Time's Barbed Arrow: Irreversibility, Crypticity, and Stored Information

James P. Crutchfield,[1,2,*] Christopher J. Ellison,[1,†] and John R. Mahoney[1,‡]

[1]*Complexity Sciences Center and Physics Department, University of California at Davis, One Shields Avenue, Davis, California 95616, USA*

[2]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA*

(Received 6 February 2009; published 28 August 2009)

We show why the amount of information communicated between the past and future—the excess entropy—is not in general the amount of information stored in the present—the statistical complexity. This is a puzzle, and a long-standing one, since the former describes observed behavior, while optimal prediction requires the latter. We present a closed-form expression for the excess entropy in terms of optimal causal predictors and retrodictors—both ϵ machines of computational mechanics. This leads us to two new system invariants: causal irreversibility—the asymmetry between the causal representations—and crypticity—the degree to which a process hides its state information.

DOI: 10.1103/PhysRevLett.103.094101 PACS numbers: 05.45.Tp, 02.50.Ey, 05.40.−a, 89.70.Cf

Constructing a theory can be viewed as our attempt to extract from measurements a system's hidden organization. This suggests a parallel with decryption whose goal is to reveal internal correlations within an encrypted data stream [1]. The hidden message is revealed only to a recipient with the correct codebook. This is essentially the circumstance a scientist faces when building a model from measurements: What are the hidden states and dynamic in the observed data?

In this view, the now-long history in nonlinear dynamics of reconstructing models from time series [2,3] is cast as a self-decoding problem, where the information used to build a model is only that available in the observed process. That is, no "sideband" communication, prior knowledge, or disciplinary assumptions are allowed. Nature speaks for herself only through the data she willingly gives up.

Here, we show that the parallel is more than metaphor: building a model corresponds directly to decrypting the hidden state information in measurements. The results show why predicting and modeling are, at one and the same time, distinct and intimately related. Along the way, we clarify the role and types of information in prediction and modeling. We show how to measure the degree of hidden information and identify a new kind of statistical irreversibility.

A process $\Pr(\overleftarrow{X}, \overrightarrow{X})$ is a communication channel with a fixed input distribution $\Pr(\overleftarrow{X})$: It transmits information from the past $\overleftarrow{X} = \ldots X_{-3}X_{-2}X_{-1}$ to the future $\overrightarrow{X} = X_0X_1X_2\ldots$ by storing it in the present. Here, X_t is the discrete random variable for the measurement outcome at time t, such as the observed z component of a spin or the symbolic dynamics of a chaotic system.

Our goal is also simply stated: We wish to predict the future using information from the past. At root, a prediction is probabilistic, specified by a distribution of possible futures $\overrightarrow{X}$ given a particular past $\overleftarrow{x}$: $\Pr(\overrightarrow{X}|\overleftarrow{x})$. At a minimum, a good predictor needs to capture all of the information I shared between past and future: $\mathbf{E} = I[\overleftarrow{X}; \overrightarrow{X}]$—the process's excess entropy ([4], and references therein).

Consider now the goal of modeling: build a representation that not only allows good prediction, but also expresses the mechanisms that produce a system's behavior. To build a model of a process, computational mechanics [5] introduced an equivalence relation $\overleftarrow{x} \sim \overleftarrow{x}'$ to group all histories that give rise to the same prediction—resulting in a map from pasts to the causal states: $\epsilon(\overleftarrow{x}) = \{\overleftarrow{x}' : \Pr(\overrightarrow{X}|\overleftarrow{x}) = \Pr(\overrightarrow{X}|\overleftarrow{x}')\}$. A process's causal states, $\mathcal{S} = \Pr(\overleftarrow{X}, \overrightarrow{X})/\sim$, partition the space $\overleftarrow{\mathbf{X}}$ of pasts into sets that are predictively equivalent. The set of causal states can be discrete, fractal, or continuous. State-to-state transitions are denoted by matrices $T^{(x)}_{\mathcal{S}\mathcal{S}'}$ whose elements give the probability of transitioning from one state $\mathcal{S}$ to the next $\mathcal{S}'$ on seeing measurement value x. The resulting model, consisting of the causal states and transitions, is called the process's ϵ machine.

Causal states have the Markovian property that they render the past and future statistically independent; they shield the future from the past [5]: $\Pr(\overleftarrow{X}, \overrightarrow{X}|\mathcal{S}) = \Pr(\overleftarrow{X}|\mathcal{S})\Pr(\overrightarrow{X}|\mathcal{S})$. In this way, the causal states give a structural decomposition of the process into conditionally independent modules. Moreover, they are optimally predictive [5] in the sense that knowing which causal state a process is in is just as good as having the entire past: $\Pr(\overrightarrow{X}|\mathcal{S}) = \Pr(\overrightarrow{X}|\overleftarrow{X})$. In other words, causal shielding is equivalent to the fact [5] that the causal states capture all of the information shared between past and future: $I[\mathcal{S}; \overrightarrow{X}] = \mathbf{E}$.

Naturally, there can be alternative models; denote their states $\mathcal{R}$. Consider the subset of these that are optimally predictive—those for which $I[\hat{\mathcal{R}}; \overrightarrow{X}] = \mathbf{E}$, where we denoted their states as $\hat{\mathcal{R}}$. Out of all optimally predictive models, the ϵ machine captures the minimal amount of information that a process must store in order to commu-

nicate all of the excess entropy from the past to the future. This is the statistical complexity [5]: $C_\mu \equiv H[\mathcal{S}] \leq H[\hat{\mathcal{R}}]$, where μ reminds us of the dependence on the dynamical system's underlying invariant measure. In short, $\mathbf{E}$ is the effective information transmission capacity of the process, viewed as a channel, and C_μ is the sophistication of that channel.

In addition to $\mathbf{E}$ and C_μ, another key (and historically prior) invariant for dynamical systems and stochastic processes is the entropy rate h_μ—a process's degree of intrinsic randomness [6]. Importantly, the ϵ machine immediately gives two of these three important invariants: a process's rate (h_μ) of producing information and the amount (C_μ) of historical information stored in doing so.

To date, $\mathbf{E}$ cannot be as directly calculated as the entropy rate and the statistical complexity. One practical consequence is that it is difficult to know when one has obtained a good estimate of $\mathbf{E}$. These are truly unfortunate, since excess entropy, and related mutual information quantities, are widely used diagnostics for processes, having been applied to detect the presence of organization in dynamical systems [2,3,7,8], in spin systems [9,10], in neurobiological systems [11,12], and even in language, to mention only a few applications. For example, in natural language the excess entropy appears to diverge with string length L as $\mathbf{E} \propto L^{1/2}$, reflecting the long-range and strongly nonergodic organization necessary for human communication [13,14].

This state of affairs has been a major impediment to understanding the relationships between modeling and predicting and, more concretely, the relationships between (and even the interpretation of) a process's basic invariants—h_μ, C_μ, and $\mathbf{E}$ [15]. Here, we clarify these issues by deriving explicit expressions for $\mathbf{E}$ in terms of the ϵ machine and C_μ, providing a unified information-theoretic analysis of stationary processes.

The above development of ϵ machines concerns using the past to predict the future. But what about the opposite, using the future to retrodict the past? Usually, one thinks of successive measurements occurring as time increases. Now, consider scanning the measurement variables not in the forward time direction, but in the reverse time direction. The computational mechanics formalism is essentially unchanged, though its meaning and notation need to be augmented.

With this in mind, the previous mapping from pasts to causal states is denoted ϵ^+ and it gave, what we will call, the predictive causal states $\mathcal{S}^+$. When scanning in the reverse direction, we have a new relation, $\overrightarrow{x} \sim^- \overrightarrow{x}'$, which groups futures that are equivalent for the purpose of retrodicting the past: $\epsilon^-(\overrightarrow{x}) = \{\overrightarrow{x}' : \Pr(\overleftarrow{X}|\overrightarrow{x}) = \Pr(\overleftarrow{X}|\overrightarrow{x}')\}$. It gives the retrodictive causal states $\mathcal{S}^- = \Pr(\overleftarrow{X}, \overrightarrow{X})/\sim^-$. And, not surprisingly, we must also distinguish a process's forward-scan ϵ machine M^+ from its reverse-scan ϵ machine M^-. They assign corresponding entropy rates, h_μ^+ and h_μ^-, and statistical complexities, $C_\mu^+ \equiv H[\mathcal{S}^+]$ and $C_\mu^- \equiv H[\mathcal{S}^-]$, respectively, to the process.

Now we are in a position to ask some questions. Perhaps the most obvious is, In which time direction is a process most predictable? The answer is that a stationary process is equally predictable in either [5]: $h_\mu^- = h_\mu^+$. Somewhat surprisingly, though, the effort involved in doing so need not be the same [16]: $C_\mu^- \neq C_\mu^+$. Naturally, $\mathbf{E}$ is mute on this score, since the mutual information I is symmetric in its variables [4].

The relationship between predicting and retrodicting a process, and ultimately $\mathbf{E}$'s role, requires teasing out how the states of the forward and reverse ϵ machines capture information from the past and the future. To do this we must analyze a four-variable mutual information: $I[\overleftarrow{X}; \overrightarrow{X}; \mathcal{S}^+; \mathcal{S}^-]$. A large number of expansions of this quantity are possible. A systematic development follows from Ref. [17] which showed that Shannon entropy $H[\cdot]$ and mutual information $I[\cdot;\cdot]$ form a signed measure over the space of events.

Using an information measure expansion, it turns out there are 15 possible relationships to consider for $I[\overleftarrow{X}; \overrightarrow{X}; \mathcal{S}^+; \mathcal{S}^-]$. Fortunately, this greatly simplifies in the case of using an ϵ machine to represent a process: There are only five relationships. (See Fig. 1.) Simplified in this way, we are left with our main results which, due to the preceding effort, are particularly transparent.

Theorem 1. Excess entropy is the mutual information between the predictive and retrodictive causal states:

$$\mathbf{E} = I[\mathcal{S}^+; \mathcal{S}^-]. \tag{1}$$

This is obtained via a simultaneous reduction of the four-variable mutual information into $I[\overleftarrow{X}; \overrightarrow{X}]$ and $I[\mathcal{S}^+; \mathcal{S}^-]$. Notably, the process's channel utilization $\mathbf{E} = I[\overleftarrow{X}; \overrightarrow{X}]$ between the past and future is the same as the utilization between the forward and reverse ϵ-machine states. Moreover, the predictive statistical complexity is given by $C_\mu^+ = \mathbf{E} + H[\mathcal{S}^+|\mathcal{S}^-]$ and the retrodictive statistical complexity by $C_\mu^- = \mathbf{E} + H[\mathcal{S}^-|\mathcal{S}^+]$.

Theorem 1 and the companion results give an explicit connection between a process's excess entropy and its causal structure—its ϵ machines. More generally, the rela-

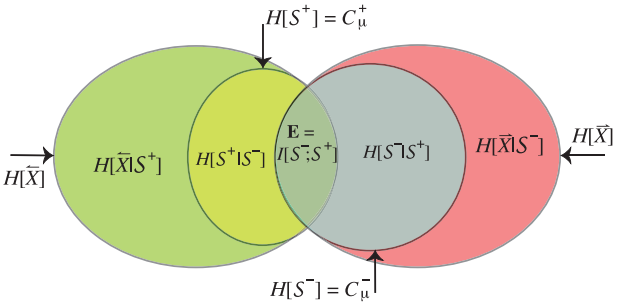

FIG. 1 (color). ϵ-machine information diagram for stationary stochastic processes. A schematic, the diagram only shows the set-theoretic relationships.

tionships directly tie mutual information measures of observed sequences to a process's structure. They will allow us to probe the properties that control how closely observed statistics reflect a process's hidden structure, that is, the degree to which observed behavior directly reflects internal state information.

At this point we have two separate ϵ machines: one for predicting and one for retrodicting. We will now show that one can do better, by combining causal information from the past and future. Consider scanning a realization, $\overleftrightarrow{x} = \overleftarrow{x}_t \overrightarrow{x}_t$, of the process in the forward direction—seeing histories $\overleftarrow{x}_t$ and noting the series of causal states $\mathcal{S}_t^+ = \epsilon^+(\overleftarrow{x}_t)$. Now change direction. What reverse causal state is one in? This is $\mathcal{S}_t^- = \epsilon^-(\overrightarrow{x}_t)$. We describe the action of changing scan direction with the bidirectional machine $M^\pm$, which is given by the equivalence relation $\sim^\pm$:

$$\epsilon^\pm(\overleftrightarrow{x}) = \{(\overleftarrow{x}', \overrightarrow{x}') : \overleftarrow{x}' \in \epsilon^+(\overleftarrow{x}) \quad \text{and} \quad \overrightarrow{x}' \in \epsilon^-(\overrightarrow{x})\}$$

and has causal states $\boldsymbol{\mathcal{S}}^\pm = \Pr(\overleftarrow{X}, \overrightarrow{X})/\sim^\pm \subset \boldsymbol{\mathcal{S}}^+ \times \boldsymbol{\mathcal{S}}^-$. That is, the bidirectional causal state the process is in at time t is $\mathcal{S}_t^\pm = (\epsilon^+(\overleftarrow{x}_t), \epsilon^-(\overrightarrow{x}_t))$. The amount of stored information needed to optimally predict and retrodict a process is $M^\pm$'s statistical complexity: $C_\mu^\pm \equiv H[\mathcal{S}^\pm] = H[\mathcal{S}^+, \mathcal{S}^-]$.

From the immediately preceding results we obtain the following simple, useful relationship: $\mathbf{E} = C_\mu^+ + C_\mu^- - C_\mu^\pm$. This suggests a wholly new interpretation of the excess entropy—in addition to the original three reviewed in Ref. [4]: $\mathbf{E}$ is exactly the difference between these statistical complexities. Moreover, only when $\mathbf{E} = 0$ does $C_\mu^\pm = C_\mu^+ + C_\mu^-$. The bidirectional machine is also efficient: $C_\mu^\pm \leq C_\mu^+ + C_\mu^-$. And we have the bounds: $C_\mu^+ \leq C_\mu^\pm$ and $C_\mu^- \leq C_\mu^\pm$. These inequalities express the compactness of the bidirectional machine in contrast to the pair of directional ϵ machines. This efficiency of representation is due to the redundancy in the predictive and retrodictive causal states.

We noted above that predicting and retrodicting may require different amounts of information storage ($C_\mu^+ \neq C_\mu^-$). It is helpful to use causal irreversibility to measure this asymmetry [16]: $\Xi \equiv C_\mu^+ - C_\mu^-$. With the above results, however, we see that $\Xi = H[\mathcal{S}^+|\mathcal{S}^-] - H[\mathcal{S}^-|\mathcal{S}^+]$. Note that irreversibility is also not controlled by $\mathbf{E}$, as the latter is scan-symmetric.

The relationship between excess entropy and statistical complexity established by Theorem 1 indicates that there are fundamental limitations on the amount of a process's stored information ($C_\mu^\pm$) directly present in observations ($\mathbf{E}$). We now introduce a measure of this: A process's crypticity is $\chi \equiv H[\mathcal{S}^+|\mathcal{S}^-] + H[\mathcal{S}^-|\mathcal{S}^+]$. This is the distance between a process's forward and reverse ϵ machines and expresses, most explicitly, the difference between prediction and modeling. To see this, we need the following connection.

Corollary 1. $M^\pm$'s statistical complexity is

$$C_\mu^\pm = \mathbf{E} + \chi. \tag{2}$$

Referring to χ as crypticity derives from this result: It is the amount of internal state information ($C_\mu^\pm$) not directly present in the observed sequence ($\mathbf{E}$). That is, a process hides χ bits of information.

If crypticity is low ($\chi \approx 0$), then much of the stored information is present in observed behavior: $\mathbf{E} \approx C_\mu^\pm$. However, when a process's crypticity is high, $\chi \approx C_\mu^\pm$, then little of its structural information is directly present in observations. Moreover, there are truly cryptic processes ($\mathbf{E} \approx 0$) that are highly structured ($C_\mu^\pm \gg 0$). Little or nothing can be learned from measurements about such processes' hidden organization.

The ϵ-machine information diagram of Fig. 1 encapsulates all of these results concisely by showing the key relationships between information production ($H[\overrightarrow{X}|\mathcal{S}^+]$ and $H[\overleftarrow{X}|\mathcal{S}^-]$), stored information ($C_\mu^+$ and C_μ^-), and excess entropy ($\mathbf{E} = I[\overleftarrow{X}; \overrightarrow{X}]$). Analyzing the 4-variable information diagram revealed a parsimonious relationship among the four variables, depicted as differently shaded ellipses. $H[\overleftarrow{X}]$ and $H[\overrightarrow{X}]$ (two largest ellipses) are the entropies of the past and future, respectively, which are the process's total information production. The information stored in the predictive ϵ machine M^+ is its statistical complexity: $C_\mu^+ \equiv H[\mathcal{S}^+]$ (small ellipse on left); likewise for M^-, $C_\mu^- \equiv H(\mathcal{S}^-)$ (small ellipse on right). The excess entropy $\mathbf{E}$ is the intersection of these sets; while the statistical complexity $C_\mu^\pm$ of the bidirectional machine $M^\pm$ is their union; the crypticity, their symmetric difference; and their signed difference, the causal irreversibility Ξ.

Consider an example that illustrates the typical process—cryptic and causally irreversible. This is the random insertion process (RIP) which generates a random bit with bias p. If that bit is a 1, then it outputs another 1. If the random bit is a 0, however, it inserts another random bit with bias q, followed by a 1.

Its forward ϵ machine, see Fig. 2(a), has three recurrent causal states $\mathcal{S}^+ = \{A, B, C\}$ and the transition matrices given there. Figure 2(b) gives M^- which has four recurrent causal states $\mathcal{S}^- = \{D, E, F, G\}$. We see that the ϵ machines are not the same and so the RIP is causally irreversible. A direct calculation gives $\Pr(\mathcal{S}^+) = \Pr(A, B, C) = (1, p, 1)/(p + 2)$ and $\Pr(\mathcal{S}^-) = \Pr(D, E, F, G) = (1, 1 - pq, pq, p)/(p + 2)$. If $p = q = 1/2$, for example, these give us $C_\mu^+ \approx 1.5219$ bits, $C_\mu^- \approx 1.8464$ bits, and $h_\mu = 3/5$ bits per measurement. The causal irreversibility is $\Xi \approx 0.3245$ bits.

Let us analyze its bidirectional machine, shown in Fig. 2(c) for $p = q = 1/2$. The interdependence between the forward and reverse states is given by

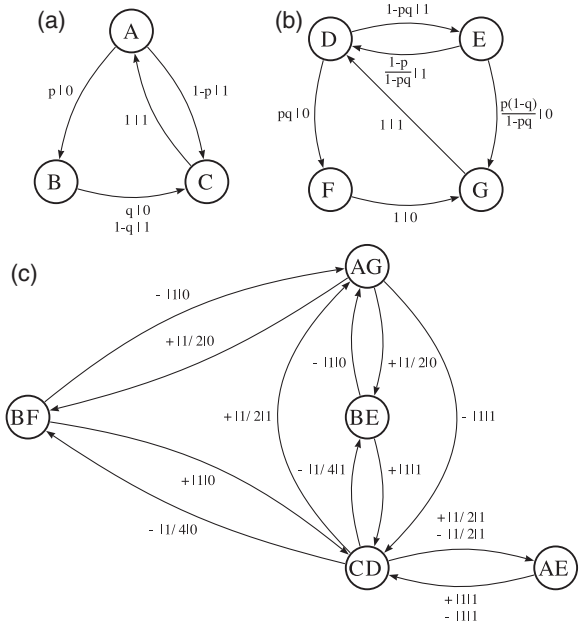

FIG. 2. Forward and reverse ϵ machines for the RIP: (a) M^+ and (b) M^-. Edge labels $t|x$ give the transition probabilities $t = T^{(x)}_{SS'}$. (c) The bidirectional machine $M^\pm$ for $p = q = 1/2$. Edge labels are prefixed with the scan direction $\{-, +\}$.

$$\Pr(\mathcal{S}^+, \mathcal{S}^-) = \frac{1}{(p+2)} \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{cccc} D & E & F & G \end{array} \\ \begin{pmatrix} 0 & 1-p & 0 & p \\ 0 & p(1-p) & pq & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \end{array}.$$

By way of demonstrating the exact analysis now possible, **E**'s closed-form expression for the RIP family is

$$\mathbf{E} = \log_2(p+2) - \frac{p\log_2 p}{p+2} - \frac{1-pq}{p+2} H\left(\frac{1-p}{1-pq}\right),$$

where $H(\cdot)$ is the binary entropy function. The first two terms on the right-hand side are C^+_μ and the last is $H[\mathcal{S}^+|\mathcal{S}^-]$.

Setting $p = q = 1/2$, one calculates that $\Pr(\mathcal{S}^\pm) = \Pr(AE, AG, BE, BF, CD) = (1/5, 1/5, 1/10, 1/10, 2/5)$. This and the joint distribution give $C^\pm_\mu = H[\mathcal{S}^\pm] \approx 2.1219$ bits, but an $\mathbf{E} = I[\mathcal{S}^+; \mathcal{S}^-] \approx 1.2464$ bits. That is, the excess entropy (the apparent information) is substantially less than the statistical complexities (stored information)—a rather cryptic process: $\chi \approx 0.8755$ bits.

To close, the main results establish that when $\chi > 0$ one cannot simply use sequence information directly to represent a process as storing **E** bits of information. We must instead store C_μ bits of information, building a causal model of the hidden state information. Why? Because typical processes encrypt their state information within their observed behavior. More particularly, observed information can be arbitrarily small ($\mathbf{E} \approx 0$) compared to the stored information (C_μ).

In deriving an explicit relationship between excess entropy and the ϵ machine, the framework puts prediction on an equal footing with modeling, allowing for a direct comparison between them [18]. Also, as we demonstrated with the RIP example, it gives a way to develop closed-form expressions for **E**. Finally and most generally, it reveals an intimate connection between unpredictability, irreversibility, crypticity, and information storage.

Practically, these results elucidate the difference between observed (mutual) information (**E**) and a process's stored information (C_μ). Analyzing a process only in terms of mutual information misses an arbitrarily large amount of a process's structure. When this happens, one concludes that a process is more random than it is and that it has little structure, when neither is true.

C. E. was partially supported by GAANN. The Network Dynamics Program funded by Intel Corporation also partially supported this work.

*chaos@cse.ucdavis.edu
†cellison@cse.ucdavis.edu
‡jrmahoney@ucdavis.edu

[1] C. E. Shannon, Bell Syst. Tech. J. **28**, 656 (1949).
[2] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, England, 2006), 2nd ed.
[3] J. C. Sprott, *Chaos and Time-Series Analysis* (Oxford University Press, Oxford, 2003), 2nd ed.
[4] J. P. Crutchfield and D. P. Feldman, Chaos **13**, 25 (2003).
[5] J. P. Crutchfield and K. Young, Phys. Rev. Lett. **63**, 105 (1989); J. P. Crutchfield, Physica (Amsterdam) **75D**, 11 (1994); J. P. Crutchfield and C. R. Shalizi, Phys. Rev. E **59**, 275 (1999).
[6] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948); **27**, 623 (1948).
[7] A. Fraser and H. L. Swinney, Phys. Rev. A **33**, 1134 (1986).
[8] *Nonlinear Modeling*, edited by M. Casdagli and S. Eubank (Addison-Wesley, Reading, MA, 1992).
[9] J. P. Crutchfield and D. P. Feldman, Phys. Rev. E **55**, R1239 (1997).
[10] I. Erb and N. Ay, J. Stat. Phys. **115**, 949 (2004).
[11] G. Tononi, O. Sporns, and G. M. Edelman, Proc. Natl. Acad. Sci. U.S.A. **91**, 5033 (1994).
[12] W. Bialek, I. Nemenman, and N. Tishby, Neural Comput. **13**, 2409 (2001).
[13] W. Ebeling and T. Poschel Europhys. Lett. **26**, 241 (1994).
[14] L. Debowski, arXiv:0810.3125.
[15] Compared to other information-theoretic measures of organization, **E** is the most general, being the "all-point" mutual information [4]. Of the alternative measures of stored information, C_μ is the most general since the ϵ machine, from which it is derived, is a minimal sufficient statistic for a process [5].
[16] J. P. Crutchfield, in Ref. [8], pp. 317–359.
[17] R. Yeung, IEEE Trans. Inf. Theory **37**, 466 (1991).
[18] As noted for human language [13,14], **E** and C_μ can diverge. Using methods developed to analyze infinite ϵ machines [4,5], we will report extensions of the present results to such cases elsewhere.