be the only "publication" on this.

[26] Equivalent under information preserving transformations: discrete permutations or continuous homeomorphisms.

[27] a.k.a. information metric, entropy metric, Crutchfield-Renyi metric. Cf. Kolmogorov distance in Section 2.2.2.

This distance is the sum of the uncertainty in X not resolved by observing Y plus the uncertainty in Y not resolved by observing X, so it measures not only what additional information is required, starting from complete knowledge of X, to completely know Y, but also what information in X must be ‘forgotten’ (erased, deleted[28]) to know *only* Y. This formalizes concerns expressed in Section 2.1.3.

This distance has advantages for evolutionary computation: clearly, for error to be minimized, residual entropy of the target given the model, and excess entropy of the model given the target, which both contribute to error, must both be minimized; and before one can usefully optimize the form of a model, first one must ensure it incorporates all and only the needed inputs.[29]

Unfortunately, Shannon distance has no inherent sense of scale: is a distance of e.g. 2 bits large or small? This can be addressed by normalization as follows.

3.2 Dissimilarity, Similarity & Equivalence

The Shannon distance[30] can be normalized in several useful ways. The Rajski distance, a metric, is normalized by joint entropy

$$\mathrm{D_S}(\mathbf{X},\mathbf{Y}) = \frac{\mathrm{H}(\mathbf{X}\,|\,\mathbf{Y}) + \mathrm{H}(\mathbf{Y}\,|\,\mathbf{X})}{\mathrm{H}(\mathbf{X},\mathbf{Y})} = 1 - \frac{\mathrm{I}(\mathbf{X};\mathbf{Y})}{\mathrm{H}(\mathbf{X},\mathbf{Y})} \qquad \mathbf{(45)}$$

A proof that this normalized distance (or *dissimilarity*) is indeed still a metric distance may be found in [59] which also discusses generalization from discrete

[28] Important especially in the context of quantum or reversible computing or more generally thermodynamics of computation.

[29] Once in a metric space, we essayed still further, attempting to apply geometry, but this has been one of the less productive activities of the research reported here. The intersection of information geometry and evolutionary computation has thus far been small [105][48][47][68] and has pursued the differential geometry approach of Amari [7]with which that of Crutchfield [30]has not yet been reconciled. We suspect that one of Crutchfield’s spaces is a hypothesis subspace of one of Amari’s spaces, fixed by parameter selection. Moraglio [71]has attempted unification of EAs in a geometric framework; however his work mostly has been in the genotypic space, whereas ours is primarily in the phenotypic space, and while most would agree that search performance is always a function of neighborhood structure, the relationship between these spaces remains conjectural.

[30] Its components, the conditional entropies, are Albert quasi-metrics; see [36].

to continuous random variables. Note that the complement of this normalized distance (the rightmost quotient) is a similarity 'metric' [64]. Whereas dissimilarity is normalized distance, similarity is normalized proximity, yielding a satisfying formalization of the latter.

This dissimilarity provides a clear sense of scale: zero indicates equivalence with respect to information (maximum mutual dependence, like a correlation coefficient of magnitude one, but able to detect non-linear dependence of arbitrary order); and one indicates statistical independence (like a correlation coefficient of zero, but universally, rather than only for Gaussian variates).

A distance or dissimilarity of zero, or a similarity of one, indicates *equivalence*. Radcliffe's forma analysis [81], a very promising but largely unexplored approach to analysis of EAs in general and ensemble (multiset [82]) EAs in particular, requires use of a predicate indicating membership in an equivalence class with respect to a basis set of equivalence relations. Such a predicate φ can be defined as being false (0 valued) except when true as follows (but see the end of Section 5.5)

$$\varphi(\mathbf{X},\mathbf{Y}) = 1 \Leftrightarrow \mathrm{d_S}(\mathbf{X},\mathbf{Y}) = 0 \qquad \textbf{(46)}$$

$$\Leftrightarrow \mathrm{D_S}(\mathbf{X},\mathbf{Y}) = 0 \qquad \textbf{(47)}$$

$$\Leftrightarrow \frac{\mathrm{I}(\mathbf{X};\mathbf{Y})}{\mathrm{H}(\mathbf{X},\mathbf{Y})} = 1 \qquad \textbf{(48)}$$

Similarity is a good overall measure of how well a model approximates a target, but by collapsing all types of divergence of the model from the target into a single scalar value, it provides limited guidance to the evolutionary process. Decomposing it into meaningful components can provide more insight.

3.3 Sufficiency and Efficiency

With regard to information content of the target versus the model, there are two possible sources of error that an evaluator should penalize:

- residual entropy in the target
- excess entropy in the model

The former is variance in the target data set that is not explained by the model. The latter is variance in the model that does not explain anything about the target. Normalizing by target entropy yields the *sufficiency*:

$$\frac{\mathrm{I}(\mathbf{Y};\mathbf{Z}_j)}{\mathrm{H}(\mathbf{Y})} \qquad \textbf{(49)}$$

This measures the fraction of target entropy captured by the model, thereby penalizing residual target entropy. Normalizing by model entropy yields the *efficiency*:

$$\frac{\mathrm{I}(\mathbf{Y};\mathbf{Z}_j)}{\mathrm{H}(\mathbf{Z}_j)} \qquad \textbf{(50)}$$

This measures the fraction of the model entropy that contributes to its explanation of the target, thereby penalizing excess model entropy.[31,32]

[31] These same expressions (but not our subsequent renormalizations of them) were previously introduced as the *coefficients of constraint* [29] or *uncertainty coefficient* [78].

[32] In previously published work, we referred to the index (ratio) in Equation 49 as *sufficiency* and that in Equation 50 as necessity, in homage to Claude Elwood Shannon and William of Ockham. Subsequent discussions with early adopters of our methods suggested a change in nomenclature to make it agree better with current parlance: the index in Equation 50 we now call *efficiency* and we advance a new definition for necessity in Section 3.4. The names don't matter; the indices do.

3.4 Necessity and Synergy-Redundancy Index

We now define *necessity* as marginal sufficiency: the sufficiency of a larger ensemble containing both a particular sub-ensemble and other components, minus the sufficiency of the larger ensemble with the particular sub-ensemble removed. This shows the contribution that the particular sub-ensemble makes to the larger ensemble, showing how indispensable is the information contained in the particular sub-ensemble to the sufficiency of the larger ensemble.

We first introduce notation for ensembles. We use $\mathbf{f}^{\mathrm{m}}$ to represent an ensemble of *m* model functions and $\mathbf{Z}^{\mathrm{m}}$ to represent their outputs. Thus necessity of the sub-ensemble $\mathbf{f}^{\mathrm{m1}}$ in the context of larger ensemble $\mathbf{f}^{\mathrm{m}}$ (comprising $\mathbf{f}^{\mathrm{m1}}$ and $\mathbf{f}^{\mathrm{m2}}$) is

$$\frac{\mathrm{I}(\mathbf{Y};\mathbf{Z}^{m1},\mathbf{Z}^{m2})}{\mathrm{H}(\mathbf{Y})}-\frac{\mathrm{I}(\mathbf{Y};\mathbf{Z}^{m2})}{\mathrm{H}(\mathbf{Y})}=\frac{\mathrm{I}(\mathbf{Y};\mathbf{Z}^{m1},\mathbf{Z}^{m2})-\mathrm{I}(\mathbf{Y};\mathbf{Z}^{m2})}{\mathrm{H}(\mathbf{Y})} \qquad \textbf{(51)}$$

The left hand side of Equation 51 is just the difference between sufficiency of the larger ensemble and that of the particular sub-ensemble, using Equation 49. The right hand side recognizes that they are normalized by the same target entropy. Combining the right hand side of Equation 51 with Equation 12 (the synergy-redundancy measure) and normalizing not by the entire target entropy but by only that portion revealed by the mutual information of the larger ensemble with the target, yields a symmetric necessity, the Synergy-Redundancy Index[33]

$${}_{\mathrm{N}}\mathrm{S}(\mathbf{Y};\mathbf{Z}^{m1},\mathbf{Z}^{m2})=\frac{\mathrm{I}(\mathbf{Y};\mathbf{Z}^{m1},\mathbf{Z}^{m2})-(\mathrm{I}(\mathbf{Y};\mathbf{Z}^{m1})+\mathrm{I}(\mathbf{Y};\mathbf{Z}^{m2}))}{\mathrm{I}(\mathbf{Y};\mathbf{Z}^{m1},\mathbf{Z}^{m2})} \qquad \textbf{(52)}$$

[33] That could equally well be called the Interaction Information Index, but we find no such name in the literature, despite McGill's definition of "interaction information" several decades prior to Chechik's use of the "synergy-redundancy measure" to define a "synergy-redundancy index" normalized by the sum of the pairwise MI terms rather than the joint MI as we do here.

The algebraic sign of this index is informative –

- Negative indicates net redundancy in the information conveyed about the target **Y** by sub-ensembles f^{m1} and f^{m2}; there might also be synergy, but it must be of lesser magnitude than the redundancy.

- Positive indicates net synergy between those sub-ensembles; there might also be redundancy, but it must be of lesser magnitude than the synergy.

- Zero can be due to true independence, where redundancy and synergy are each zero; or to merely the appearance of independence, where redundancy and synergy of equal non-zero magnitudes cancel.

This index quantifies how much of the information conveyed by an ensemble about a target is due to interaction between its constituent sub-ensembles: that is, the degree to which the whole is greater or lesser than the sum of its parts.

3.5 Model Qualities

All the definitions in the preceding sections of this chapter are without regard to the system inputs. Sufficiency, for instance, is the mutual information between a model's outputs and the target system's outputs, normalized by the target system's outputs. This normalization is an attempt to show how well, relatively, a model is doing in its explanation of the target. However, this normalization is unfair. Deterministic models cannot convey more information about a system's outputs than do the model's inputs, which are (at most) the system's inputs. The Data Processing Inequality guarantees

$$\mathrm{I}(\mathbf{Y};\mathbf{Z}) \le \mathrm{I}(\mathbf{Y};\mathbf{X}) \le \min(\mathrm{H}(\mathbf{X}),\mathrm{H}(\mathbf{Y})) \qquad \textbf{(53)}$$

Thus, to show how well a model is doing, relative to the best that it could possibly do, we renormalize many (not all) of the quantities previously introduced.

efficiency	$\dfrac{\mathrm{I}(\mathbf{Y};\mathbf{Z}^{m})}{\mathrm{H}(\mathbf{Z}^{m})}$	**(54)**
sufficiency	$\dfrac{\mathrm{I}(\mathbf{Y};\mathbf{Z}^{m})}{\mathrm{I}(\mathbf{Y};\mathbf{X})}$	**(55)**
necessity of f^{m1} in context of f^{m} with f^{m2}	$\dfrac{\mathrm{I}(\mathbf{Y};\mathbf{Z}^{m1},\mathbf{Z}^{m2})-\mathrm{I}(\mathbf{Y};\mathbf{Z}^{m2})}{\mathrm{I}(\mathbf{Y};\mathbf{X})}$	**(56)**
dissimilarity of 2 ensemble models	$\dfrac{2\,\mathrm{I}(\mathbf{Y};\mathbf{Z}^{m1},\mathbf{Z}^{m2})-(\mathrm{I}(\mathbf{Y};\mathbf{Z}^{m1})+\mathrm{I}(\mathbf{Y};\mathbf{Z}^{m2}))}{2\,\mathrm{I}(\mathbf{Y};\mathbf{Z}^{m1},\mathbf{Z}^{m2})}$	**(57)**
dissimilarity to target of ensemble model	$\dfrac{\mathrm{I}(\mathbf{Y};\mathbf{X}\,\vert\,\mathbf{Z}^{m})+\mathrm{H}(\mathbf{Z}^{m}\,\vert\,\mathbf{Y})}{\mathrm{H}(\mathbf{Z}^{m})+\mathrm{I}(\mathbf{Y};\mathbf{X})-\mathrm{I}(\mathbf{Y};\mathbf{Z}^{m})}$	**(58)**
similarity to target of ensemble model	$\dfrac{\mathrm{I}(\mathbf{Y};\mathbf{Z}^{m})}{\mathrm{H}(\mathbf{Z}^{m})+\mathrm{I}(\mathbf{Y};\mathbf{X})-\mathrm{I}(\mathbf{Y};\mathbf{Z}^{m})}$	**(59)**

If one ensemble model is our ideal model (from Section 2.1.3, Equation 25) of the relationship between the system inputs and outputs, and the other ensemble has no excess entropy (or we modify Equation 58 to normalize not by total joint entropy but rather by joint MI only), the value of the expression in Equation 57 reduces to half of that in Equation 58. The equations are consistent but for this factor of 2, which was inserted into Equation 57 to deal with the possibility of synergy that otherwise could cause this index to exceed unity. Synergy is possible between two non-ideal models, but impossible between any model and the ideal model (which already has all the information about the outputs available from the inputs). If allowing an index to range in [0,2] is acceptable, the 2 may be omitted from the denominator in Equation 57, rendering it fully consistent with Equation 58; this is our preference for cleaner generalization in Section 5.1.1.

An expression similar to Equation 59 can be formulated, quantifying the similarity of two ensemble models, but this is rarely of interest: we seek similarity to targets of models, in which quest dissimilarity to targets of models may sometimes be a useful metric; but among models we seek only diversity, for which dissimilarity is our metric, as will be discussed in Chapter 5.

We may say that a model is “epsilon sufficient” if Equation 55 falls short of unity by no more than a specified epsilon magnitude; likewise for efficiency, etc. This can be used to relax the strict equivalence relations of Radcliffe (Equations 46 ff) and enables delta-epsilon treatment of generalization (Section 8.2).

Having defined information theoretic measures for evaluating ensemble models, in Chapter 3, we now proceed, in Chapter 4, to develop algorithms for estimating these quantities and the confidence associated with those estimates.

4 Estimation of Information Theoretic Quantities

To apply information theory in the practice of evolutionary computation, it is necessary to estimate, on sampled data, the numerical values of the quantities defined in the previous chapter. This process is likely to suffer from bias even with discrete data and has much worse potential problems with continuous data. Fortunately, we are rarely concerned with entropies as such, but rather with differences and ratios that can be arranged so that biases largely cancel.

In Section 4.1, we briefly outline calculations for discrete data that come directly from the definitions of entropy and mutual information. In Section 4.2, we introduce our novel copula based algorithm for continuous and mixed data. In Section 4.3, we justify significance testing with p-values from permutation tests. In Section 4.4, we handle small sample issues with confidence bounds from bootstrapping and introduce nested resampling. In Section 4.5 we address sweeping various parameters.

4.1 Discrete Case: Straightforward Entropy Calculations

Rather than using kernel based estimators that require (in our opinion generally unjustified) *a priori* assumptions about unknown distributions, we adopt the Maximum Likelihood Estimator. This generally underestimates entropies, overestimates low and underestimates high values of mutual information [76]. Unfortunately, this reduces the dynamic range and thus the selectivity of our indicators; fortunately, it is in a sense conservative, making us less likely to discard a weakly predictive variable.

Algorithms have been implemented in Wolfram Research, Inc. *Mathematica*, Version 8.0. A source code fragment of one of the basic functionals follows.

Source Code Fragment 1.

```
mutualInformation [ x__List, ";" , y__List ] := ( H[x] + H[y] ) - H[x,y] ;
```

H[x__List] in turn calls the built-in *Mathematica* function Entropy [<k>, <list>] that uses the empirical histogram to estimate the base *k* (in our work base 2 to give units of bits) entropy of an arbitrary list of objects: integers; category labels; even real (floating point) numbers. It may be appropriate to use this function on real numbers when a data set contains a small number of distinct low precision values, each of which is repeated a significant number of times; in other cases, where real values are high precision approximations of continuous data, so many distinct values are present and there are few repeats of each (repeats would be zero probability events if we had infinite precision), another approach is required.

4.2 Continuous Data: Novel Copula Based Algorithm

With discrete data, Shannon’s entropy is well defined. For continuous data, it is not, so analogs have been defined, including “differential entropy”, which retains some of the useful characteristics of discrete entropy, but does not serve our purposes. With discrete data, state space partitions are crisp; with noisy continuous data, they are fuzzy. Algorithms for computing information dimension of continuous RVs, or analogs of discrete entropy or mutual information for continuous RVs, typically involve adaptive bin sizes, complicated bookkeeping and heuristics. To address these difficulties, a new algorithm was developed, based upon copulae and sparse arrays.

The *copula* of a set of RVs describes how their joint distribution would look if all their marginal distributions were uniform. The composition of the copula over the marginal distributions reconstructs the full joint distribution. The copula thus captures all the statistical dependencies among the RVs: it reflects all conditional probabilities after discounting prior probabilities.[34] The divergence of the copula from a uniform multivariate distribution measures mutual dependence of the RVs.

Negating the differential entropy of the copula density yields the MI (redundancy) and the related *information dimension,* the number of independent components that be required to span the given multivariate data set (which will be less than the total number of variables $\mathcal{D}_m$ if there are dependencies).

As in [2], given the $\mathcal{D}_m$ dimensional cumulative distribution $\mathbf{F}(\mathbf{x})$, we may calculate its $\mathcal{D}_m$ marginals $\mathbf{F}_i(\mathbf{x}_i)$ and thence the multivariate copula density $\mathbf{dC}(\mathbf{x})$ corresponding to the copula $\mathbf{C}(\mathbf{x})$ by

$$\mathbf{dC}(\mathbf{x}) = \frac{\mathbf{dF}(\mathbf{x})}{\prod_{i=1}^{\mathcal{D}_m} \mathbf{dF}_i(\mathbf{x}_i)} \tag{60}$$

In [2], MI is estimated from a parametric copula that approximately models the data. Again, to avoid *a priori* assumptions about unknown distributions, we prefer the direct approach based on the empirical multidimensional histogram.

[34] A machine learner may be tasked only with learning the input-output relationship, in which case the copula is the whole story, or it may be tasked also with learning the marginal distributions: an observation of one RV that makes a 1000:1 longshot on another RV 10 times as likely still leaves a 100:1 longshot.

However, the significance of the copula lies in its non-uniformity and the data may have many observed dimensions: so in the histogram, many of the bins are likely to contain too few points for robust estimates of the joint and marginal densities; but if the data were transformed so that all the marginals were uniform, the multivariate copula density could be so estimated.

A straightforward approach is to sort the $\mathcal{D}_m$ dimensional data set $\mathcal{D}_m$ times (once on each variable) and construct the $\mathcal{D}_m$ dimensional co-occurrence matrix; each bin initially will contain exactly zero or one data points. However, with N data points, this requires an array with $N^{\mathcal{D}_m}$ elements. As multivariate dependencies can be inferred reliably only on large data sets, this would require infeasible storage capacity. Fortunately, there are only N non-zero elements, so sparse array techniques can be used, reducing memory requirements from $O(N^{\mathcal{D}_m})$ to $O(N \mathcal{D}_m)$.

Computational cost of subsequent pyramidal summation of the sparse array is of the same order as that of sorting the data $\mathcal{D}_m$ times to initialize it, $O(N\ lg(N)\ \mathcal{D}_m)$. Note that $\mathcal{D}_m$ must be less than $lg(N)$ or spurious dependencies will be detected[35], so overall cost is $O(N\ lg^2(N))$.

[35] This refers not to the entire data set, but rather to any subset of its columns from which we estimate joint entropies; as evolutionary computation often uses ordinal rankings rather than absolute magnitudes, e.g. to resolve selection tournaments, and indeed has been shown in many cases to benefit from the use of order statistics, we often can limit $\mathcal{D}$ to 4.

Algorithm 1.

```
function copula ( data: array [ N, d ] ): sparse array;
zeroize N^d dimensional sparse arrays
  dataBinCounts, binFrequencies, binFreqsPrevious;
lgNumberOfBins = ceiling( lg( N ));
numberOfBins = 2 ^ lgNumberOfBins;
for j = 1 to d do {
  sort data on column j;
  replace column j data values with their ordinal ranks };
for j = 1 to N do dataBinCounts[ data[j,1],...,data[j,d]] = 1;
while lgNumberOfBins >= minLgNumberOfBins and not done do {
  binFreqsPrevious = binFrequencies;
  binFrequencies = dataBinCounts / N;
  estInfoDim = entropy( binFrequencies ) / lgNumberOfBins;
  done = we just passed 1st max of estInfoDim or its slope;
  lgNumberOfBins--;
  numberOfBins = 2 ^ lgNumberOfBins;
  contract by 1/2 the length of each dimension
       of dataBinCounts by summing counts
       from merged blocks of adjacent cells };
if not done
  then { warn "suspect results"; return binFrequencies }
  else return binFreqsPrevious
```

We later modified the algorithm to construct the entire pyramid, all the way to the top level where the number of bins along each dimension is the minimum 2, rather than adaptively deciding when to halt[36], because the heuristic proved unreliable without information from all levels. Once the whole pyramid is built, a heuristic is still used to select a level for subsequent use: starting at the pyramid's peak, we select the first minimum of the slope of the estimated information dimension[37], or the lowest level at which the expected number of samples in each multidimensional bin exceeds one, whichever comes first.

[36] Obviously this increases computation and storage costs, but the impact is minimal due to pyramiding; it also simplifies the code.

[37] The estimated joint entropy divided by the log of the number of bins along each dimension estimates the information dimension.

On synthetic data sets to which we have applied the technique, where the dependencies are known, this usually yields a reasonable compromise between false negative and false positive detections of dependencies in the data. A similar heuristic still under investigation is to select the level that estimates the greatest relative redundancy: calculate a maximum possible estimated information dimension[38]; divide the actual currently estimated information dimension by that maximum; and select the level that minimizes that ratio.

Once a level is selected, it is used as a probability mass function, and entropies are calculated in the straightforward manner, directly from the definition of the Shannon entropy, as if we had discrete data. Dimensions corresponding to RVs not of interest in any particular calculation are collapsed by summation. The scale of the estimated entropies and other information theoretic quantities derived therefrom is arbitrary, so we normalize them by other estimates *from the same calculation*[39] to get indices (e.g. similarity, sufficiency). In turn we compare indices to yield relative ranks; the ranks are more stable than the values.

Per Theorem 1 in [76], if we neither suffer excessively from repeated values, nor encounter near-perfect dependence, our empirical copula will contain at least a few sample points in each of its bins, so entropies calculated from it will converge to unbiased estimates.

[38] The maximum possible estimated information dimension at a given level of the pyramid is the lesser of: the log of the number of data points divided by the log of the number of bins per dimension; and the sum of the estimated information dimensions of each RV at that level. Absent repeats, each RV should have an information dimension of one, due to our use of equal frequency bins (data ranks) rather than equal width bins (data values) in the histogram. This is equivalent to forcing the differential entropy of each individual RV to zero (assuming that each has the maximum possible entropy) while letting the differential entropy of the entire vector of RVs go negative (due to dependencies among RVs).

[39] More precisely, from the same estimated copula density: the same level of the same pyramid of bins. It only has to be the same calculation if we allow the algorithm to adaptively choose the level of the pyramid: it might choose differently for different variables.

This algorithm has also been converted to use almost entirely integer arithmetic. This reduces storage requirements in most implementations, as integers are represented typically with 2 or 4 bytes, whereas floating point representations consume from 4 to 8 bytes or more. This may or may not reduce computation time, as some hardware has floating point acceleration from which integer math may not benefit. This may facilitate parallelization of the algorithm, especially using Field Programmable Gate Array (FPGA) hardware. In any event, it does reduce accumulated round-off errors.

An intriguing possibility using the original floating point implementation is building the pyramid from a fuzzy co-occurrence matrix: where the initial atomic bins are populated, not exclusively with zeros and ones, but rather with fuzzy membership values, reflecting neighboring data point proximity in the original coordinates; this compromises between the “equal width bin” and “equal frequency bin” approaches to the histogram. Such an endeavor is justified by noting that the sorted ranks of near (in the original coordinate space) neighbors can easily be reversed by small (high probability) noise impulses adding to the originally lesser or subtracting from the originally greater coordinate value, but the ranks of more distant neighbors can only be affected by large (low probability) noise impulses. Investigation of this possibility is left for future work.

4.3 Significance Testing: p-Values from Permutation Tests

One problem with the use of information theoretic measures of dependence rather than linear correlation is that the former are more likely to yield spuriously high values under independence, because there are exponentially many ways to construct an arbitrary lookup table with perfect non-linear dependence but only two ways to construct a lookup table with perfect correlation or anti-correlation. This can be seen easily by examples.

Example 11. Assume we have 2 unbiased binary RVs. The smallest sample that can show either independence, perfect dependence or perfect anti-dependence is 2^2=4 points; there are 4^4=256 possible joint sequences at this minimum sample size. Under the null hypothesis of independence, each of these sequences is equally likely. In 60 of these (23%), one or both of the RVs will fail to display both its values, making it impossible to assess dependence (whether with information theoretic methods or with correlation). In 24 cases (9%), the 'correct' (under the null hypothesis) similarity value of 0 will be estimated. In 48 cases (19%), an incorrect estimate of weak dependence (similarity .08) will be estimated; in another 96 cases (38%), an incorrect estimate of moderate dependence (similarity .21) will be estimated; and in 28 cases (11%), an incorrect estimate of perfect dependence (similarity 1.0) will be estimated. Even the least likely and most severely wrong answer is more likely than the right answer!

Example 12. Assume we have 2 continuous RVs (with infinite precision representations so that repeated values have zero probability). Again the smallest sample that can show a difference between independence and dependence of these RVs is 4 points. If we use not equal width but rather equal frequency bins (as in our Algorithm 1), then unlike the discrete case in Example 11 above, our sample is forced to be balanced: there are $(4!)^2$ possible temporal sequences, corresponding to 4! different arrangements with respect to the relative ordinal ranks of the 2 RVs, yielding only 3 distinct histograms; this corresponds to restricting attention to the 36 balanced of the 256 possible sequences in the discrete case (the previous example). 'Correct' (under the null hypothesis) estimation of independence occurs in 2/3 of these sequences; incorrect estimation of perfect dependence (or anti-dependence, either way a similarity of 1.0) occurs in 1/3 of them: better than the discrete case, but a wrong answer is still half as likely as the right answer. See Tables 1.1, 1.2 and 1.3.

In both the foregoing examples, the minimum number of bins per dimension and minimum number of points (for the total number of bins) makes spurious estimates of strong dependence likely, whether using information theoretic methods or linear correlation. As the number of dimensions increases, finding information theoretic dependence of one RV upon a non-linear combination of multiple other RVs becomes possible (and likely, even under independence); this is not possible with linear correlation. Even remaining at only 2 dimensions, as the number of bins per dimension increases, there is a corresponding factorial increase in the number of arbitrary lookup tables that [spuriously, under independence] show perfect dependence using the information theoretic measures, whereas the number of lookup tables that [spuriously] show perfect dependence (or anti-dependence) using linear correlation remains only 2.

Thus there is increasing “trigger-happiness” of the information theoretic techniques versus linear correlation as problems scale up. Information theoretic dependence statistics reveal no more than ordinary correlation with only 2 bins per each of 2 dimensions: it is only with larger numbers of bins that they reveal anything that we did not already know; yet with larger numbers of bins they become increasingly suspect. The first step in scaling up, from 2x2 to 3x3, is shown in the following example.

Example 13. Extending either discrete Example 11 or continuous Example 12 to 3 bins per each of 2 dimensions: under independence, 6 histograms spuriously show perfect information theoretic dependence, whereas only 2 of those also show spuriously perfect linear [anti] correlation. The number of sequences that yield each histogram is large and depends upon whether it is the discrete or continuous case; it has not been calculated here. See Tables 2.1 through 2.6.

To combat (or at least quantify) this problem, tests of statistical significance are used. Often these tests are parametric and thus based upon *a priori* assumptions about underlying (and often not truly known) distributions. In keeping with our general philosophy that we know nothing that we cannot justifiably infer exclusively from the given data sample, we instead adopt a non-parametric approach based on resampling without replacement.

Randomly permuting each of the RVs, the marginal distributions are unchanged, but the joint distribution is destroyed: replaced with some large number of random synthetic joint distributions. This enables a distribution of a statistic under the null hypothesis to be calculated (if each possible permutation is sampled exactly once – an exact test) or estimated (if a large number of random resamples that does not exhaust an even larger space of permutations is used – an approximate test). This distribution can be used in various ways; we use it to calculate or estimate a *p-value*. The interpretation of a p-value is subtle: the probability that, under the null hypothesis (independence), a statistic will be estimated as having a value at least as extreme as the value actually estimated prior to permuting the RVs.

In our work, we have not advanced the science of resampling, permutation tests, p-values, etc.; we have merely implemented and applied such techniques, after justifying the greater need for using them to assess the statistical significance of dependency strength estimates from information theoretic methods versus from linear correlation.

Source Code Fragment 2. f_ is a function that calculates a dependency strength statistic on x_ and y_. fPValue returns an approximate p-value of f_ on a sampleSize_ number of permutations of x_ and y_. Alternatively, if the length of x_ and y_ were short enough, the permutation space could be exhaustively enumerated to yield an exact p-value.

```
fSample [f_, x_List, y_List, sampleSize_Integer:741 ] :=
        Table[
                f[ RandomSample[ x ],RandomSample[ y ] ],
                {sampleSize}];

fPValue [f_, x_List, y_List, sampleSize_Integer: 741 ] :=
        Module[ {tmp},
                tmp = DeleteCases[ N[ fSample[ f, x, y, sampleSize ] ], Indeterminate ];
                N[ 1 - (Position[ Ordering[ Prepend[ tmp, N[ f[ x, y ] ] ] ],1 ] [[1,1]] - 1)
                        / Length[ tmp ] ] ];
```

4.4 Small Samples: Confidence Bounds from Bootstrapping

For error not to be dominated by bias due to small sample size, the number of data points *N* must greatly exceed the total number of bins $m^{\mathcal{D}}$, where *m* is the number of bins along each dimension and $\mathcal{D}$ is the number of dimensions.

In a selection tournament involving individual RVs, the number of dimensions typically can be limited to 4. When considering not individual RVs but rather ensembles, even in a tournament involving only 3 or 4 ensembles, $\mathcal{D}$ can grow large, as it is the total number of RVs contained in all the ensembles in the tournament. Thus achieving $N >> m^{\mathcal{D}}$ may become infeasible.

One approach to estimating sample bias/variance effects and establishing confidence bounds around calculated statistics is *bootstrapping*, a resampling technique superficially similar to the permutation tests of the previous section. A permutation test randomly resamples *marginal distributions without replacement*, bootstrapping randomly resamples the *joint distribution with replacement*.

Again in our work, we have not advanced the practice of bootstrapping as such; we have merely implemented and applied it, as illustrated in the source code fragment below, after justifying the need to do so.

Source Code Fragment 3. f_ is a function that calculates a dependency strength statistic on x_ and y_. fBootBounds returns two-sided boundFrac_ confidence bounds on f_ on a sampleSize_ number of resamples of { x_, y_ }.

```
fBootStrap[ f_, x_List, y_List, sampleSize_Integer: 741 ] :=
        Table[
                Apply[ f, Transpose[ RandomChoice[ #, Length[ # ] ]&[ listJoin[ x, y ] ] ] ],
                {sampleSize} ];

fBootBounds[ f_, x_List, y_List, boundFrac_Real: 0.10, sampleSize_Integer: 741 ] :=
        Module[ {tmp,pos},
                tmp = Sort[ DeleteCases[ N[ fBootStrap[ f, x, y, sampleSize ] ], Indeterminate]];
                pos = Max[ 1, Floor[ boundFrac * Length[ tmp ] ] ];
                Extract[ Sort[ N[ fBootStrap[ f, x, y, sampleSize ] ] ], { { pos }, { -pos } } ] ];
```

We have made one novel application of the technique – bootstrapping not only the estimation of the statistic (e.g. sufficiency index), but also estimation of its corresponding p-value. In an outer loop, we bootstrap the rows of the sample matrix to obtain a large number of different synthetic random sample matrices and estimate the information theoretic dependence strength statistic on each. In an inner loop, we permute each column of the given (bootstrap synthesized) sample matrix to obtain a distribution on the statistic under independence (and from this distribution a p-value of the dependence strength statistic). The following source code fragment implements this novel nested resampling.

Source Code Fragment 4. f_ is a function that calculates a dependency strength statistic on x_ and y_. fBBPV returns two-sided BBFrac_ confidence bounds on an approximate p-value of f_ on a PVSampleSize_ number of permutations of the marginals of a BBSampleSize_ number of resamples of { x_, y_ }.

```
fBBPV[
      f_, x_List, y_List,
      BBFrac_Real: 0.10, BBSampleSize_Integer: 102,
      PVSampleSize_Integer: 102 ] :=
          fBootBounds[
                    fPValue[ f, #1, #2, PVSampleSize ]&,
                    x, y, BBFrac, BBSampleSize ];
```

Nested resampling, especially of information theoretic functionals, is computationally very costly. However, on various small sample data sets to which we have applied these indices (see especially Section 7.2), it appears to be necessary to assess confidence in the dependency strength estimates.[40]

[40] On-line searches have not yielded any published algorithms or analysis similar to this approach. Preliminary discussions with Prof. Kishan Mehrotra, who is familiar with the statistics literature, suggest that this nesting of bootstrapping and permutation testing appears to be novel.

4.5 Order, Scale and Lag

The initial application of time series prediction can be approached using at least 3 different representations: a lag vector of values of the observable; a vector comprising smoothed values of the observable and its first few differences for constructing a single higher order differential equation; or a vector of [arbitrary presumed] state variables for constructing a system of first order differential equations.[41] In all approaches, the length of the vector must be chosen, whether manually by the user, or automatically by the modeling algorithm.

In the case of a lag vector, its minimum length is given by the Takens embedding theorem as the least integer greater than twice the fractal dimension of the attractor that produced the time series. In the case of a vector of successive derivative estimates, it is the order of the differential equation, which will typically yield a vector shorter than the lag vector representation, but only for systems that can be so represented. In the case of a vector of state variables, it is the number of degrees of freedom of the system. If what we will generally refer to as the *order* of the system is *a priori* unknown, it will be necessary to scan the data at successively increasing orders until sufficient predictive information is found or concluded absent from the data set (or at least not feasibly extractable).

[41] These state variables must be computed from the observables. If definitions of the "true" state variables are not known, it may be possible to construct an over-complete vector of surrogate state variables using the methods of "compressive sensing". Essentially a lag vector would be multiplied by a random matrix of zeroes and ones where the rows are linearly independent and "random".

The incidence of chaos, with increasing dimension, increases in continuous flows (starting at a minimum dimension of 3) but decreases in iterated discrete maps; the trends intersect at a dimension of 8 [96]. Generic modeling faces severe challenges implied by the Thom Classification Theorem if the dimension exceeds 10. The number of data samples required for meaningful statistics on non-linear multivariate dependence, and the computational time to process them, both increase exponentially with dimension. For systems that can be represented as a single higher order ODE, a lag vector length of 8 is particularly convenient, as it can be processed with a computationally inexpensive Walsh transform to estimate the observable and its first few derivatives; a lag vector of length 8 is adequate for systems with dimensions up to 4. Thus for interesting practical problems, orders of 3 to 4 should be a fertile place for investigation. System order can be estimated as the least integer exceeding the calculated information dimension $\mathcal{D}_I$: but this may include noise as an additional dimension; and will depend upon the scale (level of the pyramid) at which $\mathcal{D}_I$ was calculated.

The appropriate *scale* at which to calculate the information dimension (and the entropy, mutual information, etc.) is in general also *a priori* unknown. Our novel copula based algorithm has been used with a heuristic described above to select a pyramid level, but this heuristic is somewhat arbitrary and the algorithm itself is agnostic on this point. It also should be noted that there may not be a single most appropriate scale: in a multivariate data set, it is entirely possible to have coupling at multiple scales; e.g. at scale 2^j variables x_1 and x_2 may be strongly coupled, while at scale 2^k variables x_2 and x_3 may be strongly coupled.

Finally, related to both order and scale, the *lag* at which to look for dependence is in general also *a priori* unknown. By lag is meant the ΔT between successive elements of a lag vector in the case of uniform sampling, the ΔT between terms in the case of delay differential equations, etc. In the experiments reported in Section 6.1, the lag step size was selected to be that which yielded the first minimum of the mutual information, as the synthetic data could be generated at arbitrary step sizes. In the application reported in Section 7.1, as the real world data had been sampled at a fixed interval, dependence was plotted as a function of lag, and the lag times that maximized and minimized it for particular pairs of variables were among the findings of interest to the plant engineers.

All the statistics used herein are averages over the entire data set. It is possible to calculate Pointwise Mutual Information (PMI) over subsets of the data and at various lags and scales. This may be helpful when the dependence structure is complex or spatiotemporally varying. Exploring this is left for future work.

Having developed, in Chapter 4, algorithms for estimating information theoretic statistics, we now proceed, in Chapter 5, to illustrate their application to analysis of evolutionary algorithms.

5 Evolutionary Algorithm Analysis

We will focus upon only a few first order considerations, and view them in the new (to the field of evolutionary computation) light of information theory. Most analysis starts with *fitness*: the topic is familiar and the notion is seemingly intuitive; but the term is almost universally misused, perhaps belying the actual depth to which the notion is understood. We will start instead in Section 5.1 with *diversity*: again a familiar topic, but one where we are less likely to be led astray by untrustworthy intuition. After considering diversity, endowed with whatever new perspectives that may give us, we will turn in Section 5.2 to fitness and the closely related topic of *deception* (generally considered the primary source of problem difficulty for EAs, or at least of "intra building block difficulty" [43]).

Next we revisit one of the few quantitative theoretical results from natural biology, specifically population genetics, that has been applied to EAs: George Price's Selection and Covariance Theorem [79]. The Price Equation relates the correlation between trait value and fitness (reproductive sampling rate) with the changing preponderance of that trait in a population. Price originally assumed individual selection and non-random mating: but equipped with the novel ability, provided by our information theoretic metrics, to objectively measure the fitness of ensembles, we can predict population genetic dynamics under pair or other group selection. In Section 5.3, we extend his equation to *non-random mating*. In Section 5.4, we re-derive his original equation with a new interpretation for *effective fitness*. In Section 5.5, we reformulate his equation to identify terms quantifying fundamentally important attributes such as *evolvability*.

5.1 Diversity

Diversity is the label for several related attributes of a population among which the EA community has failed to distinguish with different labels. The natural biology population genetics community asks more specific questions, such as: How many loci have multiple alleles present in the population? What fraction of the total number of loci is this? How many different alleles are present: for each locus; and total or average? What is the statistical distribution of alleles? The value of diversity in EAs derives from several factors, including the following:

- To share limited resources, species may exploit different niches (e.g. multiple cognitive radio networks may select different spectral bands).
- Environments may be non-stationary, so alleles that are sub-optimal during one time period may become optimal during a later period (and then still later become sub-optimal again, in cycles or random walks).
- An evolutionary optimization algorithm may not have converged to a good solution and “building blocks” required to construct an as-yet-unknown good solution may be present only in species whose instances are not the most fit individuals in the current population.

Several researchers in natural biological ecosystems and EAs have used entropy as a diversity measure [110]. Various refinements of entropy may be used to measure various specific kinds of diversity. We may consider the diversity of entire populations, sub-populations, or specific individuals with respect to others. We may consider genotypes, phenotypic structures or phenotypic behaviors.

5.1.1 Phenotypic diversity of populations

In this dissertation, rather than phenotypic *structural* diversity, we address the more directly useful (for evolutionary learning of models) phenotypic *behavioral* diversity. Total information diversity, in entire populations, sub-populations, ensemble models, potential reproductive pairings, etc. is inversely related to redundancy (Equation 11, total correlation), which can be normalized as:

$$ {}_N C(Z_1, Z_2 \ldots Z_m) = \frac{C(Z_1, Z_2 \ldots Z_m)}{\sum_{i=1}^{m} H(Z_i) - \max(H(Z_1), H(Z_2) \ldots H((Z_m))} \qquad \textbf{(61)} $$

If the normalized total correlation is one and each member of a population has the same *amount* of information (entropy), then each has *exactly the same information*.[42] With a normalized total correlation of one and unequal model entropies, the model with the greatest entropy has all the information, of which each of the others has a subset. From an information theoretic perspective, all maximal entropy models in such a population are equivalent; all but one could be discarded, along with all models of lesser entropy, with no information loss. However, different individuals might represent the same information in very different ways, yielding very different behavior from a RMSE perspective.

A total correlation of zero indicates that each member of a population has information entirely distinct from that of every other member and combination of members. For every pair or larger group of individuals, the mutual information is zero. This apparent diversity is maximized by a population of random models, each conveying no information about any of the others, *or about the target!*

[42] Under, as always, a recoding (discrete permutation or continuous homeomorphism); in other words, the models are all in the same "recoding-equivalence class"; Crutchfield, *op cit*.

For ${}_{N}C()=0$ given multiple individuals that convey information about the target, each must model a different target aspect. Normalized total correlation can be minimized absent such unlikely non-overlap of the mutual information between models and a target: excess model entropy can mask a paucity of diversity in the information that different individuals convey about the target. Instead we could calculate “relevant diversity” (as we did formerly) as

$$\frac{I(Y;Z_1 \ldots Z_m)}{\sum\limits_{i\in[1..m]} I(Y;Z_i)} \qquad (62)$$

which ranges from zero to one *in the absence of synergy*. In the presence of synergy, it is unbounded; indeed, if all the information about the target results from synergy, it involves division by zero. To avoid this problem and to capture the effects of synergy, we generalize the definition of total correlation, replacing entropy with mutual information to get *model absolute total redundancy*

$$R(Y;Z_1,Z_2 \ldots Z_m) = \sum_{i=1}^{m} I(Y;Z_i) - I(Y;Z_1,Z_2 \ldots Z_m) \qquad (63)$$

and normalizing to get *model relative total redundancy*

$${}_{N}R(Y;Z_1,Z_2 \ldots Z_m) = \frac{\sum_{i=1}^{m} I(Y;Z_i) - I(Y;Z_1,Z_2 \ldots Z_m)}{I(Y;Z_1,Z_2 \ldots Z_m)} \qquad (64)$$

These are the additive inverses of the synergy-redundancy measure (Equation 12) and Synergy-Redundancy Index (SRI, Equation 52), generalized from 2 to an arbitrary number of models (as was implicit in the ensemble formulation of Equation 52). As we wish to measure not the inverse of diversity (redundancy) but rather diversity itself, we settle upon

$$ {}_{N}S(Y;Z_1,Z_2\ldots Z_m)=\frac{I(Y;Z_1,Z_2\ldots Z_m)-\sum_{i=1}^{m} I(Y;Z_i)}{I(Y;Z_1,Z_2\ldots Z_m)} \qquad (65) $$

Note that generalized SRI now ranges in [-(m-1),+1].

- A value of zero indicates that the information about the target provided by each of the models in the population is *on average* neither redundant nor synergetic with (i.e., *seemingly* independent of) the rest of the population.
- A value of negative (m-1) indicates that all the models are entirely redundant (not in their total entropy, but in the information they convey about the target).
- A value of positive one indicates that no individual model in the population conveys any information about the target, but that collectively they do convey information: this is not as unlikely as it seems (consider an ensemble comprising all the inputs to a parity circuit); the size and composition of the ensemble that conveys this information is not indicated.

Generally speaking, the more positive (more likely, less negative) the SRI, the more phenotypic (behavioral, target relevant) information diversity.

If an index that ranges in [0,1] is preferred, we may renormalize Equation 65 (as we did Equation 52 to get Equation 57) as

$$ \frac{m\,I(Y;Z_1,Z_2\ldots Z_m)-\sum_{i=1}^{m} I(Y;Z_i)}{m\,I(Y;Z_1,Z_2\ldots Z_m)} \qquad (66) $$

but this sacrifices the clear boundary between net redundant and net synergetic cases so our preference is for Equation 65.

We may foreshadow discussion of fitness by noting that the joint MI term in the numerator and denominator of should rapidly converge to I(**Y**;**X**), so variation will come only from the summation in the numerator, which equals the number of models in the population times their average MI with the target. This relates diversity with fitness (a connection lacking from most previous indicators) but desensitizes this diversity indicator as the run progresses.

Given the above noted problems, especially the confounding effects of redundancy and synergy, development of a fully satisfactory measure of diversity remains open. Nonetheless, *total and target relevant phenotypic behavioral information diversity indicators are key contributions of this thesis*.

5.1.2 Phenotypic diversity of one entity with respect to another

Assessing diversity of one entity (individual or ensemble) with respect to another can be accomplished using various indicators. For total information we may use dissimilarity (normalized information distance, Equation 45). For target relevant information we may use any of the following: necessity with respect to the target (Equation 51); necessity with respect to what the inputs reveal about the target (Equation 56); SRI (Equation 52); or the dissimilarity of Equation 57 (SRI shifted to be always positive and scaled to lie in [0,1]). As both versions of SRI are symmetric they are generally preferable to either version of necessity; but none of these satisfy the triangle inequality so unfortunately we lack a true distance. We prefer to assess model diversity using SRI from Equation 52.

5.1.3 Genotypic diversity

Genotypic representation diversity may be measured by replacing, in the above phenotypic behavioral diversity indicators, the MI with the MAI (Equation 26), Shannon entropy H with Kolmogorov complexity K, D_S with D_K (Equation 28), etc. and applying the indicators to genotypic strings rather than model outputs. Distances should use the sum rather than the maximum of the conditional complexities and be normalized by the joint rather than the maximum of the complexities, as discussed in Chapter 2, for consistency between the phenotypic behavioral and genotypic representation diversity indicators; this is not merely aesthetically appealing but also important to the conjecture stated in Section 8.1. Theoretical analyses can use Kolmogorov complexity based measures; actual measurements and calculations must replace those non-computable quantities with appropriate surrogates, such as Normalized Compression Distance (NCD).[28] EAs should preserve diverse genetic material with reproductive potential.

5.2 Fitness: Solution Quality, Reproductive Potential & Deception

In population genetics, “fitness” is defined operationally, indeed almost tautologically, as reproductive success: the relative number or frequency of viable fertile offspring (of a particular individual, lineage, genotype or phenotype). In EAs, “fitness” often refers to some measure of solution quality, which is used to drive the relative reproductive success by artificial selection. These two usages have been confounded at least since Sir Ronald Fisher used the term in 1930 [38]. Even within one of these 2 major classes of meaning, there are differences in the details, e.g.: offspring in the immediate generation or over time; solution goodness to be maximized or badness/cost/error to be minimized; etc. [85]

5.2.1 Solution quality of individual models

It is pointless to search for an error minimizing representation of a model unless that model both captures most of the available information about the target provided by the inputs and discards most of the irrelevant entropy of the inputs. Thus our metric for the overall information theoretic solution quality of an individual **f**$_j$ in the population is the similarity of its outputs **Z**$_j$ to those **Z** of our ideal model **f** (Equation 25), equal to the complement of their dissimilarity (Rajski distance, Equation 45)

$$1-\mathrm{D_S}(\mathbf{Z}_j,\mathbf{Z})=\frac{\mathrm{I}(\mathbf{Z}_j;\mathbf{Z})}{\mathrm{H}(\mathbf{Z}_j,\mathbf{Z})}=\frac{\mathrm{I}(\mathbf{Z}_j;\mathbf{Y})}{\mathrm{I}(\mathbf{X};\mathbf{Y})+\mathrm{H}(\mathbf{Z}_j)-\mathrm{I}(\mathbf{Z}_j;\mathbf{Y})} \qquad \mathbf{(67)}$$

Extending this from an individual to an ensemble model yields Equation 59. There is little difference between an individual and an ensemble model in our formalism, as even individual models can have vector valued outputs, so the outputs of an ensemble model are the concatenation of several vectors, which just forms a longer vector, and there is no limit on model output vector length.

A model can achieve a perfect score of one only by exactly explaining target variance: eliminating both residual target entropy and excess model entropy.

Like MSE, this indicator penalizes sources of error equally: it does not matter whether the target has fluctuations that the model does not reflect, or the model has jitter not found in the target; either error equally degrades indicated model fitness. A model that fully explains target entropy (to the extent possible based on the inputs), but with the same amount of excess model entropy, scores 0.5; a model that has no excess entropy, but explains only half the target entropy, also scores 0.5; and a model that explains half the target entropy, and has a like amount of excess model entropy, scores 0.333.

Definitions of information theoretic indices that objectively and justifiably evaluate ensemble models, without requiring foreknowledge of how the multiple constituent sub-models might be combined into a single more complex model, including the overall ensemble model quality expression of Equation 59 and its decomposition into sufficiency and efficiency, are key contributions of this thesis.

5.2.2 Solution quality of populations

As a population is an ensemble, its solution quality may be so calculated. One of its members may have higher or lower quality than the population as a whole: the member’s contribution to population sufficiency must be non-negative, but the member may degrade population efficiency. Considering a population as a source from which to select one model to be used *now* (not as a pool of genetic material for further evolution in hopes of producing better solutions), such an ensemble evaluation is of questionable utility. Typically more relevant as a measure of overall population solution quality in this case would be the value of

Equation 59 on its best member; from an information theoretic viewpoint, that best member is the one that yields the highest value of Equation 59.

Indeed, as both d_S and D_S are true distances, EA system performance with respect to improving solution quality can be measured in quasi-physical terms. E.g., using d_S, the rate at which the population's best individual approaches the ideal model is a "velocity" in units of bits per generation or bits per second, or a "gas mileage" in units of bits per fitness evaluation or bits per joule expended in computation. Actual physical interpretations of these are left for future work.

5.2.3 Reproductive potential of populations

By "reproductive potential" we mean high fitness in a sense somewhere in between those cited at the beginning of this section: not necessarily a high solution quality, but *containing material* that, under the genetic operators, is likely to produce offspring with high solution quality; we want to be able to evaluate this *inter alia* so we can select high potential parents for reproduction.

We may be tempted to treat an entire population as an ensemble and apply Equation 59. This is valid but generally misleading. At the end of an EA run, this index should reach or closely approach unity: however, that is a solution quality evaluation; reproductive potential is irrelevant at end of run. During an EA run, it is likely that preserving diverse genetic material likely to be needed later will require retaining (temporarily) substantial excess model entropy.

Sufficiency is the statistic most relevant to reproductive potential of a population. It measures what fraction of the information about the target output available from the inputs is [still] present in the population. That information is the most fundamental “building block” for constructing solutions in later generations.

Necessity of a population is purely an efficiency consideration, to be applied with care: models containing the same information may represent it in different forms; this *representation diversity* may be needed (in addition to information diversity) in the context of a particular evolutionary system and problem (see Section 6.2).

The population diversity index of Equation 65 is therefore important to assessing reproductive potential, as it shows to what extent information is distributed across different species in the population. High sufficiency indicates that the population contains most or all of the necessary building material; high diversity (low redundancy) indicates that the material consists of many relatively small blocks of various shapes that might be fitted together to construct the solution, rather than nearly indistinguishable large monoliths. The metaphor is accurate because multiple individuals cannot each contain a large share of the information without there being substantial overlap (much the same information in each).

5.2.4 Reproductive potential of individuals

We need means of evaluating a model’s quality and its *reproductive potential,* in combination with another potential parent model specifically and in the context of the population generally. Several of the tools we have developed can be brought to bear on the problem.

First, if the model is an ensemble, with a high *sufficiency* but a low *efficiency*, then it has high potential under mutation, which could delete from the ensemble constituent sub-models of low *necessity* that only contribute excess entropy, or recombine constituent sub-models of high necessity to filter excess entropy.

Example 14. If a target includes a bit that is the XOR of 2 input bits, and both of those input bits are present in an ensemble model, then the ensemble sufficiency will be high, the ensemble efficiency low and the necessity of each of the 2 sub-models high. A GP ensemble mutation that combines those 2 sub-models with an XOR non-terminal node will retain all the sufficiency present in the parent ensemble while reducing ensemble entropy by one bit (thus improving ensemble efficiency); it will also reduce ensemble model cardinality by one sub-model.

Second, in selection for sexual reproduction, 2 models can be evaluated as an ensemble. If the putative reproductive pairing has any objective values (e.g. sufficiency) better than either of the 2 constituent models, it has high potential. This possibility is explored informally in Example 15 and formally in Section 5.3.

Example 15. If again a target contains a bit that is the XOR of 2 input bits, and one model contains only one of those 2 bits and another model contains only the other of those 2 bits, then neither of those input bits will contribute to the sufficiency of either of those models. However, evaluating the 2 models as an ensemble will reveal their joint sufficiency. Thus they would be evaluated as having high reproductive potential if paired with each other specifically.

Third, if a model is evaluated in the context of ensembles formed with several potential partners randomly selected from the population, and some reasonable statistic (e.g. median, mean, mean plus standard deviation) on the results of those ensemble evaluations is good, then the model would be evaluated as having high reproductive potential in the context of that population generally. This possibility, its inverse and their implications are explored formally in Section 5.4.

There is no single universal tool for evaluating the reproductive potential of individuals for all purposes in all situations. However, the metrics and indices developed herein provide a general toolkit adaptable to many uses, and several of these indicators can be and have been used together in multiobjective EAs. Limiting ourselves to a single objective (for EAs with this restriction), there are 2 attractive candidates for mating pair selection: SRI (Equation 52) or the overall model quality index (Equation 59).

A high value of SRI for a pair of individuals strongly suggests that they should be mated: their synergistic elements may combine to yield higher MI with the target of a child than of either parent. Absent a significantly positive SRI, selection bias against pairs with large negative SRI at least minimizes redundancy in mating pairs, corresponding to specialization and division of labor. Use of indicators of pairing potential imitates volitional selection for mating or team formation, common in natural biology and human economic activity, but typically absent from evolutionary computation. SRI is attractive, but we argue that a better pair potential indicator for survival as well as reproductive selection is Equation 59:

- High SRI indicates that a pair is strong relative to its own members: however, those individual members might have been quite weak; high SRI does not tell us that the pair is any stronger than another pair.

- SRI is clearly justifiable for *reproductive* selection; for *survival* selection it is less so.

- Similarity to the ideal model is our chosen solution quality metric; using it also to evaluate reproductive potential of mating pairs is consistent, simple and presumably maximizes heritability of solution quality.

- Similarity to the ideal model benefits from synergetic and not from redundant MI with the target contributed by pair members, so it implicitly incorporates the diversity that SRI measures, albeit with a lesser weight.

For populations, the primary solution quality measure is *sufficiency*; for the constituent sub-models of an ensemble, *necessity* is primary but *efficiency* also matters. For ensemble models, these and other objectives are all important, suggesting true Pareto multi-objective optimization should be used with a vector of all the essential solution quality measures; otherwise the most important objectives in the vector can be collapsed into Equation 59.

If Equation 59 were calculated for each subpopulation that could be formed from the current population, the ideal evolutionary step would be to select for survival all those individuals in the subpopulation with the highest aggregate similarity to the ideal model, then recombine, mutate, etc. and repeat. However it is generally infeasible to do this: the indicators developed herein are computationally costly, and geometrically so in the number of arguments to the functional. Therefore, we do not suggest calculating functionals of more than 4 terms in EA main loops. In early stages of this work, we attempted to estimate even 3 term functionals with a heuristic. If it is known *a priori* that a particular problem involves no synergy, the following approach can be used to reduce computational costs.

Low computational cost approximation of 3-term functionals[43]

The basic elements in all these calculations are individual and joint entropies. H(**Y**) can be calculated once at the start of the run. H(**Z**$_j$) and H(**Y**,**Z**$_j$) of each model can be calculated when the model is first generated and added to the population, and recalculated only if it is structurally modified or if its parameters are changed. Only H(**Z**$_j$,**Z**$_k$) need be calculated during tournaments or other selection procedures; it then can be stored, in case it is needed again. Given these 1- and 2-term entropies, the essential 3-term entropy can be bounded:

$$\begin{aligned}&\max(\mathrm{H}(\mathbf{Y},\mathbf{Z}_j),\mathrm{H}(\mathbf{Y},\mathbf{Z}_k),\mathrm{H}(\mathbf{Z}_j,\mathbf{Z}_k))\\&\le \mathrm{H}(\mathbf{Y},\mathbf{Z}_j,\mathbf{Z}_k)\le\\&\min(\mathrm{H}(\mathbf{Y},\mathbf{Z}_j)+\mathrm{H}(\mathbf{Z}_k),\mathrm{H}(\mathbf{Y},\mathbf{Z}_k)+\mathrm{H}(\mathbf{Z}_j),\mathrm{H}(\mathbf{Y})+\mathrm{H}(\mathbf{Z}_j,\mathbf{Z}_k))\end{aligned} \quad \textbf{(68)}$$

This may then be used to bound the key indicator:

$$_{\mathrm{N}}\mathrm{I}(\mathbf{Y};\mathbf{Z}_j,\mathbf{Z}_k)=\frac{\mathrm{H}(\mathbf{Y})+\mathrm{H}(\mathbf{Z}_j,\mathbf{Z}_k)-\mathrm{H}(\mathbf{Y},\mathbf{Z}_j,\mathbf{Z}_k)}{\mathrm{H}(\mathbf{Y},\mathbf{Z}_j,\mathbf{Z}_k)} \quad \textbf{(69)}$$

[43] *This analysis should be renormalized by I(**X**;**Y**) vs H(**Y**), but that will only rescale it by a problem-specific constant.*

The bounds developed tend to be wide, so for tournament selection, we can try a heuristic. Absent the full 3-term functional, we do not know the target relevant diversity of two individuals. Applying the principle of insufficient reason, we *guess* that the proportion of overlap between the mutual information that each has with the target is the same as the proportion of overlap between their entropies, equating the following ratios:

$$\frac{\mathrm{I}(\mathbf{Y},\mathbf{Z}_j)+\mathrm{I}(\mathbf{Y},\mathbf{Z}_k)-\mathrm{I}(\mathbf{Y};\mathbf{Z}_j,\mathbf{Z}_k)}{\mathrm{I}(\mathbf{Y};\mathbf{Z}_j,\mathbf{Z}_k)}=\frac{\mathrm{H}(\mathbf{Z}_j)+\mathrm{H}(\mathbf{Z}_k)-\mathrm{H}(\mathbf{Z}_j,\mathbf{Z}_k)}{\mathrm{H}(\mathbf{Z}_j,\mathbf{Z}_k)} \qquad \textbf{(70)}$$

This can be rewritten as

$$_{\mathrm{N}}\mathrm{S}(\mathbf{Y};\mathbf{Z}_j;\mathbf{Z}_k)=-_{\mathrm{N}}\mathrm{I}(\mathbf{Z}_j,\mathbf{Z}_k) \qquad \textbf{(71)}$$

showing that our heuristic conservatively assumes *no synergy*.

Note that we do not actually have either of the 3-term functionals above; we are *assuming* a relationship involving them, in order to estimate **I** (**Y**; **Z**$_j$, **Z**$_k$) as

$$\hat{\mathrm{I}}(\mathbf{Y};\mathbf{Z}_j,\mathbf{Z}_k)=\frac{\mathrm{I}(\mathbf{Y},\mathbf{Z}_j)+\mathrm{I}(\mathbf{Y},\mathbf{Z}_k)}{1+_{\mathrm{N}}\mathrm{I}(\mathbf{Z}_j,\mathbf{Z}_k)} \qquad \textbf{(72)}$$

If this estimate exceeds the bounds previously developed, we limit it. We then use it to estimate the 3-term joint entropy, likewise limiting that to lie between its bounds. Finally, we estimate $_{\mathrm{N}}$**I** (**Y**; **Z**$_j$, **Z**$_k$) for each pairing of individuals in a tournament. Selection for reproduction (and survival) may then be performed, based not just on the relative fitness of the individuals, but also on the estimated relative fitness of their potential combinations.

Considering the individual and joint entropies of two models and a target, each region of which can contain zero or non-zero entropy, there are 128 qualitatively distinct cases. Discarding trivial cases leaves 63. Assigning 1 bit to each region of non-zero entropy, relative error of the estimated versus the exact value of ${}_{N}I(\mathbf{Y};\mathbf{Z}_j,\mathbf{Z}_k)$ ranges from +71% maximum to -36% minimum, with 16% RMS, 2% mean, 0% median and 0% mode (38 cases). *Absent synergy*, the largest errors would degrade survival selection based on ${}_{N}I()$, but not reproductive selection based on ${}_{N}S()$.

5.2.5 Deception

In EAs, deception refers to misleading evaluations of reproductive potential. Reproductive potential of an individual is usually estimated as its solution quality, without regard to population context or likely effects of genetic operators. Efforts to define and exploit notions of "effective fitness" are notable exceptions [99] but they do not directly counteract deception. Example 6 is a classic fully deceptive maximization problem: any hyperplane that divides the space of possible solutions in half, will yield an average solution quality in the half containing the global maximum that is lower than the average solution quality of the other half (which contains the global minimum); thus rewarding high solution quality individuals with reproductive opportunities will cause lineages to climb a solution quality ramp and then fling themselves off a solution quality cliff.

As pointed out by Goldberg in [43], globally deceptive problems are uninteresting because they will defeat any reasonable search or optimization strategy, but large problems that contain small deceptive sub-problems actually arise in practice and “competent” EAs can be designed to solve them. The original motivation for our foray into information theory was to avoid the deception encountered on Sprott’s simplest chaotic flow, where linear correlation of the dependent variable (jerk) with one of the observables (velocity) was strong (r^2 = .4) but with the other 2 observables (displacement, acceleration) was extremely weak (r^2 under 1E-9); yet all 3 observables were in fact required to express the defining equation (or any other equation that both approximately matched the data in the short term and when iterated long term reconstructed a chaotic attractor). MI was found to significantly outperform linear correlation and normalized root mean squared error (NRMSE) on this problem (see Section 6.1), avoiding that particular case of deception.

Yet MI and the various metrics and indices we have derived from it are still subject to forms of deception. Example 7 illustrates information theoretic deception, where noise corrupted individual models that could never be processed or recombined in any way such as to predict the outputs 100% correctly were preferred over clean individual models that could be recombined to predict the outputs 100% correctly. In Example 7 this could be avoided by doing reproductive selection based on evaluations of the joint solution quality of 2 potential parents taken as an ensemble (see Section 5.2.4 and Section 5.3). There are worse cases where ensemble evaluation does not avoid deception.

Example 16. Generate 2 input Gaussian RVs x_0 and x_1 with zero mean and unit variance. Generate 2 output variables y_0 and y_1 as the sum and difference of the input variables and rescale them by dividing by √2. Generate 2 more observables z_0 and z_1 by scaling the output variables by multiplying by 1.9, adding 2 independent Gaussian noise sources with zero mean and variance 0.1 (one noise source to each output variable) and rescaling by again dividing by √2. Calculate the efficiency, sufficiency and similarity of x_0 alone, x_1 alone, the ensemble (x_0, x_1), z_0 alone, z_1 alone and the ensemble (z_0, z_1), all against y_0 alone, y_1 alone and the ensemble (y_0, y_1). Observe that the individual noise corrupted output variables and their ensemble yield better values on all indices than the individual input variables and their ensemble, despite the fact that, by construction, y_0 and y_1 can be reconstructed 100% accurately from the latter but not from the former. See Table 3 for statistics on 4096 pseudo-random points.

Obviously the deception in Example 16 is synthetic and could be avoided by increasing the noise variance. This realistic example shows that deception can still occur, even after evaluating ensembles rather than individual variables and also decomposing the overall information theoretic model quality index into its sufficiency and efficiency components. Deception is in this case related to another of Goldberg’s core sources of problem difficulty, noise. It is unclear, in the context of the original motivating problem, how to distinguish deception from Goldberg’s third core source of problem difficulty, poor scaling. While the information theoretic evaluators developed herein do have significant advantages over more traditional EA evaluators, they are not magic bullets.

5.3 Non-random mating extension of Price's Theorem

Price's Equation can be derived from Slatkin's transmission-selection recursion as applied to Holland's canonical model of genetic algorithms [50]; this assumes individual selection for reproduction and random mating. We observe that if complementary individuals can be identified, they can be selected, not only for membership in ensemble models, but also for reproduction. Given a measurement of the fitness of an ensemble as such, a 2-member ensemble may be regarded as a potential reproductive pairing; reproductive selection can be then performed on these, rather than on individuals, favoring complementary matings. EAs implicitly process ensembles; we explicitly address the population dynamics of ensembles, using information theoretic functionals to evaluate them.

Price's Theorem can be extended to model the evolutionary population dynamics of ensembles under group selection with joint fitness (even if not separable as the product of the individual fitness of each parent). The details will depend upon: whether group selection is applied to survival, reproduction or both; for each of those, whether it acts alone or together with individual selection; whether groups are limited to potential parental pairings or are more general; and whether mating is random or not. Here we outline a proof for the simplest scenario: group selection applied to reproduction only, acting alone on potential parental pairings.

In [61] Price's original proof is followed, recast and simplified for use with GP, but still focusing on gene frequencies:

$$Q_1 = \text{frequency of gene of interest in population at generation 1}$$
$$Q_2 = \text{frequency of gene of interest in population at generation 2}$$
$$M = \text{number of individuals in population at generation 1}$$
$$z_i = \text{number of offspring of individual } i$$
$$q_i = \text{number of copies of gene of interest in individual } i$$
$$q_i' = \text{number of copies of gene of interest in offspring of individual } i$$
$$\Delta q_i = q_i' - q_i$$
$$\Delta Q = Q_2 - Q_1$$
$$Q_1 = \frac{\sum q_i}{M} = \bar{q}$$
$$Q_2 = \frac{\sum z_i q_i'}{\sum z_i} = \bar{q} + \frac{\text{Cov}(z,q)}{\bar{z}} + \frac{\sum z_i \Delta q_i}{M\bar{z}}$$
$$\Rightarrow \Delta Q = \frac{\text{Cov}(z,q)}{\bar{z}} + \frac{\sum z_i \Delta q_i}{M\bar{z}} \qquad \textbf{(73)}$$

In [6] the theorem and its proof are generalized to changes, not only in gene frequencies, but in arbitrary measurement functions, and justified directly in terms of Slatkin’s recursion and Holland’s canonical model of GAs:

$$\begin{aligned}
&\mathrm{F}(x) = \text{measurement function for propertyof interest as exhibited by genotype} x \\
&\mathrm{T}(x \leftarrow (y,z)) = \text{probability thatgenotype} x \text{ is produced by parentalgenotypes} (y,z) \\
&\mathrm{p}(x) = \text{frequency of genotype} x \text{ in populationat current generation} \\
&\mathrm{p}'(x) = \text{frequency of genotype} x \text{ in populationat next generation} \\
&\mathrm{w}(x) = \text{fitness of genotype} x \text{ in populationbiology sense (sampling rate)} \\
&\varphi(y,z) = \sum_x \mathrm{F}(x)\,\mathrm{T}(x \leftarrow (y,z)) = \text{expectation of } F() \text{ in offspringof } (y,z) \\
&\varphi = \sum_{y,z} \varphi(y,z)\,\mathrm{p}(y)\,\mathrm{p}(z) = \text{expectation of } F() \text{ in populationwithout selection} \\
&\Delta\overline{\mathrm{F}} = \overline{\mathrm{F}}' - \overline{\mathrm{F}} \\
&\overline{\mathrm{F}} = \sum_x \mathrm{F}(x)\,\mathrm{p}(x) \\
&\overline{\mathrm{F}}' = \sum_x \mathrm{F}(x)\,\mathrm{p}'(x)
\end{aligned}$$

$$\mathrm{p}'(x) = \sum_{y,z} (\mathrm{T}(x \leftarrow (y,z))\frac{\mathrm{w}(y)}{\overline{\mathrm{w}}}\mathrm{p}(y)\frac{\mathrm{w}(z)}{\overline{\mathrm{w}}}\mathrm{p}(z)) \quad \textbf{(74)}$$

$$\Rightarrow \Delta\overline{\mathrm{F}} = \mathrm{Cov}(\varphi(y,z), \frac{\mathrm{w}(y)\,\mathrm{w}(z)}{\overline{\mathrm{w}}^2}) \quad \textbf{(75)}$$

To apply Price’s Equation to reproductive selection of potential parental pairings as such, we need merely replace, in Slatkin’s recursion (74), the subexpression quantifying sampling frequencies of individual parents, with one quantifying sampling frequencies of parental pairings, which flows through to the conclusion:

$$\mathrm{p}'(x) = \sum_{y,z} (\mathrm{T}(x \leftarrow (y,z))\frac{\mathrm{w}(y,z)}{\overline{\mathrm{w}_2}}\mathrm{p}(y,z)) \quad \textbf{(76)}$$

$$\Rightarrow \Delta\overline{\mathrm{F}} = \mathrm{Cov}(\varphi(y,z), \frac{\mathrm{w}(y,z)}{\overline{\mathrm{w}_2}}) \quad \textbf{(77)}$$

This simple result is a key contribution of this thesis.

For it to be useful in GP, we need a solution quality indicator (fitness in the engineering sense) to drive the sampling rate (fitness in the population biology sense), as in Equation 78 below. Here we claim an advantage for normalized joint mutual information of the potential parents with the target: it is a computable, general, justifiable measure of how much 2 individuals jointly provide about the target. We use it here, but any other effective joint indicator could be substituted:

$$
\begin{aligned}
&V = \text{input data set}\\
&W = \text{target output data set (unknown function applied to input data set)}\\
&X = \text{output data set of genotype } x \text{ applied to input data set}\\
&Y = \text{output data set of genotype } y \text{ applied to input data set}\\
&Z = \text{output data set of genotype } z \text{ applied to input data set}\\
&\mathrm{F}(x) = \frac{\mathrm{I}(W;X)}{\mathrm{H}(X) + \mathrm{I}(W;V) - \mathrm{I}(W;X)}\\
&\mathrm{F}(y,z) = \frac{\mathrm{I}(W;Y,Z)}{\mathrm{H}(Y,Z) + \mathrm{I}(W;V) - \mathrm{I}(W;Y,Z)}\\
&\overline{\mathrm{F}_2} = \sum_{y,z} \mathrm{F}(y,z)\,\mathrm{p}(y)\,\mathrm{p}(z)\\
&\mathrm{w}(y,z) = \mathrm{F}(y,z)\\
&\overline{\mathrm{w}_2} = \overline{\mathrm{F}_2} && \textbf{(78)}\\
&\hat{\varphi}(y,z) = \frac{\mathrm{F}(y,z)}{\overline{\mathrm{F}_2}} = \text{normalized estimated expectation of } F() \text{ in offspring of } (y,z) && \textbf{(79)}\\
&\Rightarrow \Delta\overline{\mathrm{F}} = \mathrm{Cov}(\varphi(y,z), \hat{\varphi}(y,z)) && \textbf{(80)}
\end{aligned}
$$

We use mutual information as the solution quality indicator and as the basis for reproductive sampling; choosing it as our measurement function we can analyze its change. We estimate its expectation (79) in offspring of a potential parental pairing by its joint measurement on those potential parents. *In general the joint fitness (in either the engineering or population biology sense) does not factor as the product of the parental fitness values: joint solution quality may reflect synergy and/or redundancy; joint sampling may reflect non-random mating.*

The question now becomes, how accurate is our estimator (Equation 79)? Equivalently, how strong is the covariance? In other words, how heritable is our joint indicator F(*y,z*), the overall model quality of the parents as an ensemble, from Equation 59?

Heritability[44] refers to strong correlation between parental and offspring fitness across the changes typically induced by genetic operators. These changes must usually be small for heritability to be possible. "Small" in the phenotypic space is measured in terms of change in the measurement function (phenotypic property of interest). "Small" in the genotypic space can be measured by "edit distance": the minimum required number of applications of modification primitives by the genetic operators, starting from a parental form, to reach a child form. Correlation between the size of genotypic and corresponding phenotypic changes is one way of collapsing the question of search neighborhood structure to a scalar statistic, the complement of which is sometimes used as an indicator of problem difficulty.

It is apparent that mutual information *is* heritable in various simple edits, where RMSE and correlation are not. Changing additive or multiplicative constants changes RMSE, but is transparent to correlation and mutual information. Changing from an odd to an even power changes (and can altogether eliminate) correlation, but merely causes the loss of one bit or less of mutual information.

Low RMSE implies high similarity (Equation 59), but the converse does not hold: members of recoding equivalence classes with respect to similarity can differ

[44] Narrow-sense: in biology, the fraction of the phenotypic variance that can be used to predict changes in population mean; see http://en.wikipedia.org/wiki/Heritability

greatly with respect to RMSE. Most transformations that alter RMSE and many that alter correlation are transparent to mutual information, which is preserved for discrete distributions by arbitrary permutations and for continuous distributions by embeddings; the only deterministic transformations that alter MI are topological foldings (especially infinitely repeated ones due to periodic functions).

5.4 Effective fitness interpretation of Price's Theorem

It is instructive to consider defining individual fitness as the expectation of the fitness of all pairs of which the individual might be a member -- we recover the original equation, with a new interpretation:

$$\mathrm{w}(y) = \sum_{z} \mathrm{w}(y,z)\,\mathrm{p}(z) \qquad \textbf{(81)}$$

$$\Rightarrow \Delta\bar{\mathrm{F}} = \mathrm{Cov}(\varphi(y,z), \frac{\mathrm{w}(y)\,\mathrm{w}(z)}{\bar{\mathrm{w}}^2}) \qquad \textbf{(82)}$$

The fitness of an individual may not be independent of the makeup of the population of which it is a member: if the value of the individual's solution quality (or any other measurement function applied to that individual in isolation) is high, but the rest of the population does not contain individuals that, when recombined with the given individual, are likely to produce offspring with comparably high or higher values, then the individual's effective fitness is low. This relates to the concepts of "boosting" and "fitness sharing". On the other hand, if the population does contain complementary individuals, then the effective fitness depends upon the population size, the relative frequencies of the particular and the complementary species, the effectiveness of the reproductive selection operator, and any other deviations from panmixis (e.g. geographic isolation of demes), etc.: if the individual is likely to be mated with a complementary individual, effective fitness is high; if reproduction is mostly random, and only slightly biased by the joint fitness evaluation, it may not be. This affects the effective fitness of an individual genotypic line considered over multiple generations and the overall population's *evolvability*.

5.5 Evolvability, Heritability, Respect, Assortment & Geometricity

Evolvability

Evolvability has been considered at least since Sewall Wright first introduced the metaphor of the fitness landscape or adaptive landscape[45]. However, a rigorous mathematical definition remains lacking today. The general notion is

> Evolvability is loosely defined as the capacity to evolve, alternatively the ability of an individual or population to generate fit variants... Thus evolvability is more closely allied with the potential for fitness than with fitness itself; two equal fitness individuals or populations can have very different evolvabilities... Typically, researchers use some definition of evolvability based on the offspring of current individuals or populations… the transmission function of all possible offspring from a parent to define a set of metrics of evolvability... It is often argued that there may be long-term trends for evolvability to increase... However, as evolvability is more directly related to fitness potential than fitness itself, long-term change cannot be due to straight fitness selection. Thus any trend towards change in evolvability can only be understood through some second order selection mechanism by which evolution tends to retain solutions that have a more evolvable genetic system.
>
> Smith, Husbands, Layzell and O'Shea, 2002 [95]

[45] Wright used neither term, but simply "field" in [112].

Most current EAs are nowhere near expressive enough for a second-order mechanism to increase evolvability[46]. In the absence of such a mechanism, as the average solution quality of a population increases, evolvability must inevitably decrease: while effective reproductive selection may improve the relative likelihood of producing improved offspring, the overwhelming preponderance of less over more fit possible genotypes will ultimately dominate[47].

In a plot of the cumulative distribution function of fitness, the area to the left of highly fit parents will greatly exceed that to the right. Altenberg [6] states that the upper tail of the fitness distribution must grow larger with an EA than with random search for the EA to be successful.

We have already argued that similarity of an ensemble to the target is an indicator not only of solution quality (fitness in the engineering sense) but also of reproductive potential and therefore a good candidate for driving, in artificial selection, the reproductive sampling rate (fitness in the population biology sense). It is *a priori* reasonable that a joint evaluation of a characteristic of a set of parents would be more predictive of the characteristic in their offspring under recombination and mutation than the product of the values of that characteristic evaluated on each of the parents individually. This leads to several practical selection heuristics.

Pair selection for reproduction as in Section 5.3, acting in the absence of genetic operators, will increase the preponderance of informative ensembles of size 2;

[46] A more optimistic view is [5].

[47] Prof. Chilukuri Mohan has dubbed this (in unpublished discussions) “the pressure of the prior”.

these are good candidates for recombination, especially if their SRI is high. This index may be used to bias selection in favor of pairs that are not just good teams but also likely good parents: those with the potential to produce offspring better (closer to individual sufficiency and/or efficiency) than themselves.

If recombination produces offspring that are less efficient without being more sufficient, or less sufficient without being more efficient, than their parents, there appears to be no advantage to adding them to the population. Brood selection has been promoted in [3] etc.

Recombinations that trade sufficiency for efficiency or *vice versa* cannot be so easily evaluated, especially in isolation: one multiobjective heuristic is to maintain population sufficiency, try to achieve individual sub-model necessity and strive for similarity in ensemble models; ensembles evolved per this heuristic implement nonlinear Independent Components Analysis (ICA). Neutral recombinations, like neutral mutations, may be needed along the evolutionary path; they do not affect information theoretic attributes but may improve representation diversity.

Scrupulous attention to information theoretic considerations reduces the need for mutation. If population sufficiency is maintained during selection, essential information is not lost; mutation is often needed to re-introduce such information that has been lost in evolutionary algorithms that do not guarantee preservation of population sufficiency. Assuming that all individuals in the population contribute to its sufficiency, potential for useful mutation is indicated by an individual’s low efficiency. Mutation can improve efficiency by discarding input entropy that does not contribute to modeling the target.

In Section 6.1, a 3 element ensemble was required for sufficiency, using raw inputs. One of those elements was improved by squaring it, thereby discarding one bit of information (the sign): reducing its individual sufficiency (but not the ensemble sufficiency); and increasing its individual efficiency (and the ensemble efficiency); also increasing the ensemble overall solution quality. The argument for brood selection applies here also.

To justify these heuristics more rigorously, we return to Price’s Equation.

We define the improvement of the expectation of the measurement function in the brood relative to the average of the parents as

$$\mathrm{C}(x,y)=\varphi(x,y)-\frac{\mathrm{F}(x)+\mathrm{F}(y)}{2} \qquad \textbf{(83)}$$

and rewrite Equation 77 in terms of relative change to show the effects of the selection and genetic operators and their interaction as

$$\frac{\overline{\mathrm{F}_2}'-\overline{\mathrm{F}_2}}{\overline{\mathrm{F}_2}}= \qquad \textbf{relative change (84)}$$

$$\frac{\sigma_{\mathrm{F}(x,y)}}{\mu_{\mathrm{F}(x,y)}}\frac{\sigma_{\mathrm{w}(x,y)}}{\mu_{\mathrm{w}(x,y)}}\rho_{\mathrm{F}(x,y),\mathrm{w}(x,y)} \qquad \textbf{pair selection (84.1)}$$

$$+\frac{\mu_{\mathrm{C}(x,y)}}{\mu_{\mathrm{F}(x,y)}} \qquad \textbf{genetic variation (84.2)}$$

$$+\frac{\sigma_{\mathrm{w}(x,y)}}{\mu_{\mathrm{w}(x,y)}}\frac{\sigma_{\mathrm{C}(x,y)}}{\mu_{\mathrm{F}(x,y)}}\rho_{\mathrm{w}(x,y),\mathrm{C}(x,y)} \qquad \textbf{their interaction(84.3)}$$

in which we may identify the key factors affecting the essential requisites[48]

$$(\mu_{\mathrm{C}(x,y)},\sigma_{C(x,y)}) \qquad \textbf{heritability (85.1)}$$

$$\rho_{\mathrm{w}(x,y),\mathrm{C}(x,y)} \qquad \textbf{evolvability (85.2)}$$

either as averages over the space of possible evolutionary trajectories or stepwise as we proceed along one of those trajectories. Obviously this can be further generalized to an arbitrary number of parents in the selection function.

This is a key contribution of this thesis.

[48] A less detailed treatment of this, under random mating, may be found in [77].

In the case of “fitness proportionate selection”, where we set the sampling rate equal to the solution quality index (or other measurement function, as in Equation 78), Equation 84 reduces to

$$\frac{\overline{\mathrm{F}_2}' - \overline{\mathrm{F}_2}}{\overline{\mathrm{F}_2}} = \qquad \textbf{relative change (86)}$$

$$\frac{\sigma^2{}_{\mathrm{F}(x,y)}}{\mu^2{}_{\mathrm{F}(x,y)}} \qquad \textbf{pair selection (86.1)}$$

$$+\frac{\mu_{\mathrm{C}(x,y)}}{\mu_{\mathrm{F}(x,y)}} \qquad \textbf{genetic variation (86.2)}$$

$$+\frac{\sigma_{\mathrm{F}(x,y)}\sigma_{\mathrm{C}(x,y)}}{\mu^2{}_{\mathrm{F}(x,y)}}\rho_{\mathrm{w}(x,y),\mathrm{C}(x,y)} \qquad \textbf{their interaction (86.3)}$$

The genetic operators are generally thought to be neutral, on average, with respect to any phenotypic attribute measurement function, so

$$\frac{\overline{\mathrm{F}_2}' - \overline{\mathrm{F}_2}}{\overline{\mathrm{F}_2}} = \frac{\sigma^2{}_{\mathrm{F}(x,y)}}{\mu^2{}_{\mathrm{F}(x,y)}} + \frac{\sigma_{\mathrm{F}(x,y)}\sigma_{\mathrm{C}(x,y)}}{\mu^2{}_{\mathrm{F}(x,y)}}\rho_{\mathrm{F}(x,y),\mathrm{C}(x,y)} \qquad \textbf{(87)}$$

The first term will be positive. The second term will be negative (on average, and at each step after evolution has significantly improved the measurement function beyond that of a randomly generated individual). Biologists expect a population that is large enough to avoid stochastic effects to eventually fixate, with selection exactly balancing genetic variation. EA users would like to hold off such convergence as long as possible, to achieve the maximum possible solution quality. Minimizing the magnitude of the inevitably negative correlation coefficient between the fitness and its average “improvement” due to the genetic operators will extend the duration of the phase in which the effect of selection dominates that of the mostly deleterious genetic variations.

As in EA applications we care little about the mean of the population (except earlier in evolution, where we do not want good fitness individuals to be dragged back down in recombination with predominantly poor reproductive partners), it is more interesting to study the distribution of fitness than just its mean. One approach [6] is to define a local performance equation predicting growth in the population upper fitness tail. In that formulation, the key parameter limiting evolvability again is a correlation coefficient like that in Equation 85.2.

Heritability

Altenberg shows that “that there needs to be a correlation between complementary schemata of high fitness and the fitness distributions of their recombinant offspring in order for the GA to increase the chance of sampling fitter individuals” and states his intuition “that the best estimator for predicting the behavior of a GA is simply the approximation of the transmission function in the fitness domain, and that it is the rate of decay of evolvability as parents increase in fitness that is the critical feature of the transmission function for GA performance”.

We agree: the neighborhood structure of the genotypic search space, evaluated in the phenotypic fitness space, is the beginning and end of the matter of EA performance.

Thus next we look to Radcliffe’s forma analysis and Moraglio’s geometricity to justify why information theoretic ensemble fitness indices should be more heritable than traditional fitness functions, and then in Sections 6.1 and 6.2 show empirical evidence on difficult problems that these indices *remain heritable as population average fitness increases, thereby maintaining evolvability*.

Respect and Assortment

In the context of Radcliffe’s forma analysis, “respect” is defined formally. We may informally summarize that definition as: if both parents share a particular heritable attribute, all offspring of those parents (under recombination without mutation) must share it also. Likewise the formal definition of “assortment” may be informally summarized as: if one parent has a particular heritable attribute, and the other parent has another distinct heritable attribute, it must be possible (under recombination without mutation) to produce offspring with both attributes. Clearly respect and assortment are only compatible if the set of formae under consideration are separable; that is, one can establish an orthogonal or at least linearly independent basis for heritable attributes. If a genetic representation and set of genetic operators guarantee both respect and assortment, then they will provide the strong heritability and the genetic variability required for evolvability.

Geometricity

In the context of Moraglio’s attempted geometric unification of EAs, a crossover operator is considered geometric if it possesses certain formally defined properties, which we may informally summarize as: (a) it can produce only offspring lying along a shortest path (through genotypic search space) between the parents; and (b) it can produce at least one offspring lying along such a path. The former condition (a) is equivalent to Radcliffe’s definition of respect. If the latter condition (b) is strengthened to require that the operator be able to produce offspring at *all* the feasible points lying along *all* such paths between given parents, it becomes equivalent to Radcliffe’s definition of assortment. The linearly independent (preferably orthogonal) basis in Radcliffe’s formalism corresponds to the genotypic search space coordinate system in Moraglio’s.

We touched upon Kolmogorov based metrics for genotypic analyses in Chapter 2 but our emphasis throughout has been on Shannon based metrics for phenotypic evaluations; sometimes these satisfy Radcliffe’s conditions and sometimes not.

Example 17. Observable inputs are independent unbiased binary RVs $\{x_1, x_2\}$. Observable outputs are $\{y_1, y_2\}$ where $y_1=x_1$ and $y_2=x_2$. A population of ensemble models includes $\{z_1=\{x_1\}, z_2=\{x_2\}, z_3=\{x_1,x_2\}, z_4=\{\}\}$. An orthogonal basis is the pair of sufficiency recoding-equivalence class membership indicator functions (predicates) $\{\ \varphi_1=\mathrm{Floor}(\mathrm{I}(Y_1;Z^m)/\mathrm{I}(Y_1;X^m)),\ \varphi_2=\mathrm{Floor}(\mathrm{I}(Y_2;Z^m)/\mathrm{I}(Y_2;X^m))\ \}$. Recombination operators yield offspring that include all sub-models common to both parental ensembles plus a random subset of sub-models present in either. In this case: formae are separable; respect and assortment are compatible and guaranteed; recombination is geometric.

Example 18. Extend the outputs to include $y_3=x_1$ XOR x_2. Extend the basis to include $\varphi_3=\mathrm{Floor}(\mathrm{I}(Y_3;Z^m)/\mathrm{I}(Y_3;X^m))$: it is no longer orthogonal; it is still linearly independent but only to order 2. Model z_1 is a member of the forma ($\varphi_1=1$, $\varphi_2=$<don't care>, $\varphi_3=0$) and z_2 is a member of ($\varphi_1=$<don't care>, $\varphi_2=1$, $\varphi_3=0$). Recombination of these models should be able to produce offspring that are members of the forma ($\varphi_1=1$, $\varphi_2=1$, $\varphi_3=0$): recombination can indeed produce model z_3 but it is not a member of that forma; it is instead a member of ($\varphi_1=1$, $\varphi_2=1$, $\varphi_3=1$) because knowing y_1 and y_2 causes one implicitly to know y_3. In this case: formae are not separable; respect and assortment are not compatible; recombination is not geometric (in a Euclidean space).

We do not believe Example 18 to be a demonstration of a deficiency of our indicators but rather of the fundamental difficulty of parity. Formae based on our indicators in this case are not separable because information about the outputs in this case is not separable. Had we defined equivalence classes with respect to similarity (as in Equations 46 through 48), the problem would have been needlessly worse: zero dissimilarity to one output is intrinsically incompatible with simultaneous zero dissimilarity to another output (unless those 2 outputs have zero dissimilarity between them, in which case they are equivalent, effectively a single output). Equations 46 through 48 are too precise for use as a basis: that predicate indicates the recoding-equivalence classes of Crutchfield.

Note that while we are concerned with equivalence classes with respect to information, whose members may differ with respect to RMSE, it is also possible to find equivalence classes with respect to RMSE, whose members may differ with respect to information. See Table 4.

Ensemble Genetic Operator Respect of Information

There are several ways to mutate an ensemble model. Mutating one or more of its sub-model expressions will generally *not* be respectful of information. Recombining 2 of its expressions, using typical GP crossover operators, often will behave as if it were respectful of information, but this is by no means guaranteed. Duplicating one of its expressions *is* respectful of information. Composing 2 of its expressions (attaching one as a leaf of another tree) may improve efficiency and/or necessity but will reduce sufficiency (absent duplicates of the original expressions), so it is *not* respectful. Inserting a new expression, randomly generated from the terminal and non-terminal sets, is *not* respectful. Deleting a unique expression is *not* respectful. Algebraic simplification of constituent expressions and deletion of equivalent expressions *are* respectful.

When respect of information is observed, Equation 88 will hold. Primed variables denote offspring and we have deviated from our previous notation by distinguishing ensemble models with subscripts and omitting the superscripts indicating ensemble cardinality.

$$\begin{aligned}
&\mathrm{I}(\mathbf{Z}'_j,\mathbf{Z}'_k) \geq \mathrm{I}(\mathbf{Z}_j,\mathbf{Z}_k)\\
&\mathrm{I}(\mathbf{Y};\mathbf{Z}'_j) \geq \mathrm{I}(\mathbf{Y};\mathbf{Z}'_j;\mathbf{Z}'_k) = -S(\mathbf{Y};\mathbf{Z}'_j;\mathbf{Z}'_k)\\
&\mathrm{I}(\mathbf{Y};\mathbf{Z}'_k) \geq \mathrm{I}(\mathbf{Y};\mathbf{Z}'_j;\mathbf{Z}'_k) = -S(\mathbf{Y};\mathbf{Z}'_j;\mathbf{Z}'_k)
\end{aligned} \tag{88}$$

Recombining ensemble models (regrouping but not modifying their sub-model expressions, as in Examples 17 and 18) is similar to Radcliffe’s operator R3/RTR (Random Respectful Recombination / Random Transmitting Recombination), although the infinite universal set implicit in GP requires some modifications; it *is* respectful of phenotypic information, per Equation 89.

$$\begin{aligned}
&\mathrm{I}(\mathbf{Y};\mathbf{Z}'_j) \leq \mathrm{I}(\mathbf{Y};\mathbf{Z}'_j,\mathbf{Z}'_k) \leq \mathrm{I}(\mathbf{Y};\mathbf{Z}_j,\mathbf{Z}_k)\\
&\mathrm{I}(\mathbf{Y};\mathbf{Z}'_k) \leq \mathrm{I}(\mathbf{Y};\mathbf{Z}'_j,\mathbf{Z}'_k) \leq \mathrm{I}(\mathbf{Y};\mathbf{Z}_j,\mathbf{Z}_k)\\
&\mathrm{I}(\mathbf{Y};\mathbf{Z}'_j,\mathbf{Z}'_k) \leq \mathrm{H}(\mathbf{Z}'_j;\mathbf{Z}'_k) \leq \mathrm{H}(\mathbf{Z}_j;\mathbf{Z}_k)\\
&\mathrm{I}(\mathbf{Y};\mathbf{Z}'_j,\mathbf{Z}'_k) \leq \mathrm{H}(\mathbf{Z}'_j;\mathbf{Z}'_k) \leq \mathrm{H}(\mathbf{Z}_j;\mathbf{Z}_k)\\
&\mathrm{I}(\mathbf{Y};\mathbf{Z}'_j) \leq \mathrm{H}(\mathbf{Z}'_j) \leq \mathrm{H}(\mathbf{Z}'_j;\mathbf{Z}'_k) \leq \mathrm{H}(\mathbf{Z}_j;\mathbf{Z}_k)\\
&\mathrm{I}(\mathbf{Y};\mathbf{Z}'_k) \leq \mathrm{H}(\mathbf{Z}'_k) \leq \mathrm{H}(\mathbf{Z}'_j;\mathbf{Z}'_k) \leq \mathrm{H}(\mathbf{Z}_j;\mathbf{Z}_k)
\end{aligned} \tag{89}$$

Aggregation of additional non-redundant expressions into a child ensemble increases sufficiency. The raw mutual information of a child model with the target cannot exceed that of its parents jointly with the target; but the overall model quality of the child *can* exceed that of its parents, if the child has less excess entropy than the aggregate of its parents (as it often will).

Note that respect implies that redundancy raises the lower bound on the possible results of an unfavorable recombination, whereas synergy raises the upper bound on the possible results of a favorable one. Respect causes total MI between models to increase, which will likely increase their redundancy also, possibly leading to emergent speciation: an evolutionary mechanism for maximizing the likelihood of mating composable representations and minimizing the likelihood of mating non-composable representations that yield “monsters”. This is merely a conjecture![49]

From the above recounting, it is clear that respect of information is inconsistent over different genetic operations, especially when it is incompatible with the assortment needed for evolvability, but often respect is preserved, as required to maintain the heritability also needed for evolvability.

We do not claim our indicators always will pick exactly the right pairings; only that they enable selection that maximizes overall reproductive potential by commensurately measuring solution quality and relevant diversity. With their use, versus MSE or correlation, or together with MSE, high fitness building blocks are less likely to be lost, and more likely to be recombined.

Encouraged by the analysis of Chapter 5, we next present, in Chapter 6, experimental evidence for heritability of MI and evolvability using it.

[49] Might it be another explanation for Lee Altenberg’s “evolution of evolvability in GP”?

6 Experiments

This Chapter supports our claims with evidence gathered on hard synthetic problems. In Section 6.1, we revisit the original motivating problem of chaotic time series prediction. In Section 6.2, we work what is in some sense the most difficult discrete problem, parity. In Section 6.3, we return to continuous data, benchmark problems for symbolic regression.

6.1 Chaotic time series prediction

We analyze the effects of early versions of our MI based indicators on GP regression of the jerk equation that initially motivated their development. With a terminal set of 3 variables { x'', x', x } and a non-terminal set of 4 basic arithmetic operators { +, -, *, / }, discarding duplicate expressions and those that reduce to constants (with neither variance for correlation nor entropy for mutual information), we form complete populations of 3 (one term), 27 (up to 2 term) and 296 (up to 3 term) distinct expressions. We apply them all to inputs **X** and compute their RMSE, correlation and mutual information with the target **Y**. Entropy and MI are calculated using the "equal width bin" method and MI is normalized in the "unfair" sufficiency approach, dividing it by target output entropy H(**Y**). To enable use of each of the 3 indicators to estimate pair potential consistently: we normalize MI as noted, and take the reciprocal of the sum of one plus RMSE; correlation is inherently normalized.

6.1.1 Individual fitness (solution quality & reproductive potential)

Consider the initial population of all the terminals as shown in Table 5. Normalized RMSE has significantly different values for each terminal, but all of the same order of magnitude. Correlation favors the first derivative over the others by several orders of magnitude. NMI assigns similar fitness to each terminal. We immediately see an advantage of NMI: the correct solution requires all the terminals.

Next consider all the expressions of up to 2 terms as shown in Table 6. Quotient expressions were ranked poorly by all indicators, and have been dropped (here, and subsequently when we consider diversity). NRMSE ranks the product of the derivatives as the highest fitness, by a significant margin over the second ranked expression; but this is not a part of the correct solution. Correlation ranks the product of the observable and its first derivative as the highest fitness, by a significant margin over the second ranked expression. This is not a part of the defining equation, but it is a term of an alternative equation yielding the same behavior to within a constant. NMI ties the second derivative and twice that term as highest fitness; twice the second derivative is a term of the correct formula, as accurately as it can be expressed without real-valued coefficients.

Space permits tabulation and discussion of only a few of the expressions of up to 3 terms. Table 7 contains the top 5 rankings of each indicator, plus the top 10 sorted on the sum of the rankings (indicator ‘consensus’). All indicators ranked highly at least some of the optimal sub-expressions, but none gave them top rankings. The two entries in bold are maximal length correct subexpressions that can be achieved in this third generation, but are not ranked highly by the indicators; the third expression of that form received the topmost ranking by consensus of the indicators, but was not in the top five of any of the individual indicators. NMI, which is a generalization of correlation, approximately agreed on rankings as often with NRMSE as it did with correlation, which is somewhat surprising. Forms close to the correct formula were identified slightly more often by NMI than by NRMSE, both of which significantly outperformed correlation.

6.1.2 Diversity

To construct a diversity indicator from the fitness indicators of the previous subsection, the no-synergy (redundancy only) heuristic of Equation 71 was extended from NMI to correlation and normalized RMSE: this technique is admittedly suspect; but no alternative was apparent. Consider the initial population of all the terminals as shown in Table 8. Correlation ranks the first derivative better than either of the other terms by several orders of magnitude. If selection pressure were exerted on the basis of correlation as a diversity indicator, the observable and/or its second derivative would be lost from the population, precluding evolution of the correct solution.

Next we consider 2-term expressions as shown in Table 9. All indicators rank quotient expressions high on diversity; but these were ranked low on fitness by all indicators, so they are not shown in either table. RMSE ranks 20th in diversity, correlation ranks 11th in diversity, and normalized mutual information ranks 12th in diversity, the expressions each respectively ranked fittest. Thus with selection for diversity as well as fitness, both correlation and normalized mutual information are likely to preserve the individuals they ranked most fit; RMSE is not. All indicators ranked the observable very poorly on diversity. This is perhaps inevitable, as all the other expressions are derived from it; but it is unfortunate, as the observable is one of the terms of the solution.

6.1.3 Pairing & Pair Potentials

The pairing potential (reproductive selection only) and pair potential (survival selection also, especially for ensemble models) were estimated, again using the heuristic of Equation 71, corresponding formulas based on NRMSE and correlation, and approximations of the full 3-term NMI and the normalized synergy. The data tables are lengthy; hence only summary conclusions are given here. The NMI heuristic outperformed correlation, but not NRMSE, in identifying high potential pairs; the 3-term NMI approximation, even with far fewer bins, did better. SRI was a reasonably reliable indicator of when the heuristic would be an underestimate; in addition to pairing, it proved a very good indicator of pair, potential. Synergy proved to be a critical aspect of the test problem.

To evaluate the MI based indicators for use in reproductive selection operators in light of Price's Theorem, we wished to avoid the disruptive effects of particular genetic operators that could obscure the selective properties of the indicators, so they were not evaluated in the context of any given real GP system. Instead, a non-disruptive trivial recombination operation was used: for each potential pairing, combine parent trees using each of the non-terminals; measure fitness of all offspring; compute mean and max fitness of the brood. Correlation coefficients between 2nd generation parental indicators and 3rd generation offspring indicators were then calculated.

Correlation of mean fitness of the brood, with the fitness of either parent, was 0.6; with the product of the fitnesses of the parents, 0.9. Correlation of mean fitness of the brood, with the joint NMI between the parents and the target, was 0.8 in the general population of 351 potential pairings; after dropping low fitness individuals from the breeding population, reducing it to 210 potential pairings, it still held at 0.3. Correlation of improvement of the best of the brood versus the better of the parents, with the normalized synergy of the parents, was 0.2 in that breeding population; after also dropping pairings with low joint NMI against the target, reducing the breeding population to 140 potential pairings, it was 0.4.

Dropping pairings with low synergy, of the 70 pairs remaining, 45 produced broods whose best members had fitness higher than the better of their parents. Of these, 13 had offspring better than any individual in the breeding population. Of these, 10 included at least one parent that was an optimal subexpression. Of these, 4 pairs were composed of 2 optimal subexpressions. Lastly, of these, 2 sets of parents correctly paired 2 optimal subexpressions.

These findings are strong evidence of effectiveness of information theoretic fitness and diversity indicators on this "easiest hard problem": the simplest chaotic flow.

6.2 4-parity using 2-MUXes

Arbitrary logic functions may be synthesized using only multiplexers, each of which is controlled by a variable of the target function that selects between 2 inputs, which can be only zero, one or the output of another 2-MUX.

Evolutionary synthesis of 4-parity, using mutual information as a factor in the fitness function, was reported in [4]. First we illustrate the application of information theory to ensemble selection in synthesis of 3-parity.

Arranging the truth table by ones count as Table 10 shows the relationship to parity and the route to an optimal solution. A basic stage can be defined, with 2 inputs and 2 outputs; one stage per variable is required, arranged as a ladder with criss-crossed connections. This requires re-use of stage outputs, which cannot be represented in a GP tree without Automatically Defined Functions [57], assignment or the like, so a simple GP representation will have redundant MUXes in the earlier stages, as shown in Figure 18.

The same control variable is used on the left and right sides of the ladder at each stage; a different control variable is used at each stage of the ladder, each exactly once, in any order. GP pseudo-code of a stage follows.

Source Code Fragment 5.

```
(<this-stage-control-variable>,
        <previous-stage-ones-count-odd>,
        <previous-stage-ones-count-even>)
(<this-stage-control-variable>,
        <previous-stage-ones-count-even>,
        <previous-stage-ones-count-odd>)
```

One of the optimal solutions is thus the genotype

$(x_3,(x_2,(x_1,0,1),(x_1,1,0)), (x_2,(x_1,1,0),(x_1,0,1)))$

which may be more conventionally represented as

$$\sim x_3(\sim x_2 x_1 + x_2 \sim x_1) + x_3(\sim x_2 \sim x_1 + x_2 x_1)$$

and we will use this more familiar notation hereinafter.

Information theoretic insights into evolutionary learning become apparent when we compute the basic functionals and indices, first on primitive, then on more complex, models.

$$\begin{aligned}
&y = \mathrm{f}(x_1, x_2, x_3) = x_1 \oplus x_2 \oplus x_3 \\
&\forall j \in [1..3] : z_j = \mathrm{f}_j(x_1, x_2, x_3) = x_j \\
&\forall j \in [1..3] : \mathrm{H}(\mathbf{Z}_j) = \mathrm{H}(\mathbf{X}_j) = 1 = \mathrm{H}(\mathbf{Y}) \\
&\forall j \in [1..3] : \mathrm{I}(\mathbf{Y}; \mathbf{Z}_j) = \mathrm{I}(\mathbf{Y}; \mathbf{X}_j) = 0 \\
&\forall j, k \in [1..3], j \neq k : \\
&\mathrm{I}(\mathbf{Y}; \mathbf{Z}_j, \mathbf{Z}_k) = \mathrm{I}(\mathbf{Y}; \mathbf{X}_j, \mathbf{X}_k) = 0 \\
&\mathrm{I}(\mathbf{Y}; \mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3) = \mathrm{I}(\mathbf{Y}; \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3) = 1 \\
&\mathrm{S}(\mathbf{Y}; \mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3) = 1
\end{aligned} \qquad \textbf{(90)}$$

All the information contained in the inputs must be combined synergistically to model the target: but individual models, each containing all the information from a single input, convey no information about the target; neither do pairwise ensembles; only with all 3 models is the population sufficient. One key point here is that an individual may safely be removed from the population only if doing so does not compromise *population* sufficiency. Another is that an individual's necessity (contribution to population sufficiency) must be measured as such, not estimated on the basis of measured individual sufficiency.

Next consider 2 more complex models of 3-parity:

$$\begin{aligned}
z_j &= \mathrm{f}_j(x_1,x_2,x_3) = \overline{x_1}\,\overline{x_2}x_3 + \overline{x_1}x_2\overline{x_3} + x_1\overline{x_2}\,\overline{x_3} \\
z_k &= \mathrm{f}_k(x_1,x_2,x_3) =\sim(\overline{x_1}\,\overline{x_2}x_3 + \overline{x_1}x_2\overline{x_3} + x_1\overline{x_2}\,\overline{x_3} + x_1x_2x_3) \\
&\mathrm{D_S}(\mathbf{Z};\mathbf{Z}_j) \approx 0.6 \\
&\mathrm{D_S}(\mathbf{Z};\mathbf{Z}_k) = 0
\end{aligned} \tag{91}$$

Model *j* is correct in 7 of 8 cases, model *k* in none; yet the latter has a higher (indeed, a perfect) overall information theoretic solution quality index (complement of dissimilarity with the ideal model). This is because the latter may be subjected to an (*a priori* unknown) invertible transformation (here, negation) to yield the correct answer in all cases. The model with the lowest information theoretic error may not exhibit the lowest error in the problem's native output variable space, until it is appropriately transformed, and the required transformation may be neither obvious nor simple (although it was in this example). Given particular terminals, non-terminals and genetic operators, for evolvability, we may need multiple forms of a subexpression (here, a term and its negation), that is, representation diversity, despite the information redundancy.

Finally consider the progression of an ensemble, from containing all the raw inputs, to containing only a single individual that perfectly models the target, at each step maintaining sufficiency and incrementally approaching efficiency, with composition (or trivial crossover) only:

$$\begin{aligned}
&\mathrm{D_S}(\mathbf{Z};\{x_1,x_2,x_3\}) = 2.0/3.0 \\
&\mathrm{D_S}(\mathbf{Z};\{x_1\overline{x_2},\overline{x_1}x_2,x_3\}) = 1.5/2.5 \\
&\mathrm{D_S}(\mathbf{Z};\{x_1\overline{x_2}+\overline{x_1}x_2,x_3\}) = 1.0/2.0 \\
&\mathrm{D_S}(\mathbf{Z};\{(x_1\overline{x_2}+\overline{x_1}x_2)\overline{x_3},\overline{x_1\overline{x_2}+\overline{x_1}x_2x_3}\}) = 0.5/1.5 \\
&\mathrm{D_S}(\mathbf{Z};(x_1\overline{x_2}+\overline{x_1}x_2)\overline{x_3}+\overline{x_1\overline{x_2}+\overline{x_1}x_2x_3}) = 0.0/1.0
\end{aligned} \tag{92}$$

Problem inputs usually will be jointly [epsilon] sufficient but severally insufficient, and both jointly and severally far from efficiency. Assuming them to be so, the benefit of recombination is production of individuals that compose information provided by the inputs so as to enable formation of sufficient ensembles that gradually approach efficiency as evolution proceeds. These ensembles should grow ‘smaller’ during the evolutionary run: sometimes in terms of their joint entropy; sometimes in terms of the number of constituent individual models. This insight should influence survival and reproductive selection strategies.

While this example did not require mutation, it does show the benefit of neutral mutations. Both original and negated forms of the inputs are required; mutation could easily have introduced the negation. Although such a mutation would not be evaluated as neutral in terms of error, it would be in terms of information. Information theoretic fitness functions will regard far more mutations as neutral than will correlation or error based fitness functions: anything that preserves Crutchfield’s “recoding-equivalence class” membership will not alter the fitness.

Synthesis of parity is much harder than it may appear. The number of tree sizes k up to that required to implement m-parity is 2^m-1. The number of structurally different trees at each size k is C_k = (2k)! / ((k+1)! k!). The number of different assignments of m control variables to k MUXes is m^k. The number of different assignments of inputs in {0,1} to leaf MUXes is 2^{k+1}. Thus a conservative estimate of search space size is

$$\sum_{k=1}^{2^m-1} C_k m^k 2^{k+1} \qquad \textbf{(93)}$$

The number of correct solutions in that space is only $s_m = m * s^2_{m-1}$ where $s_0 = 1$ so the growth in the ratio of search space size to the number of solutions is a double exponential O(2 ^ 2^m). With only 4 inputs it is over a billion billion to one. The search space is highly epistatic. With conventional fitness functions, the adaptive landscape is not smooth. 4-parity was the problem addressed in the work reported by Aguirre and Coello: we attempted to beat their performance, but were unable even to match it; we have some idea why our EA stalled on a neutral plateau short of the final solution, but we do not know why their EA did not encounter the same problem.

We implemented a generational GP supporting uniform, truncation and rank proportional selection for reproduction and those same choices for survival. Fitness is multi-objective and the domination check is masked to regard any subset of the fitness vector: sufficiency, efficiency, overall information theoretic solution quality, inverse ensemble cardinality, correlation and inverse total ensemble size. The domination mask is dynamically set based on population fitness statistics to enable incremental evolution; for instance, to consider sufficiency only until the median individual is fully sufficient, then consider efficiency also. Expression recombination is performed by traditional GP subtree swapping. Expression mutation randomly makes one of the minimal moves in tree edit distance. Individuals are ensemble models represented as multisets with parameterized constraints on the maximum number of distinct expressions and the maximum number of copies of each expression. Ensemble recombination is

R3/RTR modified as described in Section 5.5. Ensemble mutation consists of randomly performing one of the operations listed in Section 5.5.

In experiments thus far, information theoretic GP has rapidly aggregated the raw inputs to produce fully sufficient ensembles, then more slowly composed the constituent expressions to produce ensemble models of improving necessity. The system has surprised its designers by pursuing evolutionary trajectories that were not anticipated, but which, upon inspection, proved to be following the sequentially superior building block route as prescribed by the multi-objective information theoretic fitness function.

For example, the system produced ensembles that expressed as $\{x_1 \otimes x_2, x_3 \otimes x_4\}$ (100% sufficient and 50% efficient), appeared to stall, then replicated one of the constituent expressions and grafted the copy onto the other as a leaf, improving efficiency and escaping the plateau. 3 more steps exactly like that would lead to an optimal solution; however, their probabilities are very small. Our selection may be *too neutral*: when several low probability consecutive mutations are required, and each is evaluated as neutral, there is no fitness arrow leading along the path; in true Pareto multi-objective optimization, likelihood of one tournament entrant dominating another is exponentially small in the number of objectives.

This is essentially a negative result: with similar numbers of fitness evaluations, we failed to achieve 4-parity, as Aguirre and Coello reported having done. However, we feel the problem is not in the use of information theoretic fitness but rather in some other aspect of the mechanisms of our GP. We have not given up on this problem, but have temporarily set it aside in favor of some others.

Despite not achieving 4-parity, analyzing the evolutionary trajectories that lead most of the way to it has yielded some insights. Gene *duplication* (expression duplication within an ensemble) has been found to be important, both to protect against gene deletion or disruption, and to prepare for subtree composition. *Recoding* also has been found to be important, both to provide composable representations and to transform model outputs into target coordinates. The fraction of target entropy explained only by synergy among inputs and the number of inputs that must be considered jointly to detect that synergy were found to be strong indicators of problem difficulty.

6.3 Korns' test functions

While the application of information theoretic techniques to discrete problems is straightforward, several practical questions were raised regarding its application to continuous valued data sets. To answer these questions, we adopted a continuous valued data set synthesized for testing GP regression systems [55]. It is generated by several multiple-input functions applied to pseudo-random inputs. These functions pose severe difficulties for GP or other regression techniques: they depend more strongly upon non-linear than linear interactions among multiple inputs; and they have cyclic dependencies upon the inputs, including both smooth covariations and abrupt behavioral mode switches.

The first question we address is the ability of the information-theoretic approach to distinguish relevant inputs from irrelevant inputs. Hence, unlike [55], which regressed the function outputs against all of, and only, their actual inputs, we also generated additional pseudo-random variables (“red herrings”), not actually used as inputs. From [55], we chose the 5 input versions of the 3 hardest problems:

- *cyclic* (y_0) is difficult due to periodic but mostly smooth functions
- *mixed* (y_1) compounds the difficulty with periodic abrupt mode switches
- *ratio* (y_2) further compounds difficulties with division by periodic functions

$$
\begin{aligned}
&y_{0a} = x_0 \sin x_0 \\
&y_{0b} = x_1 \cos x_1 \\
&y_{0c} = x_2 \tan x_2 \\
&y_{0d} = x_3 \sin x_3 \\
&y_{0e} = x_4 \cos x_4 \\
&y_0 = 14.65(1 + y_{0a}) - 6.73 y_{0b} \\
&-18.35 y_{0c} - 40.32 y_{0d} - 4.43 y_{0e}
\end{aligned}
$$

cyclic

(94)

$$
\begin{aligned}
&y_{1a} = \lfloor x_0 \rfloor \bmod 4 \\
&y_{1b} = 1.57 \log|x_0| - 39.34 \log|x_1| \\
&+ 2.13 \log|x_2| + 46.59 \log|x_3| + 11.54 \log|x_4| \\
&y_{1c} = 1.57 x_0^2 - 39.34 x_1^2 \\
&+ 2.13 x_2^2 + 46.59 x_3^2 + 11.54 x_4^2 \\
&y_{1d} = 1.57 \sin x_0 - 39.34 \sin x_1 \\
&+ 2.13 \sin x_2 + 46.59 \sin x_3 + 11.54 \sin x_4 \\
&y_{1e} = 1.57 x_0 - 39.34 x_1 \\
&+ 2.13 x_2 + 46.59 x_3 + 11.54 x_4 \\
&y_1 = case(y_{1a} of\, 0 : y_{1b}, 1 : y_{1c}, 2 : y_{1d}, 3 : y_{1e})
\end{aligned}
$$

mixed

(95)

$$
\begin{aligned}
&y_{2a} = \lfloor x_0 \rfloor \bmod 4 \\
&y_{2b} = \frac{1.57 x_0}{39.34 x_1} + \frac{39.34 x_1}{2.13 x_2} \\
&+ \frac{2.13 x_2}{46.59 x_3} + \frac{46.59 x_3}{11.54 x_4} \\
&y_{2c} = 1.57 x_0 + 39.34 x_1 + 2.13 x_2 + 46.59 x_3 + 11.54 x_4 \\
&y_{2d} = \frac{1.57 \sin x_0}{39.34 \tan x_1} + \frac{39.34 \sin x_1}{2.13 \tan x_2} \\
&+ \frac{2.13 \sin x_2}{46.59 \tan x_3} + \frac{46.59 \sin x_3}{11.54 \tan x_4} \\
&y_{2e} = -39.34 \log|x_1| + 2.13 \log|x_2| \\
&+ 46.59 \log|x_3| + 11.54 \log|x_4| \\
&y_2 = case(y_{1a} of\, 0 : y_{1b}, 1 : y_{1c}, 2 : y_{1d}, 3 : y_{1e})
\end{aligned}
$$

ratio

(96)

As in [55], the log function is made safe by adding .000001 to the absolute value of its input. Also as in [55], observations of function outputs y_k are corrupted by uniformly distributed gain noise with a default noise weight w of 20% according to

$$\hat{y}_k = y_k(1 + rand(-w,+w) \tag{97}$$

We calculated various normalized MI terms involving the function outputs, actual inputs and spurious ‘inputs’. As the scale of these indicators is an artifact of their calculation, affected by noise, computational precision and other factors, we regarded only relative rankings based on those values, not the absolute values themselves. We calculated how well those input selection rankings, and linear correlation based rankings, performed relative to ideal rankings. We applied several single input functions (cos, sin, mod) that ‘fold’ the data (thereby destroying information) and again ranked expressions.

In Tables 11 through 14, the total number of inputs per ensemble in the entropy calculation is m, the target function is f, the relative ranking performance score of correlation is r and the relative ranking performance score of sufficiency matching on n or more of the m inputs in the ensemble is s_n. These scores are the average of the ordinal positions of correct sublists of inputs, subtracted from that average in a random ranking, normalized by that difference in an ideal ranking; that is, a score of 1 is ideal, 0 is random and -1 is exactly backwards. The first experiment was repeated 6 times with 2^{16} points and 6 times with 2^{20} points. Little variation was observed between runs. Results were averaged to yield the tables.

In our first experiment, 10 RVs x_0 through x_9 were generated, and the first 5 were used as inputs to each of the 3 functions. In [55], one million points were generated, with each RV uniformly distributed in [-50..+50]. Initially, to limit time and memory requirements, we generated only 2^{16} points; to limit the information loss, in this reduced data set, due to the folding functions invoked (cos, sin, tan, mod, squaring), range of each RV was restricted to [-12.5..+12.5].

Single inputs are ranked using correlation and sufficiency in Table 11. Compared with linear correlation, information theoretic sufficiency worked slightly better on y_1 (*mixed*), significantly better on y_2 (*ratio*) and dramatically better on y_0 (*cyclic*).

The first experiment was repeated with 2^{20} (just over one million) data samples in [-50..+50], limiting time and memory requirements by considering only single inputs and pairs rather than larger ensembles. The larger data set confirmed the results observed with the smaller, and improved average performance of both correlation and sufficiency, except correlation on y_0 (*cyclic*), which it worsened. Single inputs are ranked using correlation and sufficiency in Table 12 and both single inputs and pairs are ranked using only sufficiency in Table 13.

Returning attention to the original data set of 2^{16} points, ensembles up to the correct size (5 inputs) and one size too large are ranked using only sufficiency in Table 14. Ranking variously sized input ensembles, sufficiency achieved on average more than 90%, and at the correct ensemble size (5) more than 99%, of ideal performance.

In the second experiment, again 10 RVs were generated, of which 5 were used as inputs to each function; but instead of using the same 5 the same way with each function, different assignments of variables to function inputs were made, as shown in Table 15. 2 RVs are ‘red herrings’, not used as inputs to any of the functions; 2 RVs are used as inputs to all the functions; 3 RVs are each used as an input to only a single function; and 3 RVs are each used as an input to all but one of the functions. The data was generated this way to evaluate information theoretic indices as detectors of dependencies between unlabelled observables where input/output relationships are *a priori* unknown.

Figures 19 and 20 are plots of sorted lists of logarithms of relative strengths of all pairwise linear correlations and information theoretic sufficiencies. They both exhibit distinct noise floors. We consider all strengths above their respective floors to be identified apparent dependencies. We compare correlation with sufficiency on the basis of their correct and incorrect identifications.

On y_1, correlation correctly identified 4 of 5 inputs, missing d_6 (its x_2). On y_2, it had 2 false negatives, d_4 (its x_0) and d_{7b} (x_4); plus it had a false positive, d_3. On y_0, it failed to identify any inputs at all. Correlation also flagged the mutual dependency between y_1 and y_2. It showed one other spurious dependency, between d_{0b} and d_6. Overall it showed 10 dependencies above its noise floor, of which 8 were valid, out of 18 actual dependencies in the data.

Sufficiency correctly identified all 5 inputs of all 3 functions, with no false negatives and no false positives. It flagged all 3 pairwise mutual dependencies of the 3 functions. Sufficiency showed one spurious dependency, between d_2 and d_3; its apparent strength was barely above the noise floor, and less than that of all real dependencies. Overall it showed 19 dependencies above its noise floor, of which 18 were valid, those being all 18 actually present in the data.

In the third experiment, RVs were generated as in the second, and to their set was added cos(d_1), sin(d_1), d_1 mod 4, cos(d_2), sin(d_2), d_2 mod 4, cos(d_4), sin(d_4) and d_4 mod 4. These are functions that 'fold' the inputs, thereby destroying information, applied to the 3 RVs that are each used in one test function only. A different 2 of these 9 terms appear in each of the 3 test functions.

The increased number of dependencies degraded detection of relevant inputs by both correlation and sufficiency; this argues for input selection as a separate pre-processing stage. On y_1, correlation and sufficiency both found dependency on d_2 mod 4, but only correlation on sin(d_2). On y_2, only sufficiency found dependency on d_4 mod 4 and sin(d_4). On y_0, only sufficiency found dependency on cos(d_1), and neither method on sin(d_1). Altogether, correlation detected 2, and sufficiency detected 4, of the 6 dependencies of the test functions on the intermediate terms introduced in the third experiment. Sufficiency also had a false positive, an apparent dependency of sin(d_4) on d_{7a}.

The foregoing experiments required calculation of many-dimensional copulae, which is computationally costly and degrades the sensitivity of the method, but is necessary for commensurate calculations of all the indices, to enable, for instance, fitness proportional selection. A less computationally costly and more sensitive alternative, that is adequate for tournament selection, is to calculate a lower-dimensional copula for each comparison. For instance, if the strength of a putative dependency of y_1 upon d_{0a} is to be compared against that of y_2 upon d_{7b}, only those variables need be considered, in a 4-dimensional copula.

In the fourth experiment, the 2 false inputs, 8 actual inputs and 3 outputs again were treated as 13 unlabelled observables. All their 78 pairwise potential dependencies were considered. Of the 3003 binary tournaments possible between strengths of putative dependencies, only 1080 were considered: those between false and actual dependencies, treating this as a classification problem; relative strengths of 2 false or 2 actual dependencies were not compared.

On these 1080 classifications, linear correlation was correct 644 times and incorrect 436 times, for an accuracy under 60%, corresponding to a performance score (relative to ideal vs random classifications, as in the relative ranking performance scores of the first experiment) under 0.20; whereas information theoretic sufficiency was correct 1059 times and incorrect 21 times, for an accuracy over 98%, corresponding to a performance score over 0.96.

Another way of looking at these results is to consider, for each of the 78 strengths, how many tournaments it would win and how many it would lose against the other 77, and using those statistics to rank them; this is the approximation of fitness proportional by binary tournament selection. Correlation correctly ranks 5 actual dependencies above its first false positive; it ranks the 18th and last actual dependency in position 76 of 78; if the top 1/3 of its rankings were selected, it would have 17 false positives and 9 false negatives. Sufficiency correctly ranks 14 actual dependencies above its first false positive; it ranks the 18th and last actual dependency in a tie of positions 23 through 25 of 78; if the top 1/3 of its rankings were selected, it would have 8 false positives and 0 false negatives.

The four experiments above were reported at GPTP-2008. Participants there, while satisfied that these results had favorably answered the question of applicability of information theoretic techniques to input selection, asked for more evidence of their applicability to fitness evaluation of intermediate forms (neither simple inputs nor final evolved models) during GP runs. Complex subexpressions of the defining equations, that are higher-order building blocks, requiring only a single additional layer of interconnection to correctly complete reconstruction of those equations, may be regarded as “penultimate forms”. For our test problems, these have been broken out as constituent top-level subexpressions in Equations 94 through 96.

In a fifth experiment conducted after the workshop, for each of the 3 test problems, on 2^{16} data points, an 11-dimensional copula (including all 5 raw inputs, all 5 penultimate forms and the target output), was calculated. For all 3 test problems, the estimated sufficiency, efficiency and similarity against the target output, of the ensemble of all 5 penultimate forms, were stronger than the corresponding indices of the ensemble of all 5 raw inputs.

Likewise, in all cases but one, the weakest of the 5 sufficiencies of the penultimate forms was stronger than the strongest of the 5 sufficiencies of the raw inputs. The one exception was on mixed, where raw input x_3 appeared to be a slightly stronger predictor of y_1 than did penultimate form y_{1d}. We “zoomed in” by calculating, on 2^{18} data points, a 3-dimensional copula involving only those 3 variables; this higher resolution measurement correctly ranked the penultimate form above the raw input on all 3 information theoretic indices.

Those familiar with information theory may object that an increase in sufficiency of the ensemble of all 5 penultimate forms over that of the ensemble of all 5 raw inputs violates the Data Processing Inequality. In response we can only remind the reader that Shannon’s entropy is defined only for discrete data, and that our estimated analog of it for continuous data is sensitive to the locations and widths of data bins relative to topological foldings of the data. This is harmful, in that intuitions learned using information theoretic techniques on discrete data cannot always be relied upon when working with continuous data; but it is also helpful, in that index values improve as topological foldings are discovered by the GP and reflected in the evolving genotypes.

We conclude that information theoretic sufficiency of continuous valued data can indeed be calculated robustly and efficiently. On this test data, it outperformed linear correlation in ranking single inputs, approached ideal performance in ranking input ensembles, and exhibited a distinct and meaningful noise floor; thus it is an appropriate pre-processor for GP input selection.

Sufficiency also correctly identified (albeit with a couple of false negatives and a single false positive) simple expressions (derived from relevant inputs by filtering with single input, information destroying, topologically folding functions) that are basic building blocks for starting to regress the test problems; thus it is also an appropriate fitness function early in GP runs. It also correctly ranked “penultimate forms” above raw inputs, and ensembles of penultimate forms above ensembles of raw inputs; thus it also appears promising as a term for a multi-objective fitness function later in GP runs.

Having presented some experimental evidence, on hard synthetic problems, of MI utility both early and late in EA runs, and MI heritability across genetic variation, Chapter 6 has supported the claim that MI enhances evolvability; next in Chapter 7 we present empirical evidence of MI utility for input selection on hard real world problems.

7 Applications

A full-blown EA based upon the principles laid out herein has not yet been applied to real world problems; however, the information theoretic indices developed herein have been applied to input selection in such problems. This Chapter supports our claim that our indicators have utility with evidence gathered on real world problems. In Section 7.1, we describe findings of dependencies among industrial plant process variables. In Section 7.2, we relate findings of dependencies between international political leadership attributes and regime stability. In Section 7.3 we outline several diverse current applications.

7.1 Industrial plant process variables

At a complex industrial plant[50] (with long material in-process times, material feedback loops, numerous sensors, many automatic control loops and human operators with manual override capabilities), an occasional minor process problem persisted, despite efforts to correct it. A data set was extracted from the plant information system, containing all the variables that were *a priori* thought potentially related: whether as advance warnings of the onset of a problem instance; or as usable in a control strategy to keep the process variable of interest in the desired range. In support of efforts by others at modeling the plant and developing such a strategy, the authors applied information theoretic techniques to identify dependencies and the time lags that maximize them.

The data set spanned 4 months, sampled at 5 minute intervals; it included time stamps and the values of process variables, controller outputs, set points and

[50] Which must remain nameless, per the wishes of the plant management that provided the data we analyzed.

status flags. In the raw spreadsheet, there were over 32k rows and 78 columns. Many ‘variables’ did not actually vary significantly during the period. Others clustered tightly at or near a small number of distinct values, with a single value appearing in over half of the samples. Although there might be dependencies involving these non- or minimally-varying variables (which might be identifiable in data from a longer or different period during which they varied more), to prevent conclusions being drawn in which there could be no statistical confidence, these variables were excluded from this analysis.

To provide a baseline against which putative dependency strengths could be compared, 6 columns of uniformly distributed, computer generated, pseudo-random numbers (“red herrings”) were inserted; this is computationally less expensive than the techniques of Sections 4.3 and 4.4, albeit less informative. The resulting 57 columns of data were processed using a version of the copula based algorithm of Section 4.2.

The strength of the dependency of the process variable of interest on each other variable (sufficiency of the latter against the former) was estimated over a range of time leads and lags; their maximum values and the lags at which they peaked were recorded. As expected, the estimated sufficiencies of the “red herring” random variables were tightly clustered, weaker than those of all the actual process variables, and generally did not vary significantly with lag. Estimated sufficiency of an actual process variable over a wide sweep in lag is shown in Figure 21 and over a narrow sweep in lag is shown in Figure 22. Wide sense behavior appears to be approximately exponential decay, as expected for a system with a finite number of parameters [11]. Narrow sense behavior appears to be decay plus random scatter; presumably the scatter is due to observational noise, and would decrease with larger numbers of samples, but this has not been confirmed.

The sufficiency estimates of several variables exhibited apparently cyclical variations over lag, of multiple distinct periods: these are oscillations, not in the variables themselves, but in the strengths of the estimated dependencies of the process variable of interest upon them. The significance of these oscillations is unknown: but it is credible that the tightness of a physical coupling may vary periodically; and the plant engineers are already aware of the existence of long period oscillations. An example is shown in Figure 23.

Basic statistics were calculated on the sufficiencies at their respective maxima. Those variables with estimated strengths exceeding the median were listed. Assuming real dependencies would have significant variation with lag, variables that lacked it were dropped. From sets of related variables, controller outputs were selected when present, yielding a list of 10 control strategy candidates.

One of these was declared physically impossible by the plant engineers, given the existence of a control loop that should eliminate any such dependence. Upon inspection, it was found that this control loop was not functioning reliably; so the mere existence of an information theoretic dependency proved itself to be informative regarding an unlooked-for fault.

Pairs of variables were then considered. The joint sufficiencies of all pairs of process variables against the variable of interest were estimated, and compared against either of those variables taken jointly with a “red herring”. 2 red herrings jointly were estimated as over 50% sufficient to model the process variable of interest; this indicates that the sample data set is too small to reliably measure joint modeling sufficiency of larger ensembles. 5 pairs stood out.

Unsurprisingly, most of the strong pairs included one of the stronger individuals, but the 5th strongest pair did not; synergy added to the moderate strengths of its constituents. It might have been expected that the pair comprised of the 2 variables whose individual dependencies were strongest would have been among the strongest pairs: but those variables proved substantially redundant – their joint sufficiency was less than the sum of their individual sufficiencies; this relates to a common pitfall in feature selection. Feature selection is typically performed greedily: the single input, feature or model that provides the greatest information gain is added to the ensemble, iteratively until some stopping criterion is met; this procedure implicitly assumes independence of features, and is known to fail to identify ensembles of synergistically interacting features when one or more of those features provides little information absent the others. This could theoretically be avoided by evaluating the joint MI between the target and each ensemble in the power set of the inputs, but this is computationally infeasible in all but very small problems. A more feasible approach is to try ensembles of gradually increasing size, starting with singleton ensembles, but even this becomes quite costly if ensembles of more than a couple of inputs are needed from a large set of observables. [52] make the seemingly reasonable suggestion “However, one can employ heuristics, for example, higher-order interactions are unlikely in the absence of lower-order interactions among the same attributes.” This is contradicted by the N-parity problem, where any ensemble of fewer than N inputs conveys no information.

These calculations were all performed using the optimal lags determined in the univariate analysis; stronger joint dependencies might be discovered if lags of both variables were jointly scanned, but that would require much more CPU time.

Joint sufficiency of 3 actual variables was not significantly higher than joint sufficiency of 3 red herrings (indeed, in many cases it was lower). This is because actual variables tend to be redundant, whereas pseudo-random numbers are (to a high order approximation) independent. With 3 random coordinates, each of which is binned into at least 32 bins (5 bits), a total of 32k bins (15 bits) are uniquely specified; as this data set has only about 32k points, this uniquely localizes each data point. The adaptive algorithm for calculating information dimension peaked at 64 bins in the investigation of dependency on 2 variables jointly, and at 32 bins in the unsuccessful attempt involving 3 variables. Thus, to determine whether combinations of 3 or more variables are able to improve significantly upon the predictive modeling power of smaller ensembles, a larger data sample would be needed.

Using the copula based algorithm of Section 5.2, the minimum number *N* of samples required, for a given level ℓ of the pyramid to yield statistics with nonzero noise margin, is easily calculated from the level of the pyramid (counting from the top ℓ=1, and corresponding directly to the log base 2 of the number of bins along each dimension at that level) and the nominal dimension $\mathcal{D}_m$ of the measurement vector (number of dimensions of the sparse array at each level of the pyramid).

$$\lg N \geq \mathcal{D}_m \ell + \lg 3 \qquad \textbf{(98)}$$

To find non-linear dependencies, it is necessary to go to at least the 2nd level of the pyramid ℓ=2, and usually further down than that to find any fine structure.

In consultation with the plant engineers, it was determined that outliers had dominated some of the dependency strength estimates, and that many of these outliers were due either to changes in the variety of product being processed (which involve changes to various controller set points throughout the mill) or to the plant going down (for planned maintenance or due to problems) and coming back up again. Several such periods were manually identified and a threshold value for one of the process variables was determined, such that if the measured value is below the threshold, it can be inferred with high confidence that the plant is down. Predicting down-time is of great importance to the mill: associated costs and other consequences can be minimized if operators are given adequate warning of an imminent event; and some down-times may be avoided if operators are alerted so they can intervene to correct a problem before it shuts the plant down. Plant down-time, especially unplanned, is expensive.

Process variables were again surveyed: searching for any that exhibited strong information theoretic sufficiency against the continuous variable that, when thresholded, served as a proxy for an explicit plant down-time indicator; then verifying discovered strong sufficiencies against the continuous variable, by calculating sufficiency against the corresponding discrete variable (the flag indicating the continuous value was below its threshold).

This yielded two interesting results. First, the variable upon which the plant down-time indicator proxy was found to be most strongly dependent, was one of those upon which the variable of interest in the process problem investigated previously, was most strongly dependent; this makes sense if a common cause can lead to that specific problem and/or plant down-time. Second, the time lag at which that variable exhibited its strongest coupling with plant down-time was verified as physically meaningful.

Finding known relationships partially validated the information theoretic indices: failing to find these relationships would have constituted false negatives. Finding a previously unknown relationship (due to the failed control loop) and a previously unknown phenomenon (the oscillations in coupling strengths) also verified the real-world utility of the indices: they provided previously unknown useful information.

7.2 International political leadership data

In the context of the interdisciplinary Forecasting Change and Societal Threats (FORECAST) working group, the Moynihan Institute of Global Affairs at the Maxwell School of Citizenship and Public Affairs of Syracuse University provided a data set for analysis. The objective was to identify attributes of leadership that make a regime crisis-prone. This information might be used to allocate assets, whether for intelligence gathering or efforts to stabilize or destabilize a regime. The data had already been analyzed using conventional statistics, which had found little of interest; it was hoped that non-linear dependencies might be found using the information theoretic techniques described hereinabove.

In the raw spreadsheet, there were 121 rows and 37 columns. Some of the columns were labels rather than data variables and so were not analyzed. Some were aliases so for each set of differently named identical variables only one was analyzed. 7 rows had “no material” for most of the discrete categorical variables and 9 rows had “leadership in dispute” and therefore no entries for most of the discrete categorical variables; these were all omitted from the analysis.

4 columns were absolute dates. The first was subtracted from each of the other 3 to yield durations or relative offsets. This yielded quantitative, if not truly continuous, variables, which were inspected together with 2 other quantitative columns. Histograms showed clusters in 3 of the continuous variables; these may be discretized to analyze consistently with the discretes in future work.

27 rows dealt with regimes still in power, for which certain entries are not applicable. For the “continuous” (quantitative variable) analysis, a special value was inserted in place of such missing data; more sophisticated techniques should be used in future work. For the discrete (categorical variable) analysis: first calculations were performed with those rows omitted; then calculations were repeated, including those rows, but omitting the columns with missing entries;

finally calculations were repeated, including those rows and substituting a special value for the missing entries; results of these 3 runs did not differ significantly.

Pairs of discrete variables with strong dependencies were found. Some proved trivial: certain spreadsheet columns had been computed from other columns. Others appeared to be non-trivial and strong, but some had low statistical confidence, due to small samples and large alphabets (numbers of categories).

Two variables whose dependence upon others is of interest are the reason a regime started and the reason that regime ended. 3 variables were identified that jointly reduced the uncertainty in the 2 variables of interest by over 2/3. This appeared to confirm a theory of one of the social scientists in the group.

The integer encodings of categorical variables were exhaustively permuted: this does not affect entropy or MI; it was intended to identify alternative encodings in which linear correlation was strong. The dynamic ranges of correlation coefficient values for some pairs of variables using different encodings were dramatic; other relationships did not manifest as correlation regardless of encoding, so they are presumably more complex conditional distributions. The only *a priori* sensible relationship that was made manifest by permutation was correlation between the reasons a regime began and ended: “live by the sword, die by the sword”.

The primary result of this effort was the development of the nested resampling technique for assessing statistical confidence described in Sections 4.3 and 4.4, which was necessitated by number of variables (25 discretes) being so large relative to the number of usable complete samples (78). This effort also led us to visualization techniques [52] to present findings. Figure 24 is a simplified information diagram, showing high confidence, strong sufficiencies of variable pairs against the reason a regime ended. Nodes are labeled with the names of variables and the individual sufficiencies of the corresponding variables against the reason a regime ended. Edges are labeled with the joint sufficiencies of the variables on their endpoints against the reason a regime ended.

information theoretic indices developed to guide input, survival, reproductive and final ensemble model selection in EAs, are good at least for the first of these purposes, on some real world data: they discovered not only suspected (political variable) and known (process variable) relationships (validating the indices), but also a previously unknown relationship (due to the failed control loop) and a previously unknown phenomenon (the oscillations in coupling strengths).

7.3 Current applications

Katya Vladislavleva Smits provided a Human Development Index (HDI) data set. Like the FORECAST international political leadership data set, it is social sciences data, with many issues. There are 147 variables but only 141 samples. Some variables are continuous, some are quantitative but discrete, others are categorical (class labels). Many nominally independent variables have trivial relationships (columns calculated from other columns per hypotheses).

A goal is to predict future values of the HDI from current and past values of other variables. Vladislavleva is applying GP, specifically DataModeler (Kotancheck[51]),

[51] A commercial data analytics package, based on Pareto multiobjective ensemble model GP, written in Wolfram Mathematica by Mark Kotanchek, about to be released by evolved-analytics.com.

which provides *inter alia* powerful but essentially heuristics tools for input (GP terminal set) selection. The data is being analyzed on an information theoretic basis to objectively measure the total statistical dependence of HDI upon single and pairs of variables (there are too few samples to consider larger ensembles). Incorporation of our indicators into DataModeler has been discussed.

Department of Defense Small Business Innovation Research (DOD SBIR) solicitation topic OSD10-L08 “Real-time Resource Allocation Co-Processor” solicited development of co-processing technology to optimize allocation of communication, computation and storage resources of and for mobile devices. Critical Technologies Inc. (CTI, this author’s corporate affiliation) proposed an Agent Resource Grant Optimizing System (ARGOS) based in part on our Evolutionary Learning with Information Theoretic Evaluations of Ensembles (ELITE^2, the name we have given to our information theoretic GP approach). The Office of the Secretary of Defense (OSD) accepted the proposal. Phase I was executed by CTI with assistance from a graduate student supported by the Syracuse University Center for Advanced Systems and Engineering (CASE). The project was managed technically by the Air Force Research Laboratory’s Information Directorate (AFRL/RI) in Rome NY. A Phase II proposal has been submitted by a team of CTI, SU, AgileX LLC, iAppFusion and Virginia Tech.

Department of Defense Small Business Technology Transfer (DOD STTR) solicitation topic AF10-BT09 “Dynamic Cross-layer Routing Using Cognitive Spectrum Allocation Techniques” solicited design and assessment of innovative methods to create adaptive cross-layer wireless networking protocols to achieve network resiliency in contested RF spectra. A team of CTI and Syracuse University proposed a CYber Cross-Layer Optimizing Publish-Subscribe (CYCLOPS) cognitive networking architecture based in part on ELITE[2]. The Air Force Office of Scientific Research (AFOSR) accepted the proposal. Phase I is being executed by the team and managed by AFRL/RI.

One common theme of current applications is handling mixed case data sets in a manner that enables meaningful comparisons, e.g. of the dependence of a discrete variable upon another discrete and a continuous variable. Some need measures of the user-, use-, time- and situation-dependent utility of information.

The HDI analysis and ARGOS and CYCLOPS contracts show that independent third parties have assessed our information theoretic approach as promising.

Chapter 7 had nothing to do with EAs, but it did present evidence that the information theoretic indices developed to guide input, survival, reproductive and final ensemble model selection in EAs, are good at least for the first of these purposes, on some real world data: they discovered not only suspected (political variable) and known (process variable) relationships (validating the indices), but also a previously unknown relationship (due to the failed control loop) and a previously unknown phenomenon (the oscillations in coupling strengths). Next, in Chapter 8, we outline promising areas for closely related future work.

8 Future Work

Closely related prior work is almost non-existent; what little we have found, we have cited above. Loosely related prior work is extensive and diverse; so much so that we have relegated discussion of related work as such to an appendix. Future work, in the sense of potential applications, is likewise diverse, so we do not address it. Future work, in the sense of theoretical development, is both easier to identify and fundamentally important, so a summary follows.

In Section 8.1, we formalize a conjecture regarding heritability of information. In Section 8.2, we develop a conjecture regarding generalization of learned models. In Section 8.3, we propose an approach to maintaining evolvability dynamically.

8.1 Heritability conjecture

Fitness Distance Correlation (FDC, [53][106][45]) is widely accepted as a necessary but not a sufficient condition for a problem to be tractable to an EA; conversely, lack of FDC is accepted as a sufficient, but not a necessary, condition for EA problem hardness. FDC has limitations that are almost as well known as its benefits; in our opinion, these result primarily from it being calculated as an average over an entire landscape rather than differentially (or at least on scales small relative to any roughness in the landscape). We conjecture that FDC, generalized to correlate not phenotypic fitness with genotypic edit distance to a global optimum but rather phenotypic Rajski distance with genotypic Kolmogorov distance on arbitrarily small scales (directly characterizing search neighborhood differential structure), is a universal measure of problem difficulty for evolutionary learning.

For any other phenotypic fitness measure to guide an EA to an optimal (e.g. minimum MSE) solution, it must [indirectly] exploit mutual information between the available inputs and the target output, and must discard all model variation (entropy) that does not contribute to modeling the target output. This is directly measured by Shannon or Rajski distance between a model and the ideal model; but that retains the FDC weakness of traversing an entire fitness landscape. Zooming in on the local search neighborhood differential structure, we can use Shannon or Rajski distance between a model and an incrementally different (potentially better or worse) model; but this ignores the target input-output relationship. Replacing total with target relevant information distance, the SR measure or SRI gives us a divergence between arbitrary models that does reflect their usefulness in modeling the target relationship; but these are not true distances as they fail the triangle inequality, which is presumably important when evaluating how far 2 models are locally from each other and from the ideal model globally. Thus our conjecture has 2 problems: we have yet to identify a phenotypic Shannon information distance with all the required properties; and our genotypic Kolmogorov distance is not computable (although, e.g., Normalized Compression Distance (NCD) is).

Despite these difficulties we present several alternative formalizations of our conjecture. Each formalization may be regarded either as an absolute *statement* about the use of a particular set of genetic operators upon a particular data set that is either true or false, or as a *property* (with a corresponding predicate) that holds on some subset of the feasible points in the space of models accessible to the evolutionary system (again in the context of a particular data set).

$\mathrm{d}_{\mathrm{K?}} =$ some TBD Kolmogorov complexity based genotypic info distance
$\mathrm{d}_{\mathrm{S?}} =$ some TBD Shannon entropy based phenotypic info distance
$\delta =$ length of an arbitrarily small move (edit) in genotypic space
$\varepsilon =$ length of an arbitrarily small move in phenotypic space
$\mathbf{f}_i, \mathbf{f}_j, \mathbf{f}_k \in$ population of input $\rightarrow$ output model functions
$\mathbf{Z}_i, \mathbf{Z}_j, \mathbf{Z}_k =$ their corresponding outputs on the same input data set

$$\forall \varepsilon : \exists \delta : \mathrm{d}_{\mathrm{K?}}(\mathbf{f}_i, \mathbf{f}_j) \le \delta \Rightarrow \mathrm{d}_{\mathrm{S?}}(\mathbf{Z}_i, \mathbf{Z}_j) \le \varepsilon \qquad \textbf{(99)}$$

$$\mathrm{d}_{\mathrm{K?}}(\mathbf{f}_i, \mathbf{f}_j,) \le \mathrm{d}_{\mathrm{K?}}(\mathbf{f}_i, \mathbf{f}_k) \Rightarrow \mathrm{d}_{\mathrm{S?}}(\mathbf{Z}_i, \mathbf{Z}_j) \le \mathrm{d}_{\mathrm{S?}}(\mathbf{Z}_i, \mathbf{Z}_k) \qquad \textbf{(100)}$$

$$\lim_{\mathrm{d}_{\mathrm{K?}}(\mathbf{f}_i, \mathbf{f}_j,) \to 0} \mathrm{d}_{\mathrm{S?}}(\mathbf{Z}_i, \mathbf{Z}_j) = 0 \qquad \textbf{(101)}$$

Equation 99 is trivially true when ε is sufficiently large. Equation 101 is also trivially true at the limit point (identical functions will have identical outputs) but may be too demanding as it approaches the limit. Equation 100 appears to be in a “sweet spot” between the other 2, but is formulated in terms of order statistics, which are less precise. Regarding any of these or similar formulations as properties, we may ask how well they hold in the evolutionary model space.

Continuing our conjecture, we assert –

- The property must be *generic* or the problem is EA-hard. That is, at any point where the property does not hold, there must be another point, arbitrarily nearby, where the property does hold, or the EA will have points from which it can never get “on track” towards a solution.

- The property must be *likely* or the problem is EA-hard. That is, the measure of all points in the space where the property does hold, must not vanish as a fraction of the total measure of all points in the space.

Note that we make no assertions about when a problem is EA-easy: either condition failing is sufficient to make the problem hard; thus both conditions holding is necessary but not sufficient to make the problem easy. Note further that “hard” does not mean that an EA cannot quickly chance upon the globally optimum solution; it means that an EA cannot be relied upon to find that solution faster than random or enumerative search.

Can we relax the first condition: allow there to be a set of points where the property does not hold, that also lack near neighbors where it holds, provided the measure of that set vanishes as a fraction of the total measure of feasible points in the space? Doing this seems to weaken the conclusion from absolute to being only in probability one; practically signifying populations might need to be large.

The assertions above are qualitative. We may make them quantitative by modifying the second condition: a problem is hard to the extent that the measure of the set of points where the property does not hold is large; under some formal assumptions, we may say that the problem difficulty is proportional to the *probability* of the property (e.g. Equation 100) not holding.

All this may be considered in the whole evolutionary model space or in regions. There might be unfortunate regions within which the property does not hold, but other regions where it does, regions large or important enough for it still to be worthwhile to use an EA on the problem. In a multi-modal problem, individuals may fall in a basin of attraction of a local optimum, from which they may not escape to find the global optimum; we refer not only to that, but also to the possibility of trackless regions within which local optimization is doomed.

Our conjecture is related to the information landscapes of [13].

The foregoing conjecture is essentially of *heritability*: small genotypic changes such as are likely under recombination of similar parents or mutation (by the particular genetic operators whose neighborhood structure we are characterizing) should lead to small phenotypic changes (as measured by information distance between input-output behaviors of the phenotypes, not by any particular output objective function nor by any consideration of the structures of the phenotypes). Evolvability requires moderate heritability: weak heritability is too explorative and degenerates into essentially random search; very strong heritability is too exploitative and degenerates into essentially local hill-climbing.

In the proposal for this research, we suggested that a general proof of heritability of mutual information would be desirable. We now realize that heritability of MI is a function of both the genetic operators and the particular optimization or learning problem to which they are being applied. The question is not “Is MI heritable?” in general but rather “How heritable is MI, under these operators, on this problem?” The answer to that more specific question is an indicator of the potential tractability of that problem using that set of operators: if MI heritability is weak or extremely strong, the problem *will be* hard for that EA; if it is moderate, the problem *might be* easy for that EA.

8.2 Generalization conjecture

In the Probably Approximately Correct (PAC) framework of Valiant and Haussler [49], the test data is required to be drawn from the same distribution as the training data. This is unrealistic: generalization is required.

Can the Rajski dissimilarity between a model and the ideal model, on the training set, be somehow updated with a distance (or divergence) between the copulae of the training and testing sets, to estimate performance on the testing set?

The Rajski dissimilarity between a model and its ideal is the complement of our solution quality metric. It measures the performance (actually, the deficiency in performance, relative to ideal) of a model on a data set. Given a performance metric, it is desirable to be able to estimate its value on a testing set, from its value on its training set, together with some measure of the difference between the training and testing sets. Many such measures of difference between statistical distributions have been defined, and some are distances.

The notion is straightforward: if we know how far the training set is from the testing set, and we know how far a model is from ideal on the former, can we not calculate (or at least bound) how far the model will be from ideal on the latter?

To calculate performance, in our information spaces, we would need not just scalar distances, but vector displacements: how far one recoding-equivalence class of information sources lies from another *and in what direction*. Our ideal model **f** is such a vector, pointing from the recoding-equivalence class of the input data set, with measure $H(\mathbf{X})$, to the class of the input-output relationship, with measure $H(\mathbf{Z})=I(\mathbf{X};\mathbf{Y})$, giving **f** a Shannon distance length of $H(\mathbf{X}|\mathbf{Z})$ $=H(\mathbf{X}|\mathbf{Y})$ in some as yet unspecified direction. Our model $\mathbf{f}_j$ is also such a vector: its base is at the same point as the base of **f**; but its tip is at the class with measure $H(\mathbf{Z}_j)$, giving it a Shannon distance length of $H(\mathbf{X}|\mathbf{Z}_j)$ in some other unspecified direction.

In Figure 25, the vector between the tips of the **f** and $\mathbf{f}_j$ vectors has Shannon distance length $H(\mathbf{Z}|\mathbf{Z}_j)+H(\mathbf{Z}_j|\mathbf{Z})$, which normalized by their joint entropy yields their Rajski dissimilarity, the complement of our performance metric; but again, in a direction that we do not yet know how to specify. If we did know how to specify direction, we would also know how to specify location of a point in information space, according to some coordinate system of basis vectors. Given that ability, not only within but also between data sets, we could form the new ideal model **f'** on the testing set, again place the bases of the vectors corresponding to our model and its (new) ideal at the same point (which can now be arbitrary, as all we want is their vector difference), and find the length of the vector between their tips. This, when normalized, would give us the Rajski dissimilarity between our model and an ideal model on the testing set, the complement of which is our (estimate of) the performance of our trained model on the testing set.

To merely bound (rather than exactly calculate) performance of the trained model on the testing set, we would need a metric distance between training and testing sets that is commensurate with the Rajski (or Shannon) distance, so that we can use the triangle inequality to give us the upper bound on Rajski dissimilarity (lower bound on performance) and the reverse triangle inequality to give us the lower bound on Rajski dissimilarity (upper bound on performance). The Jeffrey and Kullback-Leibler divergences initially appear to be candidates, as they are commensurate with Shannon distance, but the KL divergence is not symmetric (which may or may not be a problem for this application) and neither of these divergences satisfies the triangle inequality (which is clearly required).

So for exact calculations of generalization performance, we need basis vectors, or for bounding performance, we need a metric distance between distributions that is commensurate with the Rajski dissimilarity or Shannon distance, but unfortunately we have as yet been able to identify neither of these.

We did make a foray into development of basis vectors: starting with vectors, from an origin defined as knowledge of nothing (maximum uncertainty about all RVs), pointing to each of the RVs; and calculating projections of each onto each, which should correspond to their MI, using the Law of Cosines. We immediately found that RVs with no MI, although pairwise independent (geometrically perpendicular), if they have interaction information (redundancy or synergy when considered jointly against another RV), are not fully orthogonal. This yields a geometry in which the length of the 3rd side of a triangle is given not by

$$|c|^2 = |a|^2 + |b|^2 - 2|a||b|\cos\theta_c \qquad \textbf{(102)}$$

(where lengths are given as norms of vectors) but rather by

$$|c| = |a| + |b| - |a||b|\cos\theta_c \qquad \textbf{(103)}$$

which seems to corresponds to Manhattan (city block) distance, so there may be hope in some non-Euclidean space, with help from a better mathematician.

We suspect it will be necessary to reconcile the information geometry of Amari, with the information geometry of Crutchfield, with the EA geometry of Moraglio, with the respect and assortment considerations of Radcliffe, to fully geometrize important EA considerations such as heritability, generalization and evolvability.

8.3 Evolvability conjecture

In on-line learning, we may consider all the past experiences from which the system has learned to constitute the “training set”, and the very next experience encountered to be the “testing set”, in the nomenclature of Section 8.2. Thus generalization performance of a trained and tested machine learning system corresponds to performance of on-line learning, optimization and search.

Self-adaptive EAs tune their learning algorithms based on experience. The simplest example of this is the Covariance Matrix Adaptation Evolution Strategy [46], that updates its mutation operator to correspond with the experienced covariance between mutation steps and fitness function values. The idea is quite similar to the use of “momentum” in neural network learning rules, where weight updates [51] may include: in addition to the primary term, intended to reduce the particular weight's contribution to the most recently experienced error; a secondary term, that is typically an exponential moving average of previous weight update steps, intended to keep the weight moving quickly in a direction that seems to correspond to the error gradient based on recent experiences.

Evolvability depends, at least in part, on successful generalization, from what we have learned about the search neighborhoods of the past (both good and bad experiences with genetic operators), to deployment of genetic operators that will yield effective search neighborhood structures in the near future.

A defeatist or lazy attitude towards this is to oversimplify the implications of the “No Free Lunch” theorems [111] to mean that the generalization performance of all algorithms is the same. A more constructive attitude is to recognize that the assumptions of NFL are rarely fully satisfied in practice and to exploit the extent to which the fitness landscape of a real problem has some smoothness or other somewhat consistent structure.

The same difficulty is encountered in adaptive compression: Kolmogorov showed that most of the strings in the universe of all possible strings are incompressible; yet in the universe of strings that arise in practice, we compress many each day. Compressors typically build a dictionary based on sequences they have seen: when they encounter a novel sequence, they can't compress it very well; but after seeing it repeatedly, they conclude it is likely, and assign it a short proxy.

To maximize evolvability (and thus generalization and on-line performance), we wish to design evolutionary learning systems that self-adapt to maintain moderate heritability of information (balancing exploitation and exploration). The landscape metaphor offers us 2 heuristics. First, when we are on a flat plain, take long steps, and when we are on a rough surface, take small ones, as we wish to climb as rapidly as will not entail a large chance of falling off a cliff. Second, if we suddenly discover that someone has equipped us with 7 league boots, reduce the vigor with which we take each step, or if we have been hobbled, increase that vigor, to keep step length in accord with the first heuristic. These 2 heuristics are commonly muddled in an attempted direct mapping from genotype to fitness.

The foregoing metaphor can be operationalized using the phenotypic behavioral and genotypic information distances defined previously. The first heuristic deals with the change in phenotypic solution quality resulting from changes in phenotypic behavior (and often structure); in terms of our metrics this is a slope

$$\frac{\left|\mathrm{D}_{\mathrm{S}}(\mathbf{Z},\mathbf{Z}_i)-\mathrm{D}_{\mathrm{S}}(\mathbf{Z},\mathbf{Z}_j)\right|}{\mathrm{D}_{\mathrm{S}}(\mathbf{Z}_i,Z_j)} \tag{104}$$

where we evaluate this with 2 models in close proximity with each other, such as a parent and one of its offspring produced by mutation only (no recombination). The numerator is the absolute difference, between 2 models, in their respective dissimilarities to the ideal model; the denominator is the dissimilarity between the 2 considered models.

The second heuristic deals with the change in phenotypic behavior (and often structure) resulting from changes in the genotype; this is also a slope

$$\frac{\mathrm{D_S}(\mathbf{Z}_i, \mathbf{Z}_j)}{\mathrm{D_K}(\mathbf{f}_i, \mathbf{f}_j)} \qquad \textbf{(105)}$$

where again the evaluation should involve 2 models in relative proximity, and practical applications will need a computable metric such as NCD in place of $\mathrm{D_K}$.

Of course it is possible to multiply these 2 ratios, canceling one term to yield

$$\frac{\left|\mathrm{D_S}(\mathbf{Z}, \mathbf{Z}_i) - \mathrm{D_S}(\mathbf{Z}, \mathbf{Z}_j)\right|}{\mathrm{D_K}(\mathbf{f}_i, \mathbf{f}_j)} \qquad \textbf{(106)}$$

but this may lead to taking excessively large steps in regions of the search space where the solution quality changes slowly with respect to phenotypic behavior; we want our models to change relatively slowly, so that the population will contain models structurally similar enough to be compatible mates.

We conjecture that evolvability of information would be dynamically maximized in a self-adaptive EA that used $\mathrm{D_S}$ and an approximation to $\mathrm{D_K}$ as follows.

Based on recent measurements of the phenotypic solution quality change induced by phenotypic behavioral distance traversed, and the latter as a consequence of genotypic variational steps, adapt the genetic operators (or probabilities of applying particular ones, or step sizes passed to any that parameterize it) so that the likelihood of taking a given mutation step is inversely related to the phenotypic effects it is expected to have, and the most likely mutation steps are expected to have small but non-negligible phenotypic effects.

This procedure makes what we believe to be the weakest possible assumption about the adaptive landscape: that abrupt changes in its smoothness are rare.

This final conjecture, while theoretically motivated, requires no further mathematical gymnastics. It could be experimentally validated by taking an existing evolutionary learning system that already self-adapts its step size, replacing its genotypic edit distance function with NCD, replacing its fitness function (solution quality measure) with $D_S(\mathbf{Z},\mathbf{Z}_j)$, and inserting an additional phenotypic behavioral distance measurement using $D_S(\mathbf{Z}_j,\mathbf{Z}_k)$.

9 Summary of Contributions

This dissertation has made contributions in 3 clusters – indicators, algorithms, corollaries to Price's Theorem – summarized in the sections of this brief chapter. It has also made several contributions that lie outside those clusters. We have:

- ► demonstrated how to adopt an information theoretic perspective on EC;
- ► identified information transmission channels implicit in evolutionary learning;
- ► introduced nested resampling to estimate confidence in small sample statistics;
- ► defined a measure of total redundancy that generalizes total correlation;
- ► applied information theory to analysis of diversity, fitness and deception in EAs;
- ► presented evidence for heritability of information theoretic solution quality indices;
- ► presented evidence for evolvability using such indices for selection in GP;
- ► evaded deception and poor scaling on a chaotic time series prediction problem;
- ► correctly selected relevant inputs on a complex continuous function benchmark;
- ► discovered relationships in a high dimensional industrial process data set;
- ► verified hypothesized relationships in a small sample political data set;
- ► formalized conjectures regarding information distances in evolution.

9.1 Indicators

We have developed commensurate indicators of diversity and fitness (both solution quality and reproductive potential) that are computable, general, justifiable and objective. They reward individuals for their potential incremental contributions to solution of the overall problem, *inter alia* by identifying diverse, high fitness simple forms likely to be good building blocks. They are invariant under invertible transformations and only slightly degraded by most others. They have avoided deception suffered by other fitness functions on a hard problem. They have the unique ability to *objectively evaluate ensembles*, without requiring foreknowledge of how the constituent sub-models might be composed into a single more complex model. This can be used to evaluate potential reproductive pairings not only severally but also jointly, enabling non-random mating to recombine synergetic material, dramatically increasing the likelihood of producing offspring of higher solution quality than the better of their parents; this is *evolvability, the sine qua non of evolutionary computation*.

9.2 Algorithms

We have developed accurate, memory- and time-efficient, parallelizable, robust algorithms for estimating entropy, mutual information and information dimension of mixed continuous (even with repeats) and discrete (even categorical) data. These use sparse arrays to represent the maximum likelihood estimate of the empirical copula density, which fully characterizes multivariate dependencies.

9.3 Corollaries to Price's Theorem

We have extended Price's Theorem to non-random mating. We have reinterpreted it to estimate effective fitness of a genotype in the context of the population that will provide its mates. We have decomposed Price's Equation into terms that show the effects of selection, genetic variation and their interaction, and identified the key factors therein affecting heritability and *evolvability*.

We have made significant progress Towards an Information Theoretic Framework for Evolutionary Learning.

Appendix A: Biographical/Philosophical Background

This research is now focused on the application of information theory to evolutionary computation, especially machine learning of ensemble models, but it began as a foray into time series prediction. It is most naturally explained in the context of that application, the difficulty of which motivated its development.

The broad scientific objective is development and dissemination of understanding of evolution and intelligence and their interplay: how evolution can give rise to intelligence and consciousness; and how conscious action and design can then play important roles in subsequent evolution. Narrowing the focus somewhat, the “missing link” between symbolic Artificial Intelligence and non-symbolic Computational Intelligence (including machine learning approaches based on evolutionary computation) appears to be the automatic generation, maintenance and exploitation of probably approximately correct, hybrid symbolic/numeric models of the world, the self and other agents, for prediction, what-if analysis and control. Progress here would have both scientific importance and significant engineering applications (including the “sensor to symbol problem” in sensor signal processing and sensor networking).

We narrowed the focus further, to development of theory and techniques for the automatic generation of predictive models of a single observed actor or process. For clarity of analysis and ease of testing, we sought the simplest world that retained important and interesting difficulties for modeling. We chose time series emerging from “black boxes”, where the hidden systems are essentially deterministically chaotic, but with stochastic elements entering into the dynamics, and noise corrupted observables.

> … we are struck by the variety of routes that lead people to study time series. This subject, which has a rather dry reputation from a distance (we certainly thought that), lies at the heart of the scientific enterprise of building models from observations. One of our goals was to help clarify how new time series techniques can be broadly applicable beyond the restricted domains within which they evolved (such as simple chaos experiments)…
>
> Gershenfeld and Weigend, 1994 [42]

We desire techniques that produce parsimonious system descriptions, which provide insight into what the hidden dynamics *might be*. As usual, we posit that the simplest model that fits the data is 'best' (Akaike's Information Criterion [AIC], Schwarz's Information Criterion [SIC], Rissanen's Minimum Description Length [MDL], Wallace's Minimum Message Length [MML], Ockham's Razor, etc.) – but what "fits the data" depends upon what we assume to be signal versus noise.

> Our viewpoint is that the study of natural systems begins and ends with the specification of observables describing such a system, and a characterization of the manner in which these observables are linked… the concept of a model of a natural system N is a generalization of the concept of a subsystem of N… The central question surrounding the issue of model credibility is to ask to what extent 'good' predictions can be made… The answer is wrapped up in the way in which the natural system N is characterized by observables, the procedure by which observables are selected to form the subsystem, and the manner in which the subsystem is encoded into a formal mathematical system F which represents, or models, the phenomena of concern.
>
> J. Casti, 1992 [25]

To generate a useful hybrid model, we must both discover its symbolic structure and optimize its numerical parameters. To test that model, we must evaluate both its short-term point predictions and its long-term attractor characterizations.

If it can provide medium-term probability density function estimates conditional upon recent history, so much the better, although their evaluation may be difficult. Conventional approaches to modeling (see Appendix B: Related Work) have failed to yield the desired combination of parsimonious symbolic representation and accurate numeric performance even on purely deterministic, much less on partially stochastic, chaotic time series. Thus they are attractive candidates for prediction by a more general modeling technique inspired by the learning mechanisms found in natural evolutionary biology and neurobiology.

Our approach integrates several methods. Genetic Programming is used to discover the structure of differential equations that estimate the observable. Evolution Strategies are used to initialize the parameters of each ‘newborn’ individual in the population. Neural Network style learning is used to continue on-line optimization of those parameters. Not central to the approach, but of interest, is the possibility of using joint time-frequency methods, such as wavelets, to facilitate the extraction of features useful in predictive modeling.

We selected an initial test problem (Sprott’s simplest chaotic flow). We attempted to discover its defining equation using: our hybrid technique, implemented in *Mathematica*; a GECCO *TinyGP* contest entry[37], also implemented in *Mathematica;* and the *Discipulus* commercial package. All yielded formulas that grossly approximated the observable, but none found the defining equation. Indeed none found equations that, if iterated, would produce chaotic behavior. While they may have been capable of short-term point prediction, they could not reproduce long-term attractor properties, nor provide insight into the dynamics.

In short, they did not yield the models we desire. We investigated why not, and identified a form of *deception* in our initial test problem: of its several terminals, only one correlated (actually, strongly anti-correlated) with the target output; however, all terminals are found in the defining equation and make essential contributions to the behavior of the system. Without a better indicator of fitness than time domain root mean squared error or the like, and without some means of preserving essential diversity (solution-relevant information) in the population, GP would converge prematurely to populations from which all but one of the necessary terminals were absent. At this point, we could have simply replaced the error based fitness measure with one that exploited non-linear correlations; 2nd order would have sufficed for this particular test problem. However, instead, we chose to regard this as an opportunity to develop theory and techniques that can deal with dependencies of arbitrary order: we turned to information theory.

Mutual information measures how much [the distribution of] one ‘random’ variable tells us about [the distribution of] another. This captures all dependencies, whether they manifest as linear correlation, or at higher orders. Gaussian RVs that are linearly uncorrelated are independent, but this does not hold for other distributions – a linear correlation coefficient of zero does *not* generally imply independence. We developed indicators of fitness and diversity based on mutual information. This led to discoveries relating redundancy, synergy and epistasis; to joint measurements of the fitness of groups, especially potential parents, and thus to mate selection heuristics; and to multiobjective optimization using information theoretic functionals with various error measures.

More significantly, our pragmatic application of information theoretic functionals to avoid the deception encountered in our initial test problem, led us to consider the general applicability of information theory to evolutionary computation. Systematically considering each phase of problem solving using evolutionary algorithms, we found that information theory can be applied profitably throughout.

There are several perspectives from which evolutionary algorithms are analyzed. One can begin empirically or theoretically. One can emphasize micro- or macro-aspects, such as genetic operators or population dynamics. Fitness landscapes and deception are popular metaphors. Information theory can provide a new perspective.

In a natural environment, there are many resources available for exploitation by organisms (energy, raw materials, waste disposal sites, etc., all contributing to the vaguely hypothesized and debated quantity of ‘negentropy’). There are also many dangers. Survival and reproductive selection pressures are complex.

On the other hand, in an artificial environment, especially one designed for a clear engineering purpose (such as learning an input-output mapping, with clearly definable measures of solution quality), things can be made simpler. Selection pressure (whether applied to survival, reproduction or both) is designed to improve solution quality as measured in at least an upper tail of the population. Selection pressure is moderated to balance exploration (coarse sampling of grossly undersampled regions of the search space) versus exploitation (fine sampling of regions deemed attractive based on information already acquired).

The environment of a GP based learner of input-output mappings provides one essential ‘energy’ resource that models in the population must exploit (if they are to survive and reproduce): the training data set. The quantity of this resource that is available is the amount of information therein: Shannon’s entropy H of the target output data set. The quantity of this ‘energy’ that is required by an ‘organism’ to stabilize its pattern (to survive and reproduce at the maximum rate so that it comes to dominate the population) is the amount of information in the individual: Shannon’s entropy H of the model output data set. The foregoing applies implicitly to all evolutionary learning; applying it explicitly, we hope both to illuminate the evolutionary process and to solve hard problems.

We also note in passing the existence of deep connections with the other domain in which a central role is played by entropy: the explanation of thermodynamics by Boltzmann and Gibbs, statistical mechanics. Solving a problem is doing work: it consumes exergy (equivalent to information?) and increases entropy; solving a problem using EAs involves a population of individuals that might be treated as molecules of a gas, where their degrees of freedom (information dimensions) should lead, via statistical mechanics, to pressure, volume, temperature, etc. of the “grand canonical ensemble” that the population comprises.

Sources of problem difficulty must be analyzed in information theoretic terms, starting with the fundamental question: What are building blocks in information theoretic terms? Returning to the test problems and the GP attack on them…

The hybrid learning system will be applied to a series of synthetic test problems of increasing difficulty: discrete and continuous, deterministic and stochastic, linear and nonlinear, periodic and chaotic, clean and noisy functions. Ultimately, successful application to an empirical data set would prove its practical value; but that is not essential to validating our extension of EA theory.

This approach is being developed specifically to predict, and understand the possible origins of, time series that have resisted prediction by other methods, but which are potentially predictable to some extent. The most obvious sources of such data sets are processes that combine chaotic and stochastic dynamics. Chaotic dynamics are deterministic, but difficult to predict due to rapid divergence of nearby trajectories (sensitive dependence on initial conditions). Stochastic dynamics are by definition non-deterministic, but still ‘predictable’ in the sense of [conditional] probability density functions or distributions.

Systems that combine both, confound techniques developed for each. Randomness that enters into system dynamics, versus mere noise corruption of the observable, defeats nonlinear dynamical systems time series embedding techniques. Chaos allows for short-term predictions unachievable by statistical techniques, and introduces apparent nonstationarity that defeats them.

Non-invertible maps with one or more degrees of freedom, invertible maps with two or more degrees of freedom, and continuous time flows with three or more degrees of freedom, can exhibit chaos. Many well-known chaotic systems, including some that model important real-world processes, are of these minimal orders. Higher order systems are also of interest, but present greater modeling and prediction difficulties.

The ideal initial test problem is the simplest system that exhibits all of the above characteristics. It is simpler to evolve a tree than a forest, so a process described by a single higher order ODE is preferred to one described by a system of multiple first order ODEs. Sprott has investigated and catalogued many simple chaotic systems, especially jerk equations[98], of which the simplest appears to be

$$x''' = -ax'' + (x')^2 - x \qquad \textbf{(A-1)}$$

As the parameter *a* is reduced from 2.082, the system follows the common “period doubling route to chaos”. From 2.0577 to 2.0168 there is the usual structure of bands of chaos interrupted by periodic cycles.

To produce a test problem with the desired characteristics, these deterministic dynamics are made stochastic by ‘randomly’ varying the parameter *a* over the entire period doubling and chaotic interval [2.0168 .. 2.082]. Experiments are being conducted on personal computers, where *use of the nominally random numbers available within the environment would amount to self-deception*.

Therefore initial experiments have been conducted using sinusoidal variation of the parameter: this still makes the process nonstationary in the short term (albeit both wide sense and cyclo-stationary), so the interesting problems can still be studied; better sources of random numbers have been identified [113][114] and will be inserted. Another way of making the dynamics stochastic would be to add a random term ***r***, not itself directly observable, into the formula for the observable:

$$x''' = -ax'' + (x')^2 - x + r \qquad \textbf{(A-2)}$$

We will also test with other chaotic jerk systems, including the WINDMI attractor:

$$x''' = -ax'' - x' - e^x + b \qquad \textbf{(A-3)}$$

[96] and the Moore-Spiegel oscillator:

$$x''' = -x'' - (T - R + Rx^2)x' - Tx \qquad \textbf{(A-4)}$$

Chaotic systems can be extremely deceptive. Fitness evaluations may require many iterations of candidate model equations, calculation of attractor dimensions, etc. as well as information measures. Multiobjective methods thus involve fitness evaluations of widely varying costs. Techniques are needed to decide, during the course of an evolutionary run, when evaluation of the more expensive objectives is justified in terms of information gain.

Pending application of decision theoretic techniques, expensive fitness functions can be treated as constraints: rather than calculating them accurately, they can be estimated quickly, and individuals with grossly bad performance penalized. Akaike's Information Criterion is often used to penalize model complexity, but would be redundant with normalizing indicators by total entropy.

In addition to the evolutionary parameters that must always be selected when applying GP, there are additional design decisions involved in modeling an unknown system using differential (or difference) equations or when embedding an attractor. Will a single higher order ODE, or several first order ODEs, be used? If the former, what order; if the latter, how many? What is the lag vector length? These all relate to various definitions of system dimensionality: the integer number of degrees of freedom in the configuration space (of the unknown equations of state describing the hidden dynamics); the (sometimes smaller) integer number of dimensions of the solution manifold; the (usually smaller, usually non-integer) dimension of the attractor; etc. [96] These are all, in general, unknown *a priori*. The lag vector length is usually determined empirically by the method of false nearest neighbors. The time delay step size between elements of that vector is usually determined empirically from the first minimum of the mutual information (a usage of that functional unrelated to that we make for EAs).

While all of these are important in practice, they are not germane to our research. Therefore we evade them, by using synthetic data for which the correct answers to these questions are known. We use chaotic jerk equations, the simplest flows that can exhibit chaos, so we need find only a single higher order ODE rather than multiple first order ODEs, and so that single ODE can be of minimum order.

We eschew Principal Components Analysis (Karhunen-Loeve transformation) of the data: while it removes linear correlations of the elements of the embedding vector, it is likely to obscure the non-linear relationships we seek to discover. Our technique is related to the generalization of PCA, not just to low degree nonlinearities, but to nonlinear Independent Components Analysis.

A question that cannot be so easily evaded is that of overfitting. We need to ‘clearn’ the data: concurrently *clean* (de-noise) the data using a tentative model, and *learn* (discover, construct) the model from that data; but when do we stop? Any filtering must be strictly short length Finite Impulse Response, as long FIR or Infinite Impulse Response filtering will alter its dynamical structure.

This is related to the need to withhold testing data from the training set. The amount of data to be withheld depends upon the relative frequency of testing versus training events, and the length of GP runs. A related issue is the distinction between testing of the parameter values to which an individual has trained, and testing of the structure of the individual: should different testing sets be used for short term point prediction tests and tests of the variants of attractors reconstructed by long term iteration of the model equations?

Wolfram Research, Inc. publishes *Mathematica*, software that they describe as a “fully integrated environment for technical computing”. It has substantial power applicable to the problem of nonlinear stochastic systems prediction and analysis, and is based upon a flexible and extensible programming language. It has been considered for GP work by many, but actually used by few, because it constantly strives to reduce expressions, interfering with their explicit maintenance and manipulation by GP systems and their programmers. [37][100] Its advantages appear to outweigh its disadvantages as a platform for our research, so the hybrid evolutionary system is being developed in *Mathematica*.

Appendix B: Related Work

B-1. Statistical Techniques

Approaches that assume partial predictability, partial randomness, stationarity, ergodicity and linearity have dominated statistical time series prediction efforts. Most techniques attempt to fit the data to models of various classes, most of which have been variations on the Auto Regressive Moving Average (ARMA) described by Box and Jenkins [14]. These work well within their domain of applicability, where their assumptions hold; but they break down quickly if applied outside that domain, even with seemingly minor violations of the assumptions.

> … *if and only if* the power spectrum is a useful characterization of the relevant features of a time series, an ARMA model will be a good choice for describing it. This appealing simplicity can fail entirely for even simple nonlinearities if they lead to complicated power spectra (as they can). Two time series can have very similar broadband spectra but can be generated from systems with very different properties, such as a linear system that is driven stochastically by external noise, and a deterministic (noise-free) nonlinear system with a small number of degrees of freedom… Nonlinearities are essential for the production of 'interesting' behavior in a deterministic system; the point here is that even simple nonlinearities suffice.
>
> Gershenfeld & Weigend, *op cit [42]*

Traditional statistical approaches, which can model linear dynamics, even when almost entirely masked by noise, often fail utterly when confronted with nonlinear dynamics, even when completely noise-free. Operating at shorter time scales, where noise corruption of the observables or stochastic dynamical components dominate, statistical techniques are still likely to be confounded by the apparent non-stationarity and consequent non-ergodicity introduced by nonlinearity. Nonlinear dynamics researchers have thus had to develop alternative methods.

B-2. Dynamical Systems Techniques

Approaches that attempt to provide short term point predictions, long term attractor reconstructions and characterizations of invariant properties, assuming nonlinear but strictly deterministic underlying dynamics, have been developed over the last 30 years. The practice of these techniques has no universally accepted name, but has been called ‘embedology’ by some of its practitioners.

> The impact of the discovery of chaos lies in the realization that nonlinear systems with few degrees of freedom, while deterministic in principle, can create output signals that look complex, and mimic stochastic signals from the point of view of conventional time series analysis… The key fact is that short-term prediction is not ruled out for chaotic systems, if there are a reasonably low number of active degrees of freedom. It is the ability to predict a short time ahead that is the basis of the new techniques…
>
> Ott, Sauer and Yorke, 1994 [75]

These techniques, while remarkably successful in providing accurate, deterministic short term point predictions in some cases that had been intractable with previous techniques, also have limitations and weaknesses. The models they produce typically: are not symbolic but rather numeric; are not global but rather consist of large numbers (thousands or more) of local submodels; and are linear, or at most low degree polynomial, despite arbitrary nonlinearities being their entire *raison d’etre*. They suffer greatly from noisy observables.

> Global models of the source of one's observations are in some sense the eventual goal of the analysis of any data set… The development of such global models is unlikely to proceed in an algorithmic sense… The very interesting theoretical work on signal separation has not been enormously successful in its application to its original goal: reduction of noise in observations of chaotic systems.
>
> H. Abarbanel, 1995 [1]

B-3. Population Learning Techniques

Statistical and dynamical systems techniques try to fit data to standard models. Some techniques use a single model, and the only flexibility is in its parameters; other techniques have a library of models, from which a selection is made, either by the user or by an algorithm. Control systems engineers call this model selection process "system identification". For greater generality, some researchers have applied Genetic Programming, effectively using an infinite library of models; but this model space must be searched efficiently.

Genetic Programming is well suited to the discovery or construction of useful and sometimes meaningful structure. One classical means of describing a system is by a single, higher order, ordinary differential (or difference) equation (ODE). Another classical dynamics description is a system of several, first order, ODEs. A GP can evolve a tree representing a single ODE, or a forest representing a system of ODEs, providing the desired insight into the hidden system dynamics.

> “Inducing equations based on theory and data is a time honoured technique in science. This is usually done manually... In this work… human induction of equations is compared with the use of genetic programming… The genetic programming induced equation was competitive with the best of the human-induced equations... Particularly important… was the possibility to (i) interpret the equation, and (ii) improve it using theoretically motivated considerations… the genetic programming engine was used as a hypothesis generator… answers are produced through automatic means, not as a black box, but as a tentative expression that can be used as a basis for analysis. This is potentially a much more useful result, as it shows that the symbolic nature of genetic programming can be used to build up knowledge in a problem domain.”
>
> Keijzer, Babovic, Baptist and Uthurburu, 2005 [54]

Mutual information between the target and models (individual and ensemble) in the population can be used in all phases of solving a problem by GP. It is well suited to the identification of explicit building blocks and combinations of building blocks with high potential to contribute to the solution. Evolutionary algorithms that explicitly identify and recombine high fitness building blocks are claimed superior to those which process building blocks only implicitly [33]. Explicit learning of appropriate linkages between variables is also claimed beneficial [94].

It has recently been recognized that only a small fraction (generally the most fit) of the initial population contributes substantially to the final population [31]. It has also recently been recognized that the most fit members of the initial population tend to be small and simple [56].

Composing these findings, it is tempting to create an initial population of only the terminals themselves, or at most simple combinations thereof, which are likely to *be* the needed building blocks, and evolve successively more complex combinations only from the most fit of the simple initial forms. Such approaches have been proposed by others [25].

In place of random mating, various mate selection heuristics have been suggested [41]. Calculation of the joint mutual information with the target of a pair of potential parents is attractive as such a heuristic: offspring produced by recombination without mutation can do no better than the joint observation provided by their parents, except for reduction of excess model entropy; mutation induces variance about this, which would otherwise be an upper bound.

Survival selection has also been studied [8]; joint mutual information provides an upper bound on the information that can be provided by an ensemble, without foreknowledge of how information from the different individuals in the ensemble might be combined. In this context, multiobjective GP is appealing [84].

While GP is strong in discovering symbolic structures, it is weak in optimizing any numerical parameters that may need to be embedded in such structures for them to fit the data accurately. This can be mitigated with another population learning technique, Evolution Strategies. Another attractive aspect of ES is self-adaptation of evolutionary process parameters.

As foreshadowed (in Appendix A: Background), certain problems, including our initial test problem, exhibit deceptiveness of one form or another; this affects various evolutionary algorithms in different ways. We argue that deception of an EA is always *with respect to a given fitness function*, and may cause more or less difficulty for an EA *depending upon its effectiveness in preserving diversity*.

Deception may be avoided by more ‘insightful’ fitness functions and/or mitigated by more effective mechanisms for preserving diversity.

The pitfall of current techniques that our innovations avoid is loss of essential information from the population. Such loss is often due to deception of fitness functions that implicitly depend on linear correlation between targets and individuals in the population, or due to failure to preserve *relevant* diversity in the population. Such loss is especially likely when that essential information is distributed across multiple individuals but not apparent in them individually, due to epistasis or synergy among them.

B-4. Individual Learning Techniques

Among biologically inspired machine learning techniques that are not population based, artificial Neural Networks dominate: in generality of problem-solving power; and in popularity with researchers, who have applied them often to time series prediction, with varying degrees of success. Their great strength, relative to GP, is in numerical performance; but their great weakness is that, aside from a very few special purpose constructive methods, they are model-free estimators rather than explicit model builders. Natural biology inspires a hybrid approach: exploit the synergy of phylogenetic (population) learning of structures with GP, and ontogenetic (individual) learning of parameters with neural networks.

Individuals in a GP population are function trees, which in the NN literature would be called sparse high order networks; while standard NN learning rules cannot generally be applied, gradient descent or hill climbing of some sort can optimize the parameters of those networks. The adaptive individuals are likely to become trapped in local minima of the error function, but the evolving population is hoped to sample the fitness landscape sufficiently densely to locate a basin in which may be found a global optimum, providing the desired point prediction accuracy. Structural sampling of the model space can be performed using GP, and numerical sampling of the parameter space can be performed using ES to initialize and then NN style learning to further optimize the parameters.

Many researchers have employed such hybrids; the most interesting efforts have explicitly evolved ensemble models [66]. A GP population is inherently an ensemble model, whether that is explicitly recognized not. Recognition of the fact enables its exploitation: for group selection, to improve the fitness of an entire population as such, above and beyond improvements that result from individual selection of its members; and for use of a population, at any generation, to model all the various aspects of the target, even if the submodels of each of those aspects have not [yet] been recombined into a single NN.

In more general machine learning theory, the Probably Approximately Correct framework [49] is widely used. Its classical application is to state machine based learners of binary sequences. It is appealing to apply it to GP, NNs, ensemble models, hybrid learners, etc., but it is unclear how, or how it might need to be extended, to do so.

B-5. Joint Time-Frequency Representations

Pure time (frequency) domain approaches may miss features readily evident in the frequency (time) domain. Thus, we planned to use wavelets to facilitate detection of features irrespective of their ‘native’ domain. We identified several ways to integrate wavelets with NN and GP techniques. Our transforms were motivated by several conflicting objectives:

- to mitigate the effects of measurement noise in the data;
- to facilitate the discovery of hidden structure in the series;
- to represent discovered relationships in a form amenable to interpretation.

The literature is replete with applications of joint time-frequency representations to time series prediction. However, we have tabled work in this area, as our research focus has shifted away from the time series prediction application (which, despite its practical importance, was for us always only a test problem), toward the application of information theory generally to evolutionary learning.

B-6. Information Theoretic Indicators of Fitness

Deception is an important aspect of evolutionary computation. Our initial attempts at time series prediction with GP suffered from a particular form of deception: all terminals in the set selected for the test problem appear in the hidden defining equation, which we wish to discover; but only one of them correlates (actually, anti-correlates strongly) with the observable. This led to rapid loss from the population of building blocks essential to evolution of the correct solution. It was apparent that the deception was due to the implicit linear correlation inherent in typical error measures; so mutual information was investigated, as it can capture arbitrary higher-order statistical relationships.

Information theoretic functionals have been used as fitness functions, or components of fitness functions, by a few GP researchers. In [4], a variant of normalized mutual information (different from our formulation only by a constant) was a factor in a formula that collapsed multiple objectives into a scalar fitness; superiority of normalized over raw mutual information, and of the GP using mutual information over prior GA results, were reported. Despite the favorable reported results, the evolutionary computation community does not appear to be adopting mutual information broadly: we assume this is because no one has yet shown information theoretic approaches to be generally applicable to EAs; showing this has become a major goal of our work.

B-7. Information Theoretic Indicators of Diversity

Mutual information has also been used as a diversity indicator by a few evolutionary learning researchers. In [66], fitness sharing was based on the concept of set covering and implemented by minimization of mutual information between NN models comprising ensembles: favorable results were reported; but note that what was being maximized here, we label as only *apparent* diversity, which may not all contribute to modeling the target.

We find no evidence in the literature of recognition or exploitation of certain key insights regarding mutual information based indicators of fitness and diversity:

- they are justifiable, general, computable and *commensurate*

- they enable measurement of *relevant* diversity

- they enable joint estimation of the reproductive potential of groups

Practical and theoretical exploitation of these insights is now our primary focus.

B-8. Information Theory in Artificial & Natural NNs

Information theory has been used more in machine and biological learning research outside the field of evolutionary computation. The principle of maximum information preservation has been recognized in the design of self-organizing artificial NNs [65]. Neurobiologists have found that redundancy and synergy are both exploited in neural coding [26].

B-9. Information Theory in Natural Evolutionary Biology

Numerous references to information theory can be found in the highly suspect literature of "intelligent design", but few in that of evolution science. In both, the usage is more philosophical and allegorical than rigorous or literal.

> "Environmental selection emerges as a result of a lack of sensitivity (low mutual information) between the genetic system and the environment. Sexual selection, by contrast, emerges as a result of increased sensitivity (high mutual information) between members of the genetic system."
>
> D. Brooks, 2001 [16]

Whether this claim is intended to be mathematical or notional, its general accuracy, and its specific applicability to our work, are all as yet unclear. Likewise relevance to us of complexity theory and statistical mechanics, the other areas where entropy is considered, is unclear and not currently being pursued.

Lastly we must admit that much of the related work adopts simpler and less computationally costly approaches to modeling, which are often adequate. Our information theoretic evolutionary learning approach is justified only by the labor costs of manually exploring complex data sets and defining explicit algorithms, and the lack of less expensive automated procedures with sufficient generality.

Charts and Diagrams

Table 1.1: Independence (384 sequences)

	y=-1	y=+1	∑
x=-1	1	1	2
x=+1	1	1	2
∑	2	2	4

Tables 1.2 & 1.3: Perfect [anti]dependence (96 sequences each)

	y=-1	y=+1	∑
x=-1	2	0	2
x=+1	0	2	2
∑	2	2	4

	y=-1	y=+1	∑
x=-1	0	2	2
x=+1	2	0	2
∑	2	2	4

Tables 2.1 & 2.2: Perfect [non] linear dependence

	y=-1	y=0	y=1	∑
x=-1	3	0	0	3
x=0	0	3	0	3
x=+1	0	0	3	3
∑	3	3	3	9

	y=-1	y=0	y=1	∑
x=-1	0	0	3	3
x=0	0	3	0	3
x=+1	3	0	0	3
∑	3	3	3	9

Tables 2.3 & 2.4: Perfect non-linear (only) dependence

	y=-1	y=0	y=1	∑
x=-1	3	0	0	3
x=0	0	0	3	3
x=+1	0	3	0	3
∑	3	3	3	9

	y=-1	y=0	y=1	∑
x=-1	0	3	0	3
x=0	0	0	3	3
x=+1	3	0	0	3
∑	3	3	3	9

Tables 2.5 & 2.6: Perfect non-linear (only) dependence

	y=-1	y=0	y=1	∑
x=-1	0	3	0	3
x=0	3	0	0	3
x=+1	0	0	3	3
∑	3	3	3	9

	y=-1	y=0	y=1	∑
x=-1	0	0	3	3
x=0	3	0	0	3
x=+1	0	3	0	3
∑	3	3	3	9

Table 3: Statistics on Example 16 deceptive data

	similarity			sufficiency			efficiency		
	y0	**y1**	**(y0,y1)**	**y0**	**y1**	**(y0,y1)**	**y0**	**y1**	**(y0,y1)**
x0				0.139	0.152				
x1				0.145	0.158				
(x0,x1)	0.312	0.321	0.578	0.712	0.729	0.732	0.357	0.364	0.732
z0	0.702			0.825			0.825		
z1		0.714			0.833			0.833	
(z0,z1)			0.722			0.839			0.839

Table 4: Inequivalent information despite equal error

x	y	z1	z2	e1	e2
0	0	1	1	1	1
1	1	2	2	1	1
2	2	3	3	1	1
3	3	0	0	-3	-3
4	0	1	3	1	3
5	1	2	0	1	-1
6	2	3	1	1	-1
7	3	0	2	-3	-1
			mean	0	0
			RMS	1.7320508	1.7320508
			dissimilarity	0	0.6666667

Table 5: Fitness indicators applied to 1-term expressions

	NRMSE	rank	abs(r)	rank	NMI	rank
x"	0.00378	1	0.00001	3	0.32918	1
x'	0.00248	2	0.62345	1	0.28534	3
x	0.00167	3	0.00002	2	0.29051	2

Table 6: Fitness indicators applied to 2-term expressions

	NRMSE	rank	abs(r)	rank	NMI	rank
x'*x"	0.00474	1	0.52760	5	0.29225	9
x"-x'	0.00377	3	0.50975	7	0.30890	5
x'-x"	0.00228	8	0.50975	7	0.30890	5
x'	0.00248	5	0.62345	2	0.28534	15
x-x'	0.00167	12	0.30025	10	0.31249	3
x'-x	0.00144	14	0.30025	10	0.31249	3
x*x'	0.00091	19	0.82145	1	0.30413	7
x"	0.00378	2	0.00001	25	0.32918	1
2*x"	0.00249	4	0.00001	25	0.32918	1
2*x'	0.00157	13	0.62345	2	0.28534	15
x'+x"	0.00228	7	0.50975	6	0.28091	19
x'^2	0.00226	9	0.57310	4	0.25740	21
x"^2	0.00240	6	0.00001	27	0.29437	8
x	0.00167	11	0.00002	20	0.29051	12
x+x'	0.00144	15	0.30027	9	0.28020	20
x+x"	0.00194	10	0.00002	19	0.28439	18
x^2	0.00032	21	0.05984	12	0.28811	14
x"-x	0.00140	16	0.00002	22	0.29080	10
x*x"	0.00102	18	0.05452	13	0.28530	17
x-x"	0.00140	17	0.00002	22	0.29080	10
2*x	0.00087	20	0.00002	20	0.29051	12

Table 7: Fitness indicators applied to 3-term expressions

	NRMSE	rank	abs(r)	rank	NMI	rank
x'^2-x	0.0028	11	0.4885	49	0.3287	9
2*x"-x'	0.0025	15	0.3610	60	0.3358	3
x'*(x"-1)	0.0047	4	0.8159	12	0.2947	63
x"-x'	0.0038	7	0.5097	47	0.3089	25
x'-x"	0.0023	19	0.5097	48	0.3089	26
x-x'^2	0.0019	38	0.4885	50	0.3287	10
x"-(x*x')	0.0013	73	0.8214	7	0.3105	20
(x'-1)*x"	0.0030	9	0.3474	63	0.3078	29
(x'-x")*x"	0.0029	10	0.3833	55	0.3037	36
x'-2*x"	0.0019	40	0.3610	61	0.3358	4
x'*x"	0.0047	3	0.5276	40	0.2923	71
-x'	0.0050	2	0.6234	28	0.2853	101
x'*(x'-x)	0.0010	93	0.8692	1	0.3031	40
x'*(x"-x)	0.0011	82	0.8542	3	0.2991	50
(1-x)*x'	0.0017	47	0.8335	5	0.2894	86
(1+x')*x"	0.0053	1	0.5750	35	0.2800	120
(x-x')*x'	0.0008	118	0.8692	2	0.3031	41
x'*(x-x")	0.0008	111	0.8542	4	0.2991	51
x*x"-x'	0.0010	90	0.2213	81	0.3490	1
x'-x*x"	0.0010	95	0.2213	82	0.3490	2
x'^2-x"	**0.0021**	**28**	**0.5110**	**44**	**0.2735**	**129**
x"	0.0038	5	0.0000	287	0.3292	5
x+2*x"	**0.0021**	**30**	**0.0000**	**258**	**0.2753**	**124**

Table 8: Diversity indicators on 1-term expressions

	NRMSE	rank	abs(r)	rank	NMI	rank
x"	0.00237	2	0.39038	2	0.28729	1
x'	0.00245	3	0.00005	1	0.31786	2
x	0.00154	1	0.39040	3	0.32584	3

Table 9: Diversity indicators on 2-term expressions

	NRMSE	rank	abs(r)	rank	NMI	rank
x*x"	0.00073	8	0.43379	17	0.26316	9
x'*x"	0.00140	20	0.17278	7	0.25024	8
x*x'	0.00078	9	0.24939	11	0.28457	16
x'^2	0.00128	13	0.24758	8	0.28368	15
x"^2	0.00134	18	0.41493	16	0.22746	7
2*x'	0.00129	16	0.24812	9	0.31197	22
x+x'	0.00132	17	0.45985	20	0.26904	11
x'+x"	0.00158	24	0.39962	14	0.26600	10
x^2	0.00027	7	0.47402	23	0.29705	19
x'-x	0.00104	12	0.47178	21	0.29638	17
x"-x'	0.00129	15	0.37171	12	0.31437	24
2*x"	0.00142	21	0.45822	18	0.27386	12
x-x'	0.00128	14	0.47178	21	0.29638	17
x+x"	0.00157	23	0.40659	15	0.27894	14
x"	0.00163	26	0.45822	18	0.27386	12
x"-x	0.00096	11	0.49553	26	0.30235	20
x'-x"	0.00154	22	0.37171	12	0.31437	24
x'	0.00171	27	0.24812	9	0.31197	22
2*x	0.00079	10	0.47654	24	0.32125	26
x-x"	0.00135	19	0.49553	26	0.30235	20
x	0.00162	25	0.47654	24	0.32125	26

Table 10: Parity arranged by ones count

x1	*x2*	*x3*	*y*
0	0	0	0
0	0	1	1
0	1	0	1
1	0	0	1
0	1	1	0
1	0	1	0
1	1	0	0
1	1	1	1

Table 11: Correlation vs sufficiency, 2^{16} points

m	f	r	s1
1	cyclic	0.56	**0.97**
1	mixed	0.90	**0.94**
1	ratio	0.64	**1.00**

Table 12: Correlation vs sufficiency, 2^{20} points

m	f	r	s1
1	cyclic	0.45	**1.00**
1	mixed	0.93	**1.00**
1	ratio	0.85	**1.00**

Table 13: Sufficiency of up to 2-ensembles, 2^{20} points

m	f	s1	s2
1	cyclic	**1.00**	-
1	mixed	**1.00**	-
1	ratio	**1.00**	-
2	cyclic	0.69	0.88
2	mixed	0.69	0.86
2	ratio	0.69	0.90

Table 14: Sufficiency of up to 6-ensembles, 2^{16} points

m	f	s1	s2	s3	s4	s5
1	cyclic	**0.97**	-	-	-	-
1	mixed	**0.94**	-	-	-	-
1	ratio	**1.00**	-	-	-	-
2	cyclic	0.66	0.83	-	-	-
2	mixed	0.67	0.86	-	-	-
2	ratio	0.69	0.90	-	-	-
3	cyclic	0.62	0.79	0.92	-	-
3	mixed	0.61	0.76	0.87	-	-
3	ratio	0.63	0.81	0.93	-	-
4	cyclic	0.59	0.73	0.86	0.93	-
4	mixed	0.58	0.70	0.82	0.92	-
4	ratio	0.59	0.72	0.84	0.94	-
5	cyclic	0.57	0.66	0.77	0.88	**1.00**
5	mixed	0.57	0.66	0.77	0.87	**0.99**
5	ratio	0.57	0.68	0.79	0.90	**1.00**
6	cyclic	0.55	0.63	0.72	0.82	0.99
6	mixed	0.55	0.63	0.72	0.82	0.97
6	ratio	0.56	0.65	0.75	0.85	0.99

Table 15: Roles of unlabeled inputs

datum	y0	y1	y2
d0a	-	-	-
d0b	-	-	-
d1	-	-	x0
d2	-	x0	-
d3	-	x1	x1
d4	x0	-	-
d5	x1	-	x2
d6	x2	x2	-
d7a	x3	x3	x3
d7b	x4	x4	x4

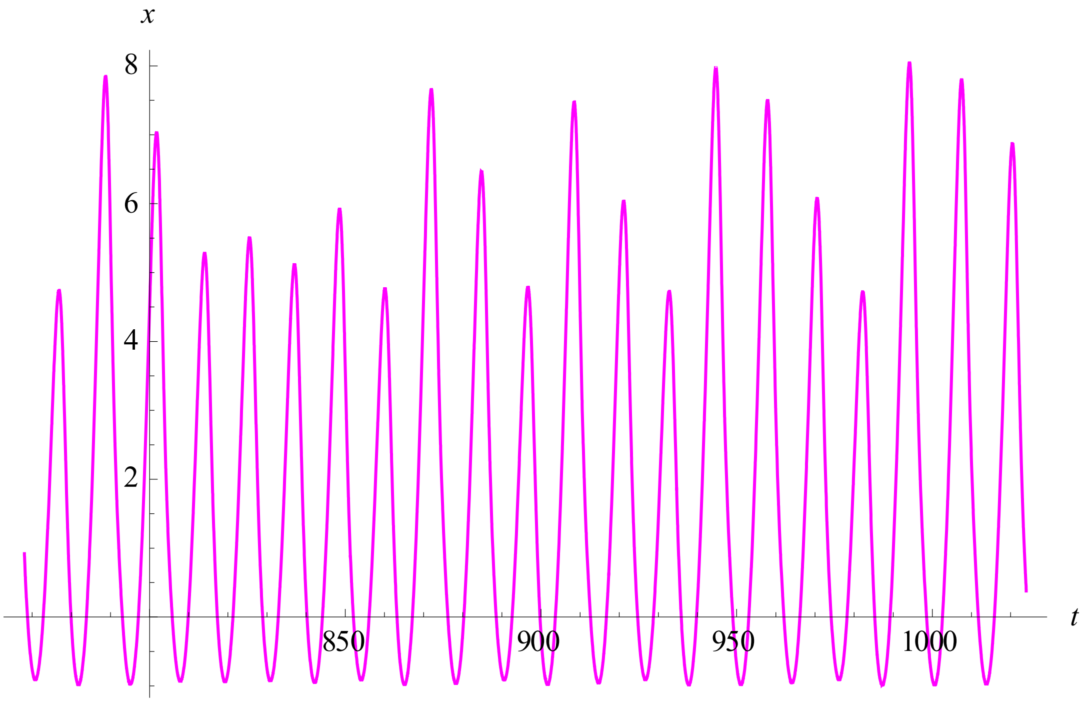

Figure 1: Sprott's chaotic jerk system: time domain

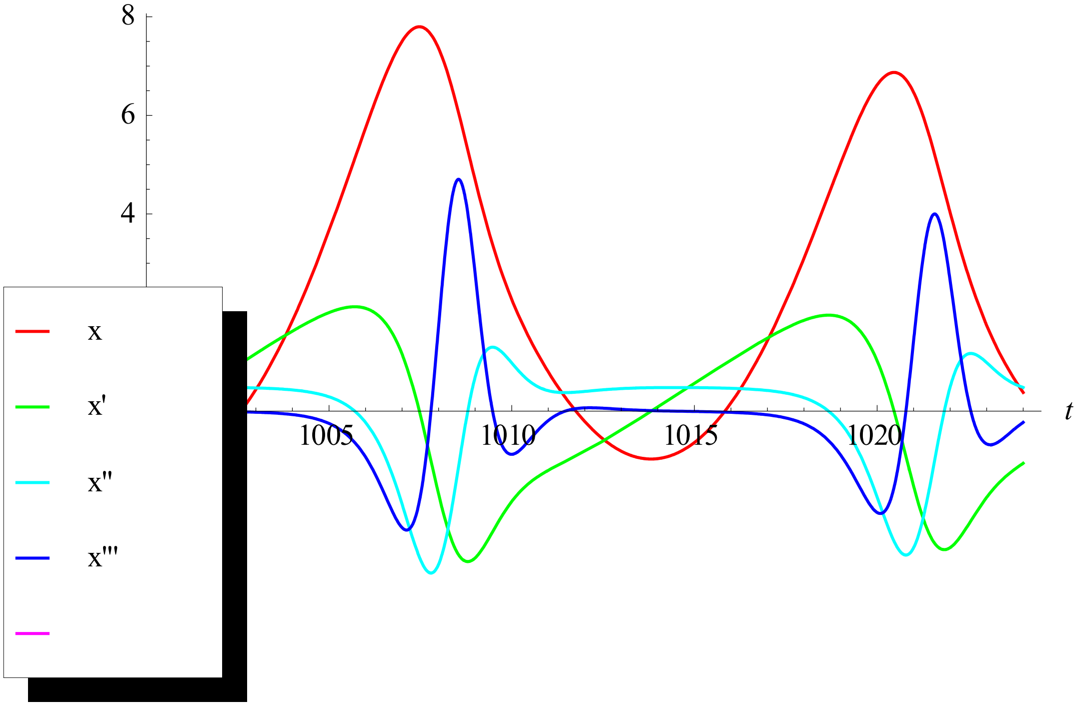

Figure 2: Sprott's chaotic jerk system: time domain derivatives

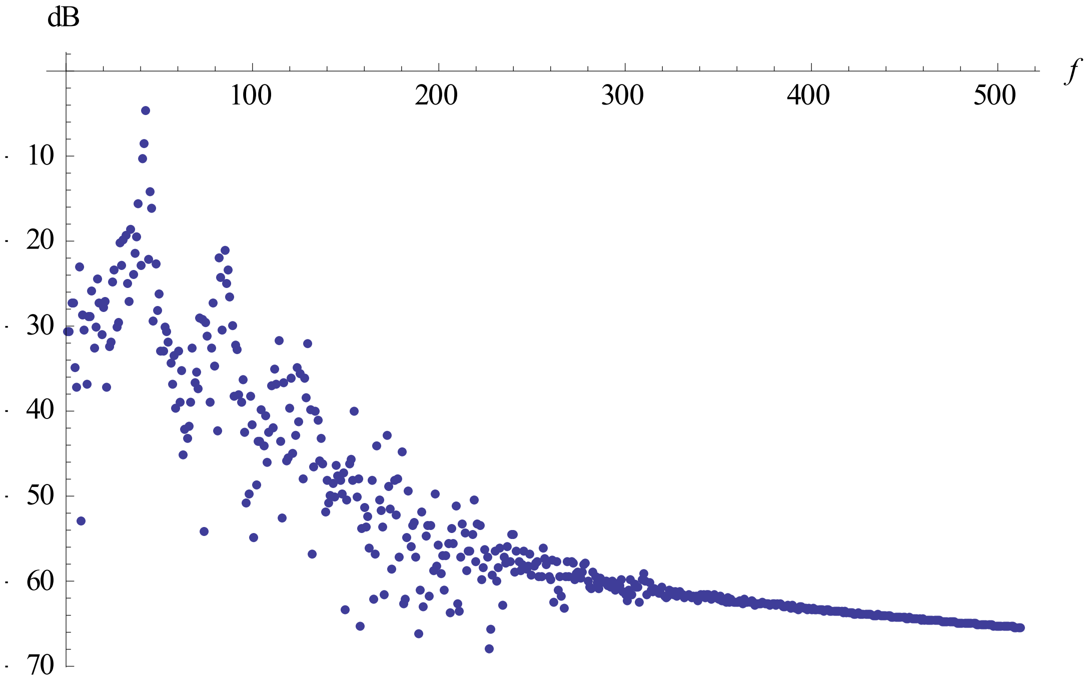

Figure 3: Sprott’s chaotic jerk system: frequency domain

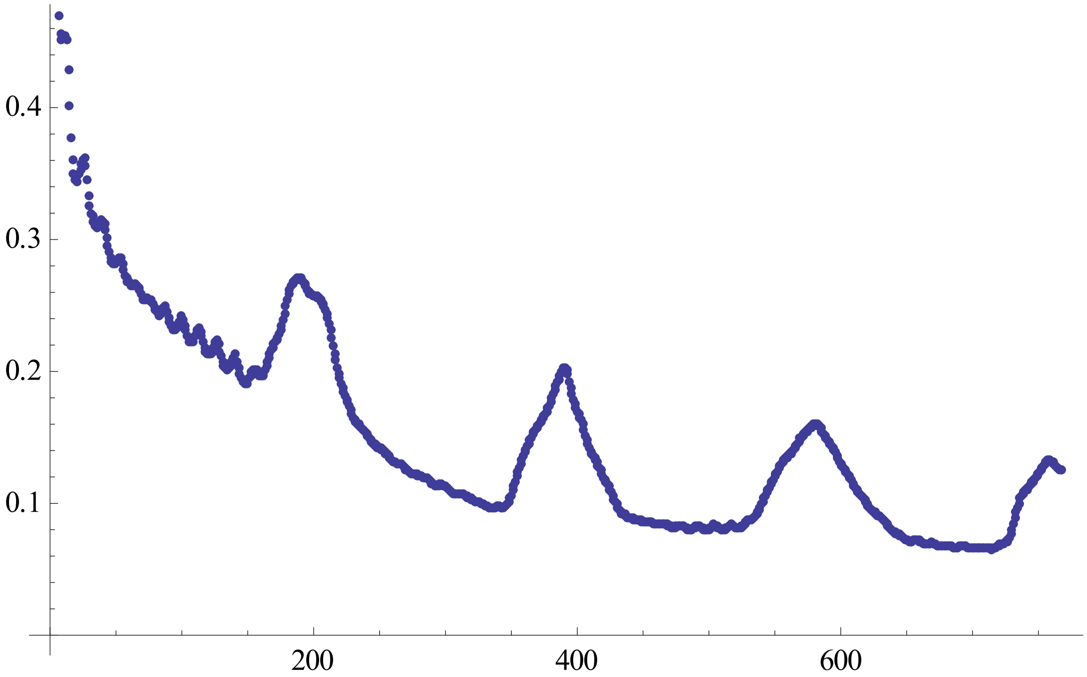

Figure 4: Sprott’s chaotic jerk system: lag-self-MI

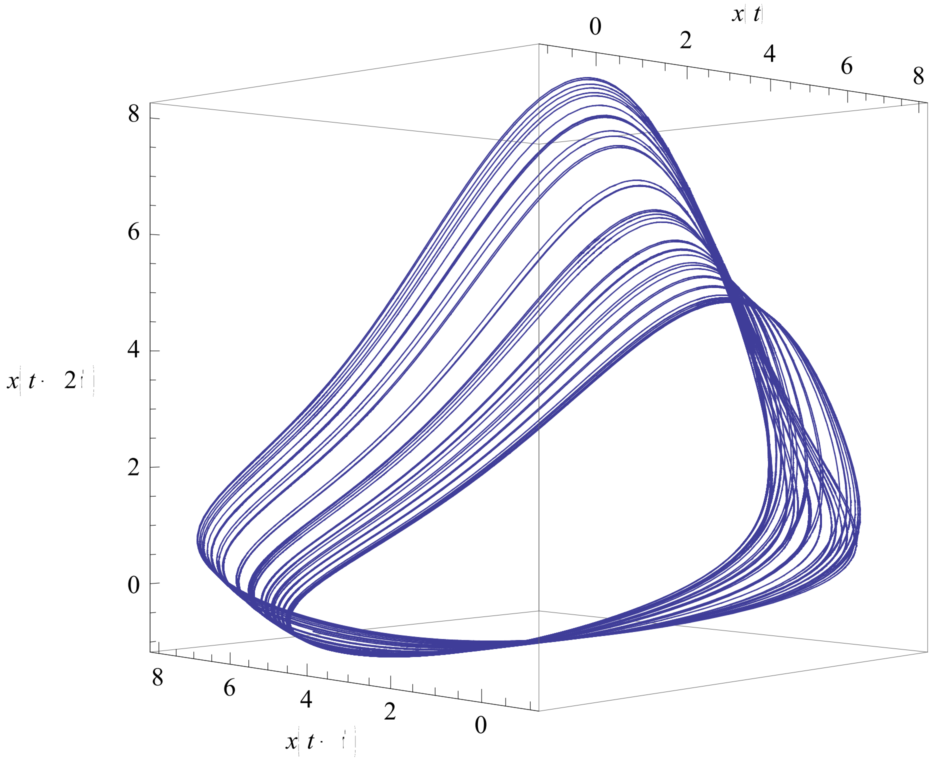

Figure 5: Sprott’s chaotic jerk system: lag space

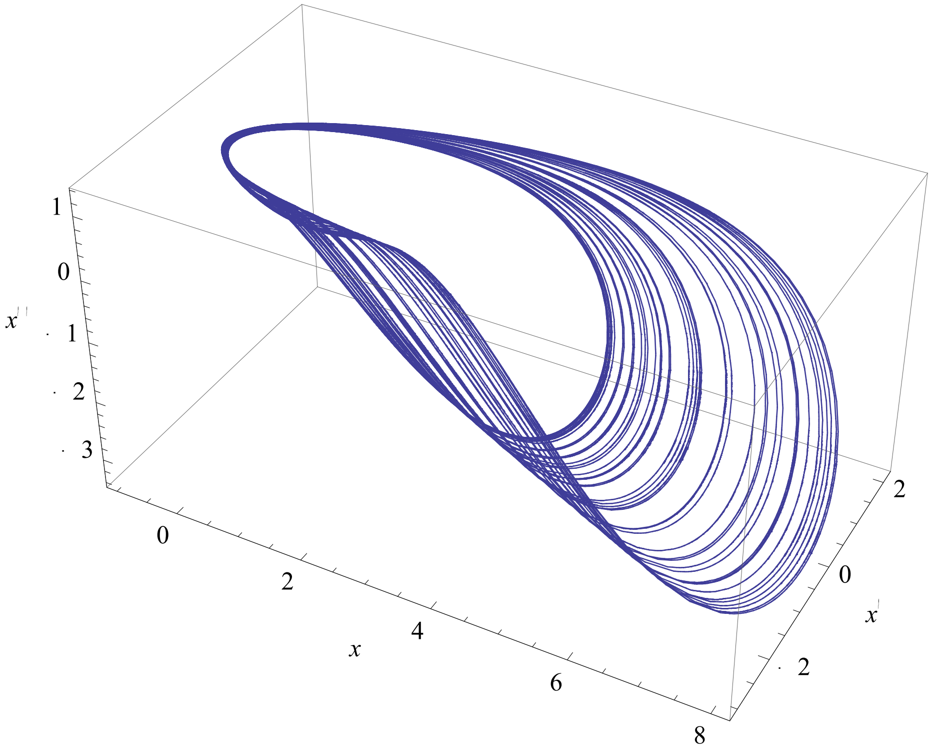

Figure 6: Sprott’s chaotic jerk system: state space

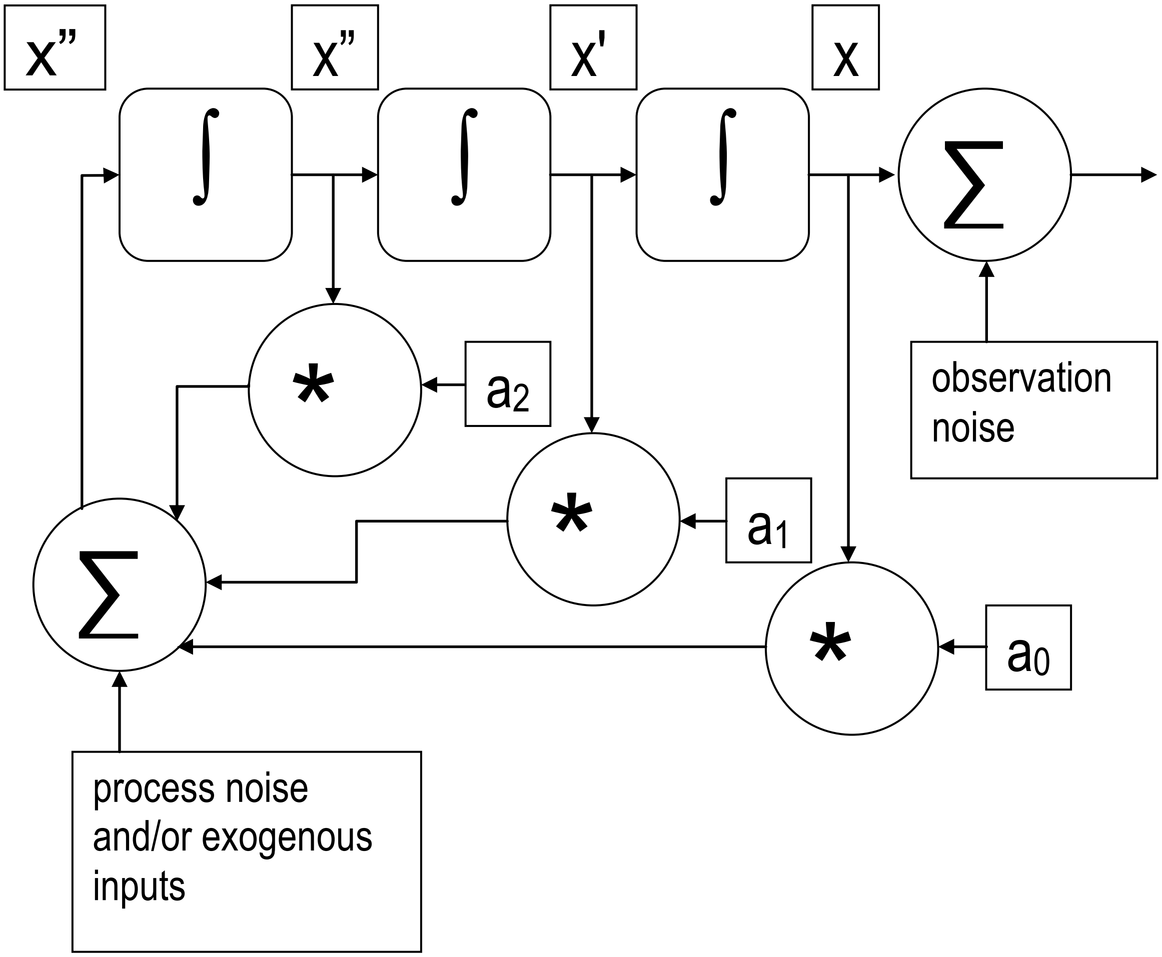

Figure 7: Jerk system

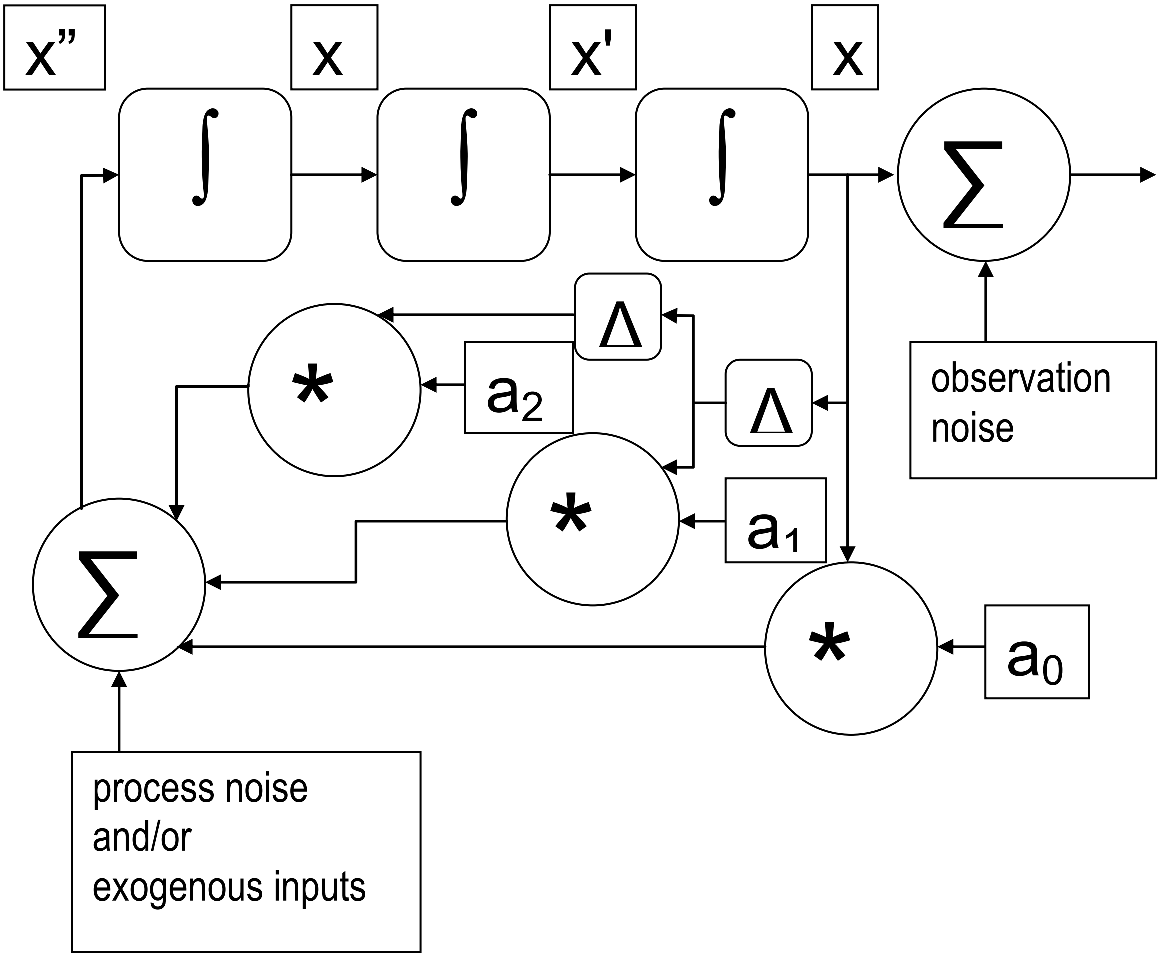

Figure 8: Inverted jerk system

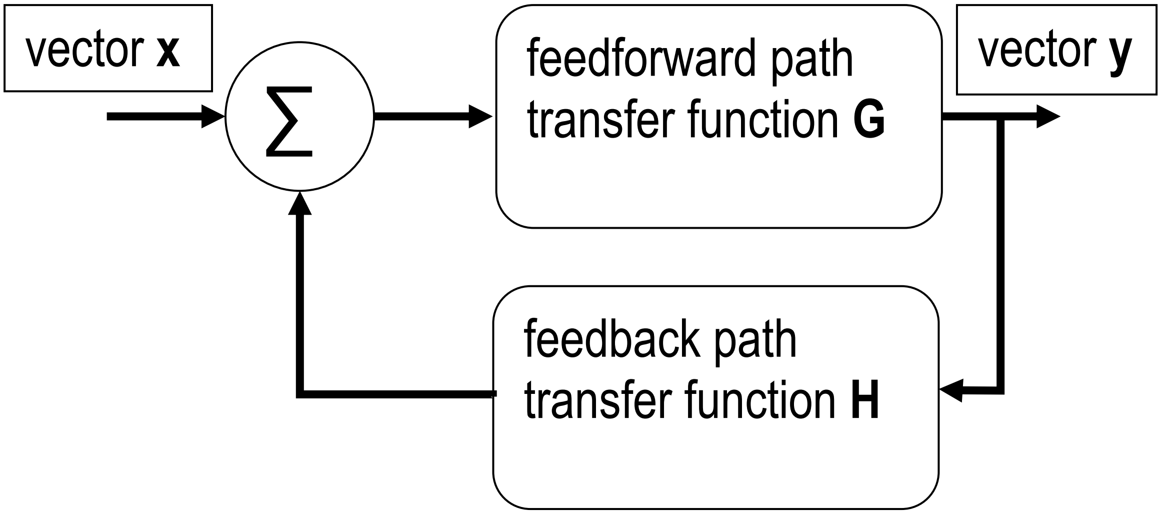

Figure 9: General feedback system

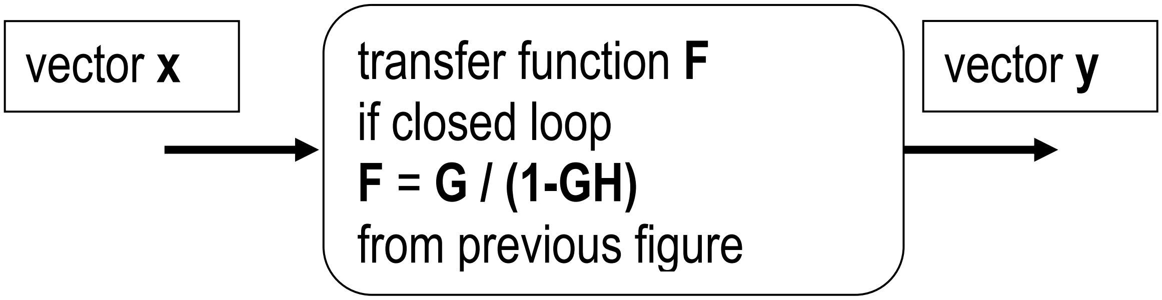

Figure 10: Open loop or unobserved feedback process

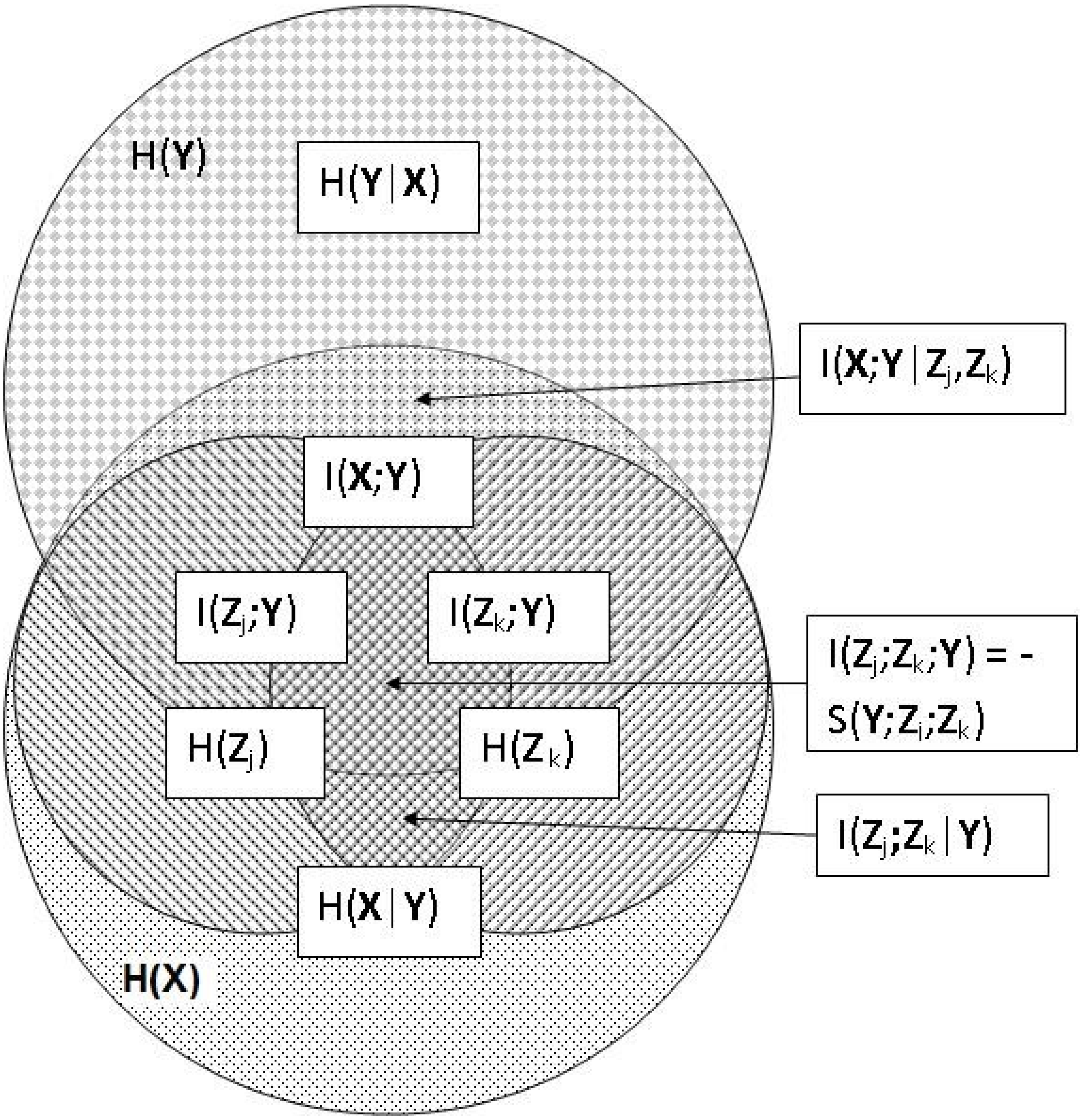
$H(\mathbf{Y})$
$H(\mathbf{Y}|\mathbf{X})$
$I(\mathbf{X};\mathbf{Y}|Z_j,Z_k)$
$I(\mathbf{X};\mathbf{Y})$
$I(Z_j;\mathbf{Y})$
$I(Z_k;\mathbf{Y})$
$I(Z_j;Z_k;\mathbf{Y})$ = - $S(\mathbf{Y};Z_i;Z_k)$
$H(Z_j)$
$H(Z_k)$
$I(Z_j;Z_k|\mathbf{Y})$
$H(\mathbf{X}|\mathbf{Y})$
$H(\mathbf{X})$

Figure 11: Input, target output & 2 non-ideal models

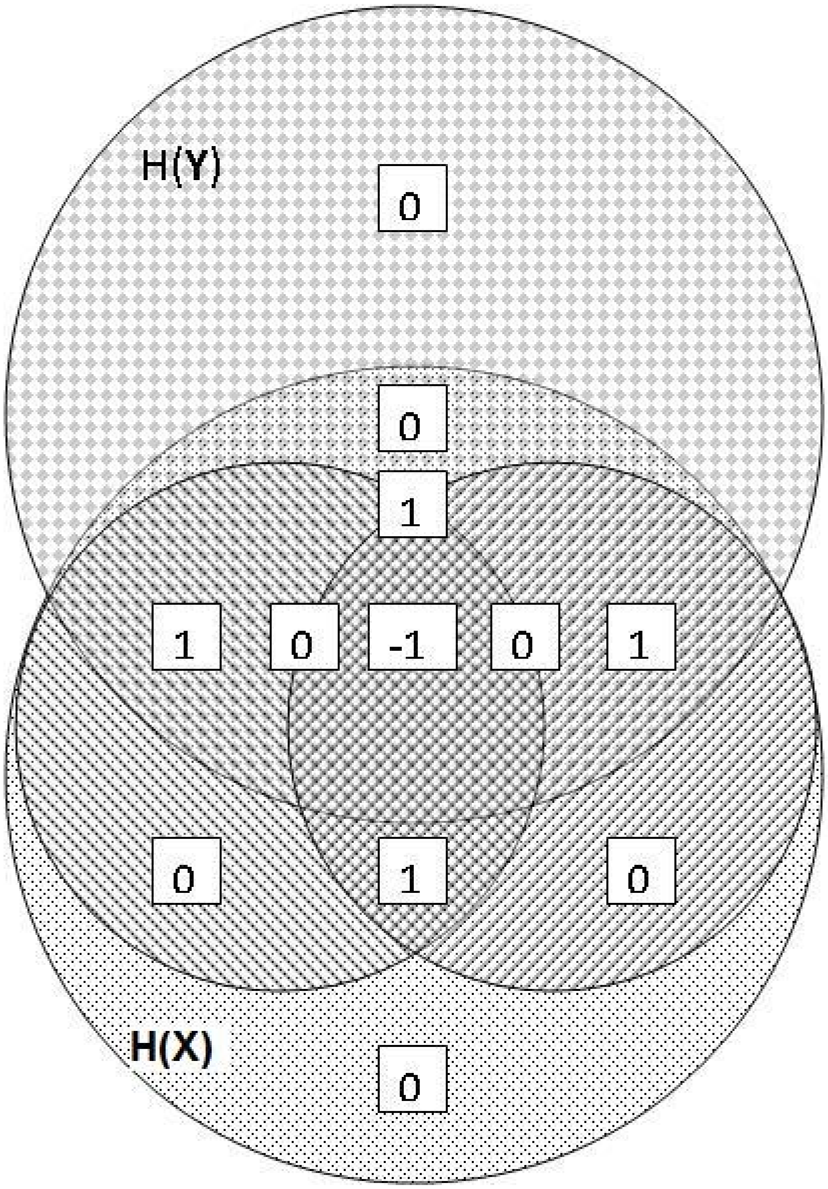

Figure 12: Example 1: maximal synergy

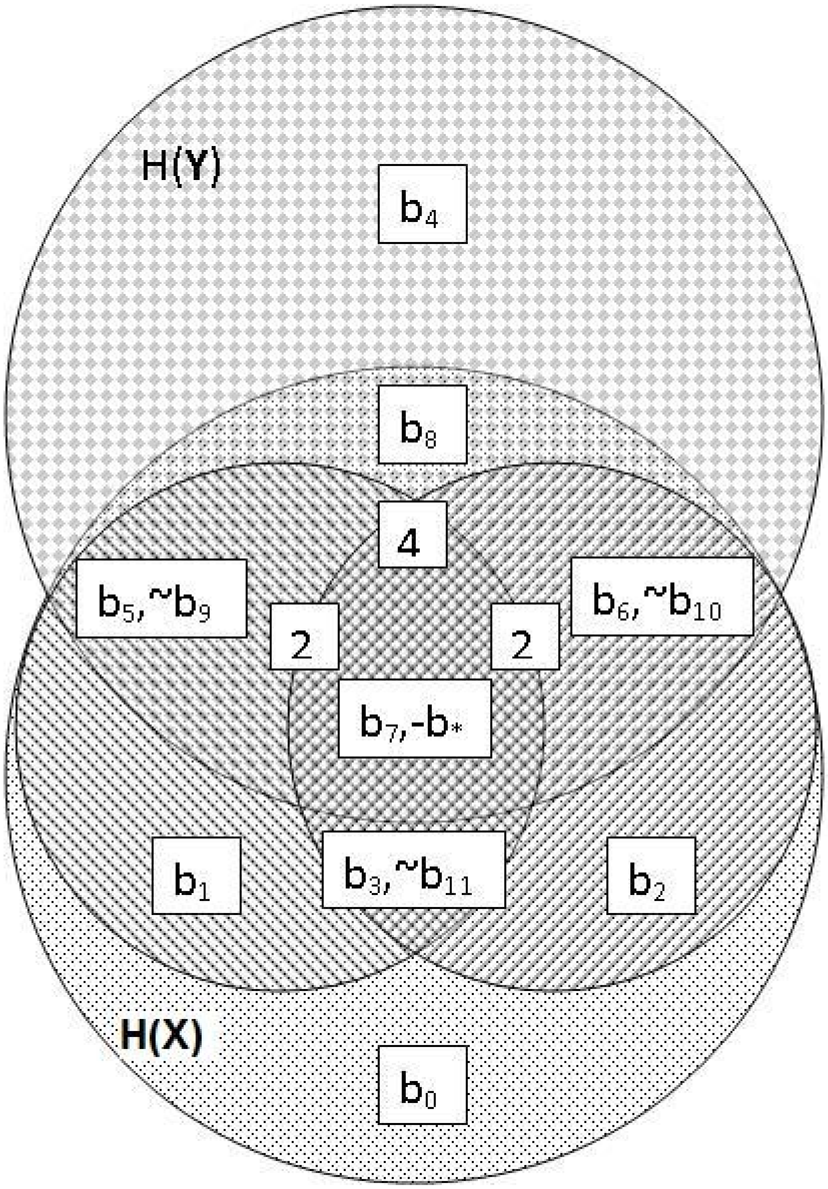

Figure 13: Example 2: redundancy = - synergy (pseudo-independence)

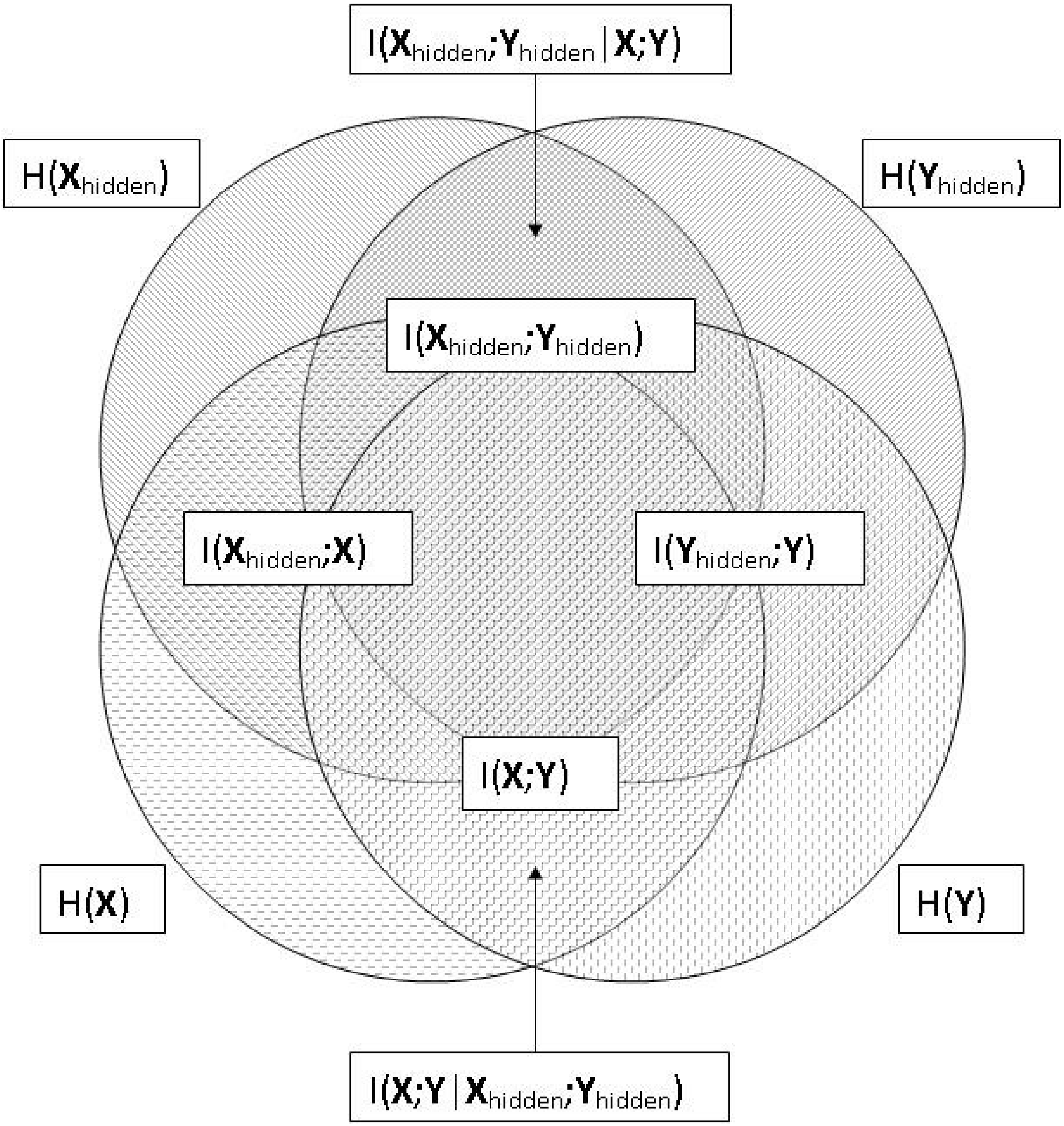

Figure 14: Hidden vs observed process

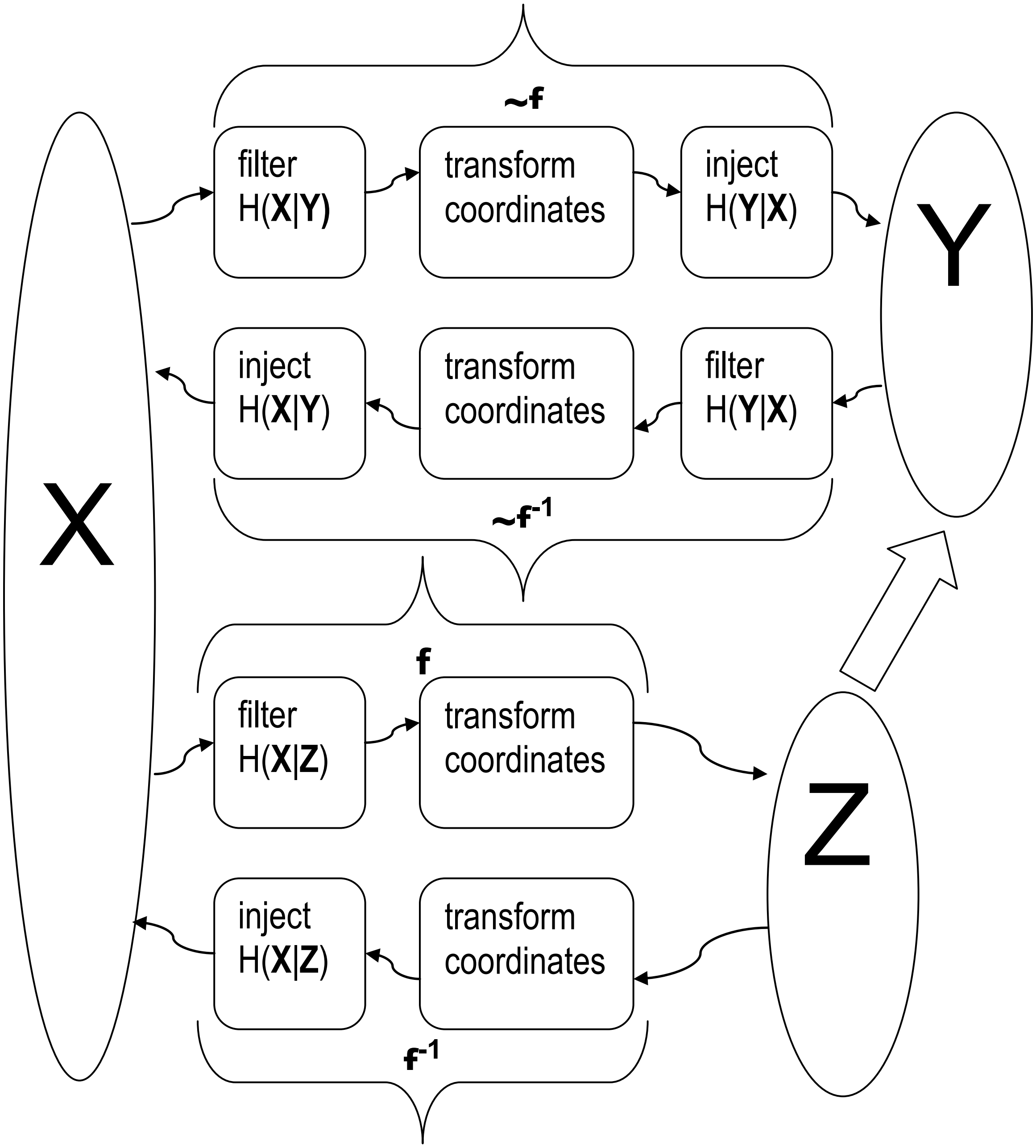

Figure 15: Ideal modeling information flow

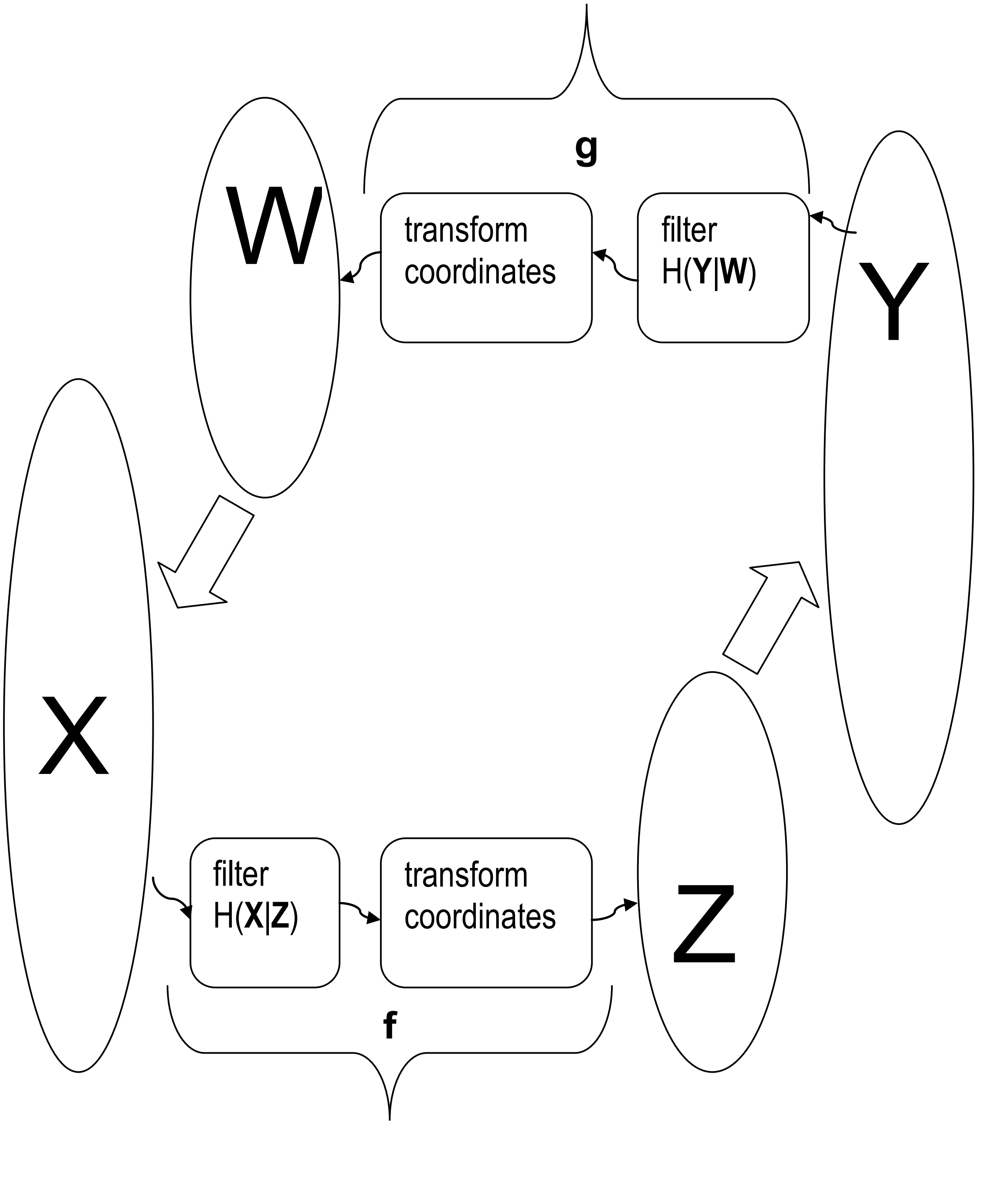

Figure 16: Forward and reverse modeling

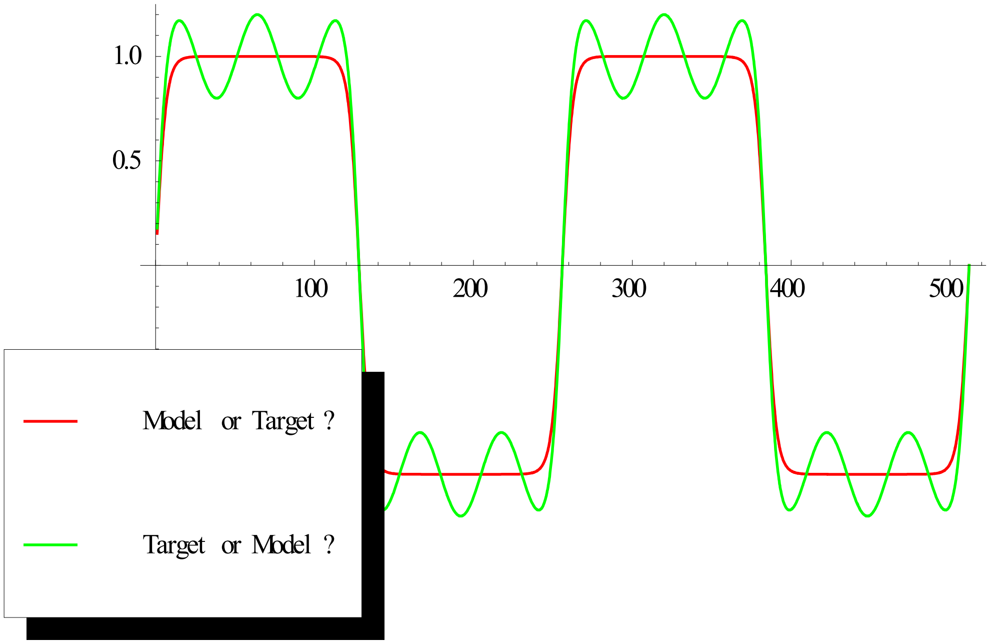

Figure 17: Error due to residual target or excess model entropy

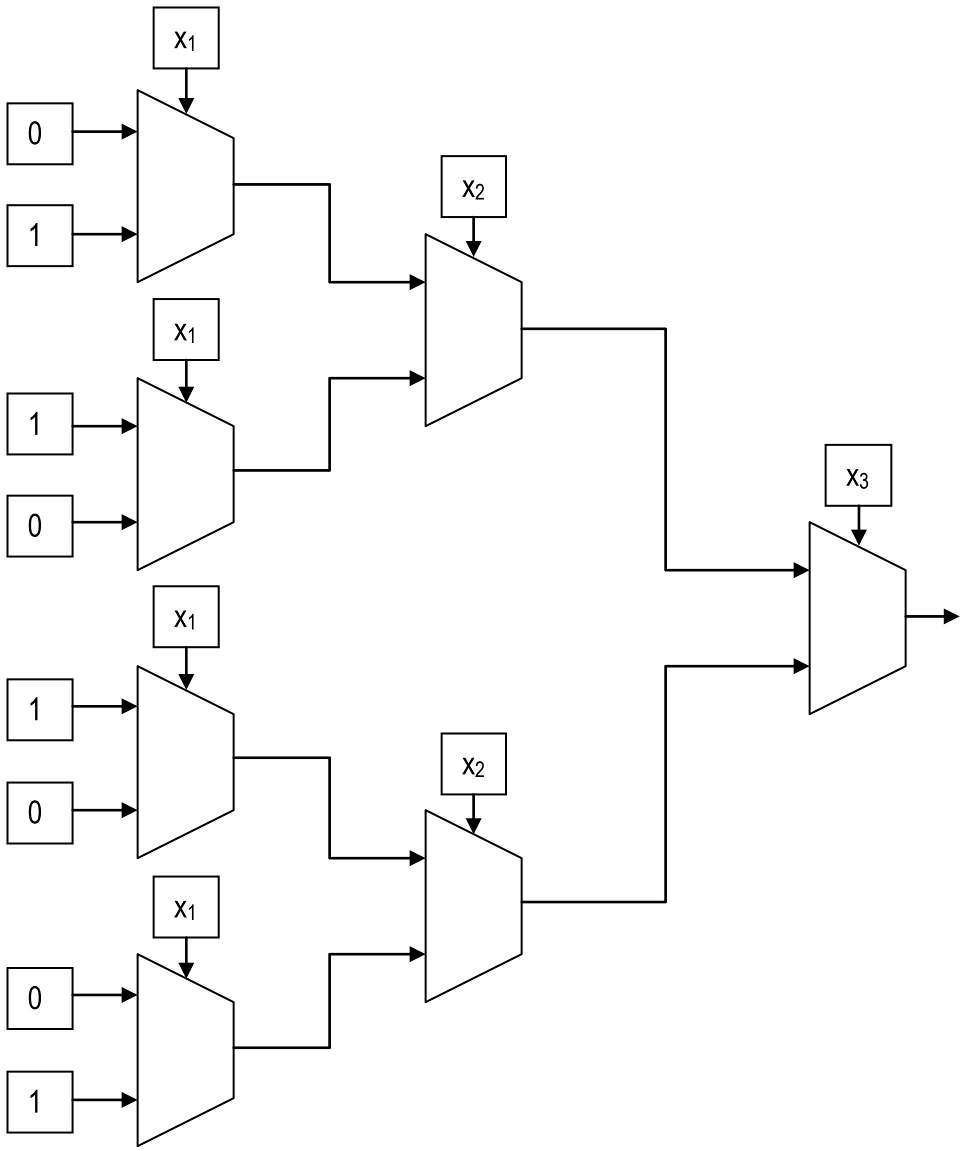

Figure 18: Parity constructed from multiplexors

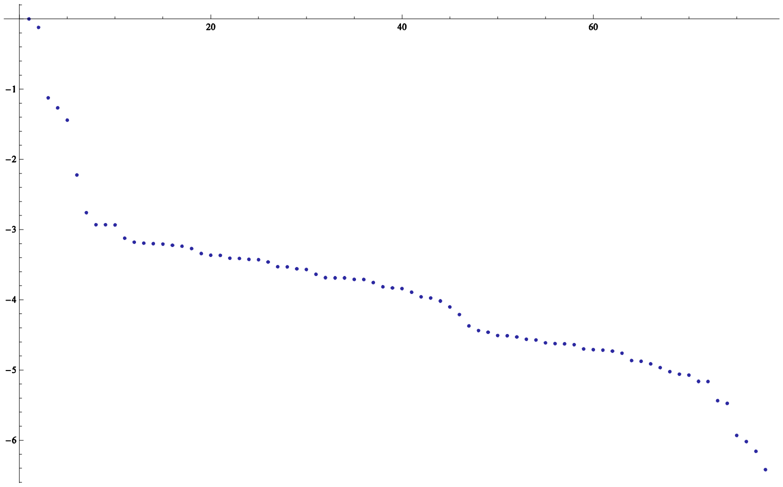

Figure 19: Sorted logarithms of correlations

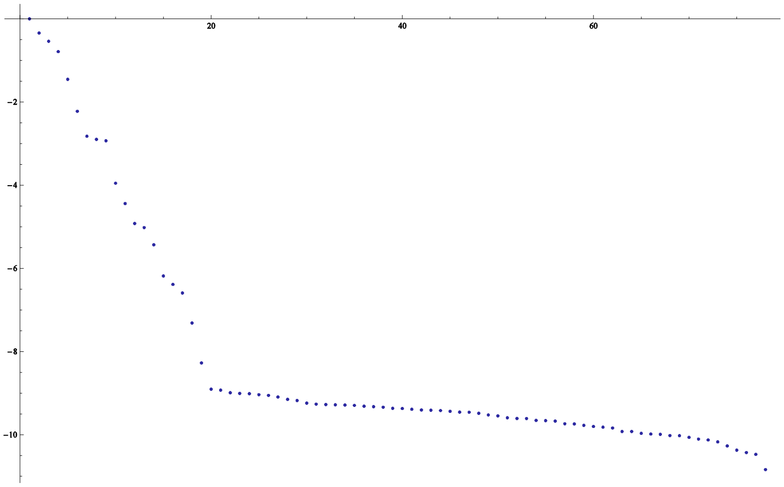

Figure 20: Sorted logarithms of sufficiencies

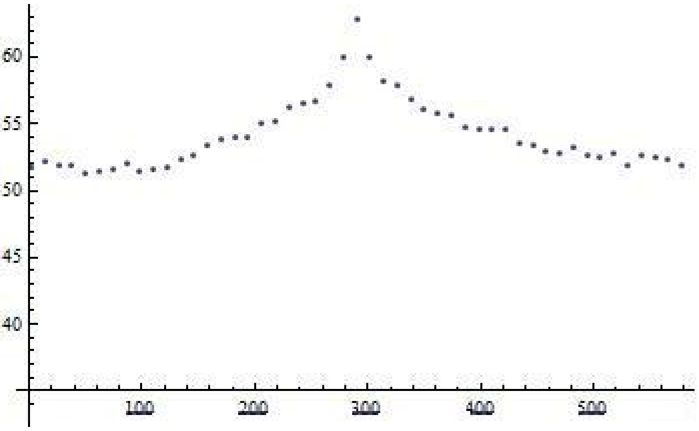

Figure 21: Sufficiency vs lag, wide sweep

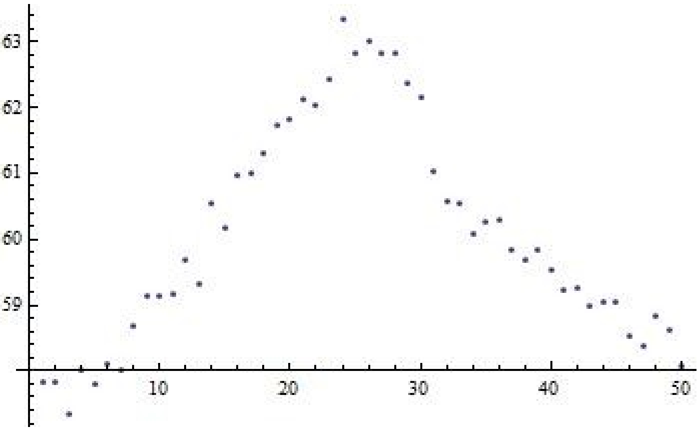

Figure 22: Sufficiency vs lag, narrow sweep

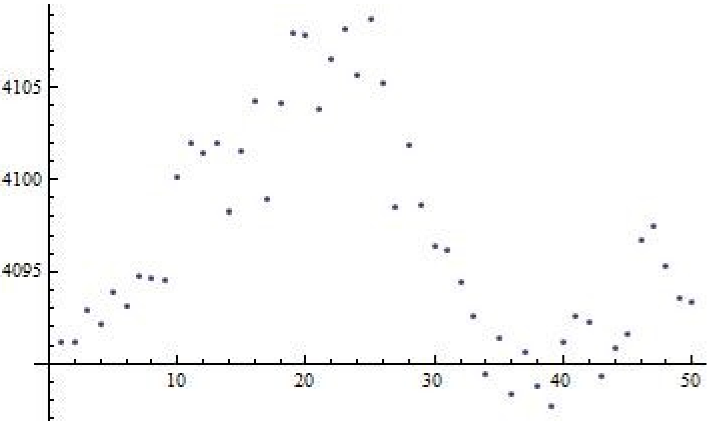

Figure 23: Sufficiency vs lag, multi-period cycles

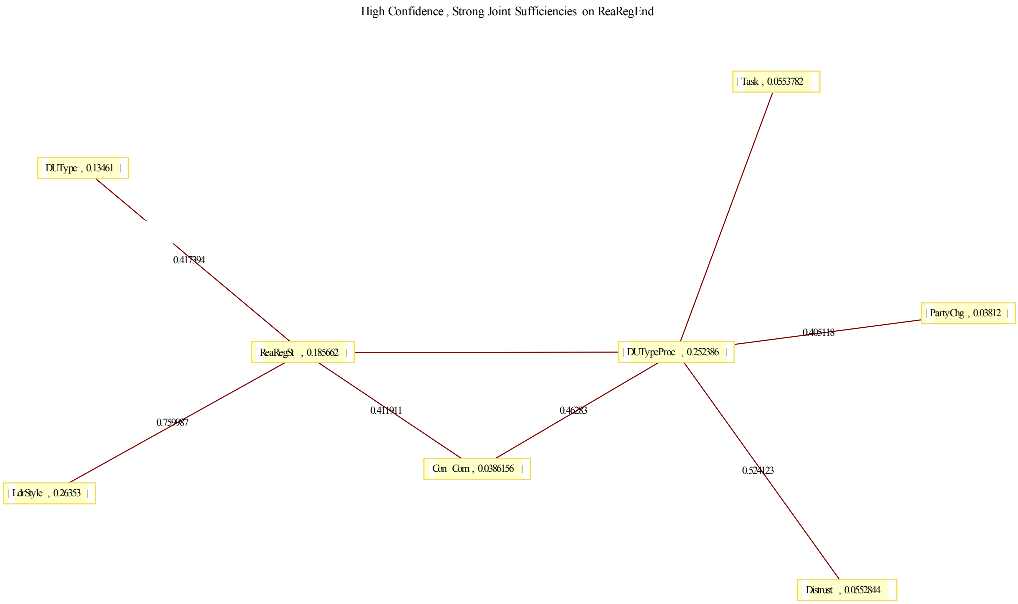

Figure 24: Information graph of political variables

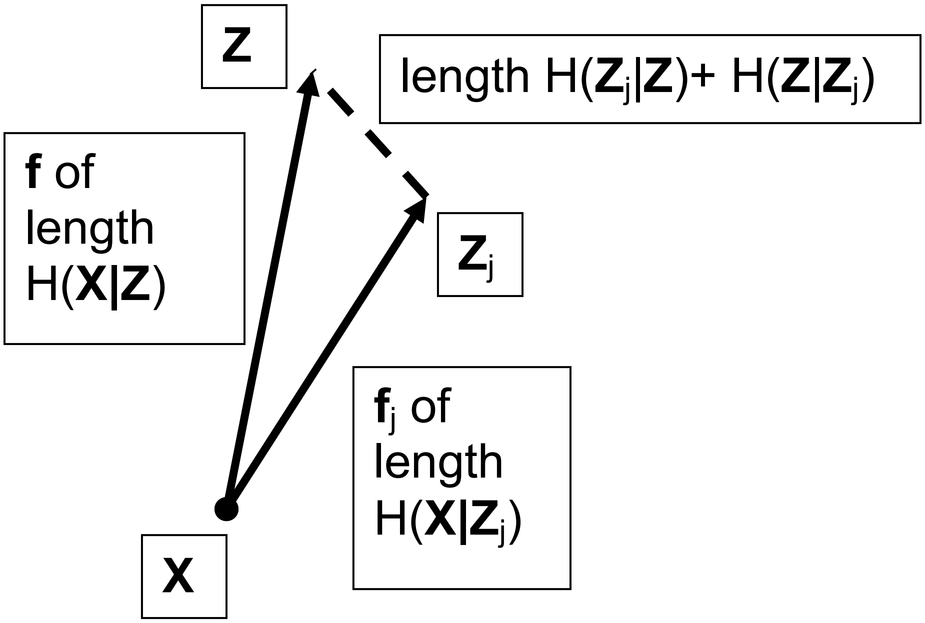
Z
length H($\mathbf{Z}_j$|$\mathbf{Z}$)+ H($\mathbf{Z}$|$\mathbf{Z}_j$)
f of length H(**X**|**Z**)
$\mathbf{Z}_j$
$\mathbf{f}_j$ of length H(**X**|$\mathbf{Z}_j$)
X

Figure 25: Functions as vectors in information space

Bibliography

[1] H. Abarbanel. *Analysis of Observed Chaotic Data.* Springer 1995

[2] Abayomi, K.A., Lall, U., and de la Pena, V.H. Copula based independent component analysis. Working paper, Duke University and Columbia University. http://ssrn.com/abstract=1028822 retrieved 2007

[3] Affenzeller, M., Wagner, S. and Winkler, S. Goal-Oriented Preservation of Essential Genetic Information by Offspring Selection. In *Genetic and Evolution Computation Conference (GECCO)* (Washington DC 2005) ACM Press

[4] Aguirre, A. and Coello, C. Mutual Information-based Fitness Functions for Evolutionary Circuit Synthesis. In *Congress on Evolutionary Computation (CEC)* (Portland OR 2004) IEEE 1309-1316

[5] Altenberg, L. The Evolution of Evolvability in Genetic Programming. in *Advances in Genetic Programming*, MIT Press, 1994, pp 47-94

[6] Altenberg, L. The Schema Theorem and Price's Theorem. in Whitley, L., ed. *Foundations of Genetic Algorithms 2.* Morgan Kauffman, San Mateo 1993

[7] Amari, S. and Nagaoka, H. Methods of information geometry. *Translations of mathematical monographs, 191*, American Mathematical Society 2000

[8] Bassett, J., Potter, M. and DeJong, K. Applying Price's Equation to Survival Selection. in *GECCO* (Washington DC 2005) ACM 1371 – 1378

[9] Baudrillard, J. *Cool Memories*. Verso, 1990

[10] Bennett, C., Gacs, P., Li, M., Vitanyi, P. and Zurek, W. Information Distance. *IEEE Trans. on Information Theory, 44* (4) (1998) 1407-1423

[11] Bialek, W., Nemenman, I. and Tishby, N. Predictability, Complexity, and Learning. *Neural Computation, 13* (2001) MIT 2409–2463

[12] Bongard, J. and Lipson, H. Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences, 104* (24) (2007)

[13] Borenstein, Y. and Poli, R. Information landscapes. in *GECCO* (Washington DC 2005) 1515–1522

[14] Box, G. and Jenkins, F. *Forecasting and Control*. Holden-Day 1976

[15] Evolution in the Information Age: Rediscovering the Nature of the Organism

[16] Brooks, D. Evolution in the Information Age: Rediscovering the Nature of the Organism. *SEED Journal,1* (1) (2001)

[17] Card, S. Information Distance based Fitness and Diversity Metrics. in *GECCO* (Montreal, Quebec 2010) Workshop on Entropy, Information and Complexity

[18] Card, S. and Mohan, C. Multiobjective information theoretic ensemble selection. in *SPIE Defense, Security and Sensing* (Orlando FL 2009)

[19] Card, S. and Mohan, C., An application of information theoretic selection to evolution of models with continuous valued inputs. in *Genetic Programming Theory & Practice (GPTP) VI* (Ann Arbor MI 2008) Springer 29-42

[20] Card, S. and Mohan, C. Towards an information theoretic framework for genetic programming. in *GPTP V* (Ann Arbor MI 2007) Springer 87–106

[21] Card, S. and Mohan, C. Ensemble selection for evolutionary learning with information theory and Price's Theorem. in *GECCO* (Seattle WA 2006) poster paper 103

[22] Card, S. and Mohan, C. Information Theoretic Indicators of Fitness, Relevant Diversity and Pair Potential. in *CEC* (Edinburgh 2005) IEEE vol 3 pp 2545–2552

[23] Card, S. Time Series Prediction by Genetic Programming with Relaxed Assumptions in *Mathematica*. in *GECCO* (Seattle WA 2004) Graduate Student Workshop

[24] Card, S. Genetic Programming of Wavelet Networks for Time Series Prediction. in *GECCO* (Orlando FL 1999) Graduate Student Workshop

[25] Casti, J. *Reality Rules: I (Picturing the World in Mathematics – The Fundamentals).* Wiley-Interscience 1992

[26] Chechik, G. *An Information Theoretic Approach to the Study of Auditory Coding.* PhD thesis, Hebrew University 2003

[27] Ching, J., Wong, A. and Chan, K. Class-dependent discretization for inductive learning from continuous and mixed-mode data. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 17* (7) (1995) 641-651

[28] Cilibrasi, R. and Vitanyi, P. Clustering by Compression. *IEEE Trans. on Information Theory, 51* (4) (2005) 1523-1545

[29] Coombs, C. H., Dawes, R. M. & Tversky, A. *Mathematical Psychology: An Elementary Introduction*. Prentice-Hall, Englewood Cliffs NJ 1970

[30] Crutchfield, J. Information and its Metric. in *Nonlinear Structures in Physical Systems – Pattern Formation, Chaos and Waves*, Springer-Verlag 1990

[31] Daida, J. Towards Identifying Populations that Increase the Likelihood of Success in Genetic Programming. in *GECCO* (Washington DC 2005). ACM vol 2 pp 1627–1634

[32] Davy, M. and Doucet, A. Copulas: A new insight into positive time-frequency distributions. *IEEE Signal Processing Letters,10* (2003) 215-218

[33] Day, R. and Lamont, G. An Effective Explicit Building Block MOEA – The MOMGA-IIa. in *CEC* (Edinburgh 2005) IEEE paper 413

[34] Defaweux, A., Lenaerts, T., van Hemert, J. and Parent, J. Complexity Transitions in Evolutionary Algorithms: Evaluating the impact of the initial population. in *CEC* (Edinburgh 2005) IEEE paper 196

[35] Deignan, P., Meckl, P., and Franchek, M. The MI-RBFN: Mapping for Generalization. in *American Control Conference* (2002)

[36] Deza, M. and Deza, E. *Encyclopedia of Distances*. Springer-Verlag, Heidelberg 2009

[37] De Vylder, B. A Tiny GP System. in *GECCO* (2004) competition entry

[38] Fisher, R. The Genetical Theory of Natural Selection. At The Clarendon Press, Oxford, England 1930

[39] Fraundorf, P. Heat capacity in bits. *Amer. J. Phys.* **71** (11) (2003) 1142-1151

[40] Fraundorf, P. Information physics: From energy to codes (2003). arXiv:physics/9611022 retrieved 2007

[41] Fry, R., Smith, S. and Tyrrell, A. A Self-Adaptive Mate Selection Model for Genetic Programming. in *CEC* (Edinburgh 2005) IEEE paper 260

[42] Gershenfeld, N. and Weigend, A. *Time Series Prediction: Forecasting the Future and Understanding the Past.* Addison-Wesley 1994

[43] Goldberg, D. *The Design of Innovation: Lessons from and for Competent Genetic Algorithms*. Kluwer 2002

[44] Grunwald, P. and Vitanyi, P. Shannon Information and Kolmogorov Complexity. http://arxiv.org/abs/cs/0410002 (2004) retrieved 2009

[45] Gustafson, S. and Vanneschi, L. Crossover Based Tree Distance in Genetic Programming. *IEEE Trans. On Evolutionary Computation, 12* (4) (2008) 506-524

[46] Hansen and Ostermeier . Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. in *CEC* (1996) 312-317

[47] Harper, M. Escort evolutionary game theory. http://arxiv.org/abs/0911.1764v4 revised 2010 retrieved 2011

[48] Harper, M. Information geometry and evolutionary game theory. http://arxiv.org/abs/0911.1383v1 revised 2009 retrieved 2011

[49] Haussler, D. Part 1: Overview of the Probably Approximately Correct Machine Learning Framework. www.cbse.ucsc.edu/staff/haussler_pubs/smo.pdf retrieved 2005

[50] Holland, J. Adaptation in Natural and Artificial Systems. MIT Press 1992. 1st edition Univ. of Michigan Press 1975

[51] Jacobs, R. Increased rates of convergence through learning rate adaptation. *Neural Networks, 1* (4) (1988) Elsevier 295-307

[52] Jakulin, A. and Bratko, I., Quantifying and visualizing attribute interactions: an approach based on entropy. *Journal of Machine Learning Research* (2003)

[53] Jones, T. and Forrest, S. Fitness distance correlation as a measure of problem difficulty for genetic algorithms. in *6th International Conference on Genetic Algorithms* (1995)

[54] Keijzer, M., Babovic, V., Baptist, M. and Uthurburu, J. Determining Equations for Vegetation Induced Resistance using Genetic Programming. in *GECCO* (Washington DC 2005)

[55] Korns, M.F. Large-scale, time-constrained symbolic regression/classification. in *GPTP V* (Ann Arbor 2007) Springer 53-68

[56] Koza, J., Al-Sakran, S. and Jones, L. Cross-Domain Features of Runs of Genetic Programming Used to Evolve Designs for Analog Circuits. Optical Lens Systems, Controllers, Antennas, Mechanical Systems, and Quantum Computing Circuits. in *NASA/DoD Conference on Evolvable Hardware* (Washington DC 2005) IEEE 205-214

[57] Koza, J. Genetic Programming II: Automatic Discovery of Reusable Programs. MIT Press 1994

[58] Koza, J. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press 1992

[59] Kraskov, A. and Grassberger, P., MIC: Mutual Information Based Hierarchical Clustering. in Emmert-Streib, F. and Dehmer, M. ed. *Information Theory and Statistical Learning,* Springer, New York NY 2008 pp 101-123

[60] Kraskov, A., Stogbauer, H. and Grassberger, P. Estimating Mutual Information. *Phys. Rev. E, 69* (6) (2004) and Errata *Phys. Rev. E, 83* (1) (2011)

[61] Langdon, W. and Poli, R. Foundations of Genetic Programming. Springer-Verlag, Berlin 2002

[62] Learned-Miller, E. Hyperspacings and the Estimation of Information Theoretic Quantities. UMass Technical Report 04-104 (2004)

[63] Li, M. and Vitanyi, P. *An Introduction to Kolmogorov Complexity and its Applications*. Springer-Verlag, New York NY 2008

[64] Li, Ming, Chen, Xin, Li, Xin,Ma, Bin, and Vitanyi, Paul. The similarity metric. in *14th ACM-SIAM Symposium on Discrete Algorithms* (2003) 863–872.

[65] Linsker et al, cited in "Self-Organizing Systems III: Information Theoretic Models," in Haykin, S. *Neural Networks: A Comprehensive Foundation*, IEEE 1994 pp 444ff

[66] Liu, Y., Yao, X., Zhao, Q. and Higuchi, T. Evolving a Cooperative Population of Neural Networks by Minimizing Mutual Information. in *CEC* (Korea 2001) IEEE 384-389

[67] Ma, Jian and Sun, Zengqi. Mutual Information Is Copula Entropy. *Tsinghua Science and Technology, 16* (1) (2011) 51-54

[68] Malago, L., Matteucci, M. and DalSeno, B. An Information Geometry Perspective on EDAs: Boundary Analysis. in *GECCO* (Atlanta GA 2008)

[69] McGill, W.J. Multivariate information transmission. *Psychometrika 19*, (1954) 97-116

[70] Meir, R. and Ratsch, G. An Introduction to Boosting and Leveraging. in *Advanced lectures on machine learning,* Springer-Verlag 2003

[71] Moraglio, A. Towards a Geometric Unification of Evolutionary Algorithms. PhD thesis, University of Essex, 2007

[72] Muharram, M.A. and Smith, G.D. (2004). Evolutionary feature construction using information gain and gini index. In *7th European Conference on Genetic Programming (EuroGP)* (Coimbra, Portugal 2004) *LNCS*, 3003 Springer-Verlag 379–388

[73] Oakley, H. Two Scientific Applications of Genetic Programming: Stack Filters and Non-Linear Equation Fitting to Chaotic Data. in Kinnear, K. ed. *Advances in Genetic Programming*, MIT Press 1994

[74] O'Reilly, U. Using a Distance Metric on Genetic Programs to understand Genetic Operators. MIT AI Lab 1997

[75] Ott, E., Sauer,T. and Yorke, J. *Coping with Chaos: Analysis of Chaotic Data and the Exploitation of Chaotic Systems*. Wiley 1994

[76] Paninski, L. Estimation of Entropy and Mutual Information. *Neural Computation, 15*, MIT 2003 pp. 1191–1253

[77] Potter, M., Bassett, J. and DeJong, K. Visualizing Evolvability with Price's Equation. in *CEC* (Canberra Australia 2003)

[78] Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T., *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, Cambridge England 1988 p 634

[79] Price, G. Selection and Covariance. *Nature, 227* (1970) 520-521

[80] Principe J., Fisher III, and Xu D. Information Theoretic Learning. in *Unsupervised Adaptive Filtering*, Wiley-Interscience 2000

[81] Radcliffe, N. The Algebra of Genetic Algorithms. *Annals of Maths and Artificial Intelligence, 10* (1994)

[82] Radcliffe, N. Genetic set recombination. in *Foundations of Genetic Algorithms 2*, Morgan-Kauffman 1993

[83] Rissanen, J. *Information and Complexity in Statistical Modeling*. Springer, New York NY 2007

[84] Rodriguez-Vazquez, K., Fonseca, C. and Fleming, P. Multiobjective Genetic Programming: A System Identification Application. 1997

[85] Roff, D. Defining fitness in evolutionary models. *Journal of Genetics, 87* (4) (2008) 339-348

[86] Rosca, J. Entropy Driven Adaptive Representation. in *Workshop on GP* (1995)

[87] Schmidt, M. and Lipson, H. *Learning Noise.* in *GECCO* (London England 2007)

[88] Schoenauer, M., Larranaga, P. and Lozano, J. ed. Special Issue on Estimation of Distribution Algorithms. *Evolutionary Computation*, MIT Press 2005

[89] Seo, Dong-Il and Moon, Byung-Ro. An information-theoretic analysis on the interactions of variables in combinatorial optimization problems. *Evolutionary Computation, 15*(2) (2007)

[90] Shalizi, C.R. and Crutchfield, J.P. Computational Mechanics: Pattern and Prediction, Structure and Simplicity. *Journal of Statistical Physics, 104* (2001) 817-881

[91] Shalizi, C. 2003. Complexity, Entropy and the Physics of gzip. http://www.cscs.umich.edu/~crshalizi/notabene/cep-gzip.html revised 2003 retrieved 2008

[92] Shannon, C. A Mathematical Theory of Communication. *The Bell System Technical Journal, 27* (1948) 379-423 and 623-656

[93] Sklar, A. Fonctions de répartition à n dimensions et leurs marges*. Publ. Inst. Statist. Univ. Paris, 8* (1959) 229-231

[94] Smith, R. An Iterative Mutual Information Histogram Technique for Linkage Learning in Evolutionary Algorithms. In *CEC* (Edinburgh 2005) IEEE paper 133

[95] Smith, T., Husbands, P., Layzell, P. and O'Shea, M. Fitness Landscapes and Evolvability. *Evolutionary Computation, 10*(1) (2002) 1-34

[96] Sprott, J. *Chaos and Time-Series Analysis*. Oxford University Press, New York 2003

[97] Sprott, J. Algebraically Simple Chaotic Flows. *International Journal of Chaos Theory and Applications, 5*(2)

[98] Sprott, J. (1997) Simplest dissipative chaotic flow. *Physics Letters A 228*, Elsevier Science 1997

[99] Stephens, C.R. and Vargas, J.M. Effective Fitness as an Alternative Paradigm for Evolutionary Computation. *Genetic Programming and Evolvable Machines, 1* (4) (2000)

[100] Suleman, H. *Genetic Programming in* Mathematica. MS thesis, University of Durban-Westville 1997

[101] Szathmary, E. and Smith, J. The major evolutionary transitions. *Nature, 395* (1995)

[102] Takens, F. Detecting Strange Attractors in Turbulence. *Lecture Notes in Mathematics*. Springer, Berlin 1981

[103] Tishby, N., Pereira, F.C. and Bialek,W. The Information Bottleneck method. in *The 37th annual Allerton Conference on Communication, Control, and Computing* (1999) 368-377

[104] Torkkola, K. On Feature Extraction by Mutual Information Maximization. in *International Conference on Acoustics, Speech, and Signal Processing* (2002) IEEE

[105] Toussaint, M. Notes on information geometry and evolutionary processes. http://arxiv.org/abs/nlin/0408040v1 submitted 2004, version on author's site revised 2006

[106] Vanneschi, L. and Tomassini, M. A Study On Fitness Distance Correlation and Problem Difficulty for Genetic Programming. in *GECCO* (New York NY 2002) 307-310

[107] Vos Post, J. http://necsi.edu/wiki/index.php/Evolutionary_channel_capacity

[108] Watkins, C. Selective Breeding Analysed as a Communication Channel: Channel Capacity as a Fundamental Limit on Adaptive Complexity. in *Symbolic and Numeric Algorithms for Scientific Computing: Proc. of SYNASC* (2008) IEEE 514-518

[109] Watkins, C. The Channel Capacity of Evolution: Ultimate Limits on the Amount of Information Maintainable in the Genome. in *3rd International Conference on Bioinformatics of Genome Regulation and Structure* (Novisibirsk, Russia 2002) vol 2 pp 58-60

[110] Wineberg, M. and Oppacher, F. The Underlying Similarity of Diversity Measures Used in Evolutionary Computation. in *GECCO* (Chicago IL 2003)

[111] Wolpert, D. and Macready, W. No free lunch theorems for search. Technical Report SFI-TR-95-02-010, Santa Fe Institute 1995 *et seq*

[112] Wright, S. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. in *6th International Congress on Genetics* (1932) 355–366

[113] http://www.fourmilab.ch/hotbits

[114] http://www.random.org

Vita

NAME OF AUTHOR: Stuart William Card

PLACE OF BIRTH: Herkimer, New York

DATE OF BIRTH: May 28, 1963

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

Syracuse University, Syracuse, New York

Utica College, Utica, New York

United States Naval Academy, Annapolis, Maryland

DEGREES AWARDED:

Master of Science in Electrical Engineering, 1990, Syracuse University

Bachelor of Science in Computer Science, 1985, Utica College

AWARDS AND HONORS:

Order of the Purple Cross, York Rite Sovereign College of North America, 2011

US Patent # 7,203,732, “Flexible remote data mirroring” (coinventor), 2006

Mohawk Valley Engineering Company of the Year (cofounder), 2002

Mohawk Valley 40 Under 40, 2001

UK Patent # GB 2 251 502 B, “Data-loss prevention products” (coinventor), 1993

Rensselaer Math & Science Medal, 1981

Bausch & Lomb Science Award, 1981

PROFESSIONAL EXPERIENCE:

Principal Engineer, Northrop Grumman, 2010-PRESENT

Chief Scientist, Critical Technologies Inc., 1989-PRESENT

Teaching Assistant, Computer Science, Syracuse University, 1996-1997

Adjunct Lecturer, Comp. Science, Physics & Calculus, Utica College, 1991-1995

Adjunct Lecturer, Computer Science, SUNY Institute of Technology, 1990

Program Engineer, GE Aerospace Electronics, 1985-1989

United States Navy (Honorable Discharge), 1981-1985

THESIS-RELATED PROFESSIONAL ACTIVITIES:

World Conference on Computational Intelligence (WCCI) (Brisbane, Australia 2012) Workshop on Entropy, Information and Complexity (proposed organizer)

GECCO (Montreal, Quebec 2010) Workshop on Entropy, Information and Complexity (organizer)

Multi-objective information theoretic ensemble selection. in *Mohawk Valley Technology Symposium* (Utica NY 2010) (seminar presenter, NY State Society of Professional Engineers professional development hours awarded to attendees)

2 invited reviews of related papers for *Evolutionary Computation* journal

1 invited review of related paper for *IEEE Trans. on Information Theory*

OTHER PROFESSIONAL ACTIVITIES:

CET/EET Curricular Advisory Committee, SUNY Institute of Technology

past Trustee, Central NY Chapter, U.S. Naval Academy Alumni Assoc.

past chair, Rome-Utica Chapter, Association for Computing Machinery (ACM)

past chair, Mohawk Valley Chapter, Computer Society of the IEEE (IEEE-CS)

past chair, Mohawk Valley Section, Inst. of Elect. & Electronics Eng's (IEEE)

past director, Mohawk Valley Engineers Executive Council (MVEEC)

past director, Mohawk Valley Applied Technology Commission (MVATC)

THESIS-RELATED PUBLICATIONS WITH C. MOHAN:

Multiobjective information theoretic ensemble selection. in *SPIE Defense, Security and Sensing* (Orlando FL 2009) (invited paper)

An application of information theoretic selection to evolution of models with continuous valued inputs. In *Genetic Programming Theory & Practice (GPTP) VI*, (Ann Arbor MI 2008) Springer 29-42

Towards an information theoretic framework for genetic programming. In *GPTP V* (Ann Arbor MI 2007) Springer 87–106

Ensemble selection for evolutionary learning with information theory and Price's Theorem. In *Genetic and Evolutionary Computation Conference (GECCO)* (2006) poster paper 103

Information Theoretic Indicators of Fitness, Relevant Diversity and Pair Potential. in *Congress on Evolutionary Computation (CEC)* (Edinburgh 2005) IEEE vol 3 pp 2545–2552

THESIS-RELATED PUBLICATIONS:

Information Distance based Fitness and Diversity Metrics. In *GECCO* (Montreal, Quebec 2010) Workshop on Entropy, Information and Complexity

Time Series Prediction by Genetic Programming with Relaxed Assumptions in *Mathematica*. in *GECCO* (2004) Graduate Student Workshop

Genetic Programming of Wavelet Networks for Time Series Prediction. in *GECCO* (1999) Graduate Student Workshop

OTHER PUBLICATIONS:

Real-Time Resource Allocation Co-Processor. in *IEA/AEI* (Syracuse NY 2011)

Concurrent Multipath Routing & Transport in a Mobile Wireless Gateway. in *Military Communications Conference (MILCOM)* (Monterey CA 2004) IEEE

Transparent Gateway for Delayed Intermittent Asymmetric Wireless Networks. in *MILCOM* (Monterey CA 2004)

Data Mining a High-Speed Bursty Stream on a Limited Buffer in Pseudo-Stationary States. in *International Workshop on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS)* (Lviv, Ukraine 2003) IEEE (coauthor)

Towards Data Mining Temporal Patterns for Anomaly Intrusion Detection Systems. in *IDAACS* (Lviv, Ukraine 2003) IEEE (coauthor)

Service Specific Coordination Function for Transparent Assured Delivery with AAL5. in *MILCOM* (Atlantic City NJ 1999) IEEE (coauthor)

Bridging Legacy Protocols with DAMA/ATM on RF Media. in *Dual-Use Technologies and Applications Conference* (Syracuse NY 1996) IEEE Mohawk Valley Section

Neural Network Range Mapper Hardware Prototype. in *Dual-Use Technologies and Applications Conference* (Syracuse NY 1996) IEEE Mohawk Valley Section

Neural Network Based Optic Flow Extraction. in *Dual-Use Technologies and Applications Conference* (Utica NY 1994) IEEE Mohawk Valley Section

PUBLICATIONS IN DEVELOPMENT:

Tactical/Airborne Network Authentication: Analysis of Zero Knowledge Proof Based Alternatives. submitted to *MILCOM* (Baltimore MD 2011) IEEE (coauthor)

Tactical/Airborne Network Resource Access Control: Zero Knowledge Proofs of Varinymous Credentials for Agent Capability Security. submitted to *MILCOM* (Baltimore MD 2011) IEEE

Genetic Programming with Information Theoretic Evaluations of Ensembles. (proposed title of journal article invited by editor of *Genetic Programming and Evolvable Machines* journal)

Entropy, Information and Complexity in Evolutionary Learning. (proposed title of textbook invited by editor at Springer)