DIFFERENTIAL HEBBIAN LEARNING

Bart Kosko
VERAC, Inc., 9605 Scranton Road, San Diego, CA 92121-1771

ABSTRACT

The differential Hebbian law $\dot{e}_{ij} = \dot{C}_i \dot{C}_j$ is examined as an alternative to the traditional Hebbian law $\dot{e}_{ij} = C_i C_j$ for updating edge connection strengths in neural networks. The motivation is that concurrent change, rather than just concurrent activation, more accurately captures the "concomitant variation" that is central to inductively inferred functional relationships. The resulting networks are characterized by a kinetic, rather than potential, energy. Yet we prove that both system energies are given by the same entropy-like functional of connection matrices, Trace($\dot{E}$ E). We prove that the differential Hebbian is equivalent to stochastic-process correlation (a cross-covariance kernel). We exactly solve the differential Hebbian law, interpret the sequence of edges as a stochastic process, and report that the edge process is a submartingale: the edges are expected to increase with time. The submartingale edges decompose into a martingale or unchanging process and an increasing or novelty process. Hence conditioned averages of edge residuals are encoded in learning though the network only "experiences" the unconditioned edge residuals.

INTRODUCTION

Synaptic connections are causal connections. Their modification is an act of inductive inference. Edge connection strengths are inferred from node (neuron, processing element, etc.) behavior. This suggests that the modification criteria should, at minimum, reflect the logico-causal criteria of scientific method used for attributing a functional relationship among variable quantities. And what are these criteria but that the quantities should move or change in the same or opposite directions and that the "cause" temporally precedes the "effect?" Eighteenth century empiricist philosopher David Hume[1] observed that we habitually make causal ascriptions when we observe sustained "constant conjunctions of events." In his System of Logic, nineteenth century empiricist philosopher John Stuart Mill[2] refined Hume's observation. Mill observed that the causality we attempt to inductively infer is simply the "concomitant variation" of the variable quantities, and this formulation, often restated in the jargon of statistical inference, remains the operative notion of causality today.

THE DIFFERENTIAL HEBBIAN LAW

The task is to specify the dynamical equation $\dot{e}_{ij} = f(C, E)$ of the edge e_{ij} that connects node (causal variable) C_i to C_j in a network, where $C^T = (C_1, \ldots, C_n)$ is the node state vector and $E = [e_{ij}]$ is the matrix of edge connections. We assume the transfer equation $\dot{C}_i = g(C, E)$ is given for each node (and is dominated by an inner product of input edges and

0094-243X/86/1510277-6$3.00

nodes): $$\dot{C}_i = \sum_k C_k e_{ki} + D_i , \quad (1)$$

where D_i contains all other terms. This functional form, a sort of neural OR gate, predominates in neural modeling. We assume that the righthand side of (1) contains terms internal and external to the network. The simplest internal terms are the inner product minus the current activation C_i. The external terms are an observation or sensor term O_i and an advice or teacher or expert-response term R_i, both of arbitrary structure. Hence $D_i = O_i + R_i - C_i$.

The standard selection of f in $\dot{e}_{ij} = f(C, E)$ is attributed to the "correlation learning" hypothesis of Hebb[3], which is simply that concurrent activation of nodes increases the "synaptic efficacy" or strength of the connection between them. In a nutshell,

$$\dot{e}_{ij} = C_i C_j . \quad (2)$$

Typically the current strength e_{ij} is subtracted from the righthand side to represent "forgetting" or "memory decay" (or to slow the otherwise exponential growth?). This linear appendage does not affect the current analysis and for notational simplicity we omit it.

Equation (2) is widespread in the neural net literature. It occurs in the famous Grossberg[4-6] equations and the related equations of Hopfield[7] and in similar form in the adaptive equations of most neural modelers. But apart from referencing Hebb's (nonmathematical) conjecture, it seems the operative argument for using (2) is simply that everyone uses it. For surely the problems with (2) warrant investigating alternatives. To begin, (2) promotes spurious causal associations. If any two processors or nodes are active in a network, no matter how big the network, how far apart the nodes, or how independent their patterns of activation, the Hebbian law (2) grows a causal connection between them. Concomitant activation replaces concomitant variation. Worse, the spurious causal attributions tend to grow exponentially fast (as can be seen from the exponential form of the exact solution of (2) when the forget term $-e_{ij}$ is appended). In practice this necessitates "hardclipping" of interconnects both during training and classification sessions. Finally, transfer functions must first be integrated before (analytically) including them in (2). This integration is never easy.

A natural alternative to (2) is the differential Hebbian law:

$$\dot{e}_{ij} = \dot{C}_i \dot{C}_j . \quad (3)$$

The differential Hebbian measures concomitant variation. It imputes causality according to (lagged) conjunctions of event changes. As a result, it truly behaves in correlation fashion. For although the functions C_i are nonnegative, their derivatives are not. Hence the connection e_{ij} strengthens iff both nodes agree in sign, hence iff both nodes move in the same direction. (Note this implies concurrent activation.) Negative causality accumulates if they move in opposite directions. Moreover, transfer functions such as (1) can be directly plugged

into (3), allowing many properties to be determined analytically. Below we exploit this fact to solve the system (1) and (3) for $e_{ij}(t)$.

The idea of using rates of change in learning laws is spreading. Two especially noteworthy cases are the drive-reinforcement model of Klopf[8-9] and the backward-error-propagation model of Rumelhart, Hinton, and Williams. Klopf reports that a wide array of Pavlov-like learning behaviors is accurately predicted (retrodicted) by a change-based law. Rumelhart uses a time derivative of input activation (essentially $\dot{C}_j$) in his "generalized delta law" (subsuming the classical perceptron convergence theorem) to solve the exclusive-or problem, the parity problem, and a variety of others.

KINETIC ENERGY CONNECTIONS

The energy of a network is the sum of the eigenvalues of the product connection matrix $E\dot{E}$. Here $\dot{E}$ is the n-by-n symmetric matrix of connection changes $[\dot{e}_{ij}]$. For Hopfield networks, Abu-Mostafa and St. Jaques[11] have shown that the number of energy minima (memory sites) is no more than n. We conjecture that these and comparable equilibria correspond to the eigenvalues of $E\dot{E}$.

To prove the eigenvalue theorem, define the potential energy (P.E.) and kinetic energy (K.E.) of the network [C, C, E, E] as follows:

$$\text{P.E.} = C^T E C = \sum_i \sum_j C_i C_j e_{ij} , \tag{4}$$

$$\text{K.E.} = \dot{C}^T E \dot{C} = \sum_i \sum_j \dot{C}_i \dot{C}_j e_{ij} . \tag{5}$$

The trick is that if the Hebbian law (2) is in force, then $\dot{e}_{ij}$ replaces $C_i C_j$ in (4). If the differential Hebbian law (3) is in force, then $\dot{e}_{ij}$ replaces $\dot{C}_i \dot{C}_j$ in (5). Hence then P.E. = K.E. !

THEOREM.
$$\sum_{i=1}^{n} \lambda_i = \begin{cases} \text{P.E.} & \text{if } \dot{E} = C C^T \\ \text{K.E.} & \text{if } \dot{E} = \dot{C} \dot{C}^T \end{cases} , \tag{6}$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $E\dot{E}$ $(\dot{E}E)$.

PROOF. By basic linear algebra, the sum of eigenvalues of $E\dot{E}$ equals the sum of diagonal elements, the trace Trace($E\dot{E}$). Then

$$\text{Trace}(E\dot{E}) = \sum_i (E\dot{E})_{ii} = \sum_i \sum_j e_{ij} \dot{e}_{ji} = \sum_i \sum_j e_{ij} \dot{e}_{ij}$$

equals P.E. or K.E. according as $\dot{E}$ equals $C C^T$ or $\dot{C} \dot{C}^T$. Q.E.D.

Two comments are in order. First, when a no weight-change analysis is desired--as it often is in Grossberg and Hopfield networks--simply invoke the Hebbian or differential Hebbian law to replace $\dot{e}_{ij}$ with $C_i C_j$ or $\dot{C}_i \dot{C}_j$. Second, Trace($E\dot{E}$) can be viewed as an entropy-like functional. For, in

learning networks, the edges or neural pathways adapt slowly to the activation or signals flowing through them. I.e., $\dot{e}_{ij}$ tends to be smaller than e_{ij} at any time t. Let us model this property with the hypothesis that $\dot{e}_{ij} \cong \ln e_{ij}$. Now recall that Von Neumann[12] defined the entropy of a (quantum-mechanical) system with Trace(P lnP), where P is a positive semidefinite matrix with Trace(P) = 1, thus generalizing the log-of-probability entropy of Boltzmann and others (including Shannon). Watanabe[13] has kept this entropy measure current by showing how its minimization corresponds to pattern recognition in many cases. On our learning hypothesis, $\text{Trace}(\dot{E}E) \cong \text{Trace}(P \ln P)$ for suitable P. However, since Trace(E) = Trace(Ė) = 0, E cannot be normalized to 1 and thus an exact identity between the two trace functionals cannot be expected to hold.

CORRELATION AND AN EXACT SOLUTION

Let us interpret the node vector C as a random vector. Let each node C_i be a stochastic process: $C_i : T \times \Omega \rightarrow [0, \infty)$, where T is an index set of time values and (Ω, A, P) is the probability space. Suppose we know the present value of $C_i(t)$ $(C_i(t, \omega))$. Then given no other information, and being true scientific empiricists, what is our best prediction of $C_i(t + 1)$? Surely it is just the present value $C_i(t)$. Let us call this the quasi-martingale assumption:

$$E_P(\, C_i(t + 1)\,) \quad = \quad C_i(t) \quad , \tag{7}$$

where E_P is the expectation with respect to probability measure P. If we now use the discrete differential Hebbian, we arrive at

$$\begin{aligned} e_{ij}(t + 1) &= e_{ij}(t) + \Delta C_i(t)\ \Delta C_j(t + 1) \qquad (8) \\ &= \sum_{s=0}^{t} \Delta C_i(s)\ \Delta C_j(s + 1) \\ &= \sum_{s=0}^{t} [C_i(s) - E(C_i(s))][C_j(s+1) - E(C_j(s+1))], \end{aligned}$$

which has the form of a nonnormalized cross-covariance kernel--the key term in the definition of statistical correlation. This identity establishes a direct connection with correlation and our operative definition of causality, the differential Hebbian (3).

Since the differential Hebbian uses derivatives of transfer functions, we can plug in dynamical equations such as (1) and solve directly for e_{ij}. We demonstrate that the network [C, C, E, Ė] given by (1) and (3) can be easily solved. The trick is that e_{ij} can be pulled out of the inner product in (1). (3) then takes the form $\dot{e}_{ij} + a\, e_{ij} = b$ --a first-order inhomogeneous ordinary differential equation with variable coefficients, one of the most tractable differential equations. Details of this expansion can be found in Kosko[14-15]. Here we state the exact solution:

$$e_{ij}(t) = e^{\int_0^t p(s)\,ds}\left(K + \int_0^t q(s)\, e^{-\int_0^s p(u)\,du}\,ds\right), \quad (9)$$

where the functions p and q contain the grouped terms that occur in the manipulation when C_i and C_j are expanded with (1), which for brevity we omit. The essential point is that p only contains the observation term O_i and contains it linearly and q contains both O_i and O_j and the product $O_i O_j$. The constant K is given by $K = e_{ij}(0)$.

EDGES AS SUBMARTINGALES

The edge e_{ij} is expected to increase in strength in time as information accumulates under the differential Hebbian hypothesis (3). This answers the question where edges tend on average, or, put another way, how connected networks become on average. To formalize such speculations, we interpret e_{ij} as stochastic process--$e_{ij}: T \times \Omega \to (-\infty, \infty)$--on the probability space (Ω, A, P). We superscript with t to index the random variables $\{e_{ij}^t\}$.

External input enters the network through the observation terms O_i in (1). A hearty scientific empiricism dictates that, fundamentaly, we know nothing of the future flux of experience. Experimentation is the resultant coping device. What does this mean for future values of the random variable O_i^t when we know the present value O_i^s, $s < t$? More precisely, suppose we have an increasing sequence of sets of information in the form of a "filtration" of sigma-algebras $A_s \subset A_t \subset A$, $s \leq t$. Then what is the conditional expectation $E(O_i^t \mid A_s)$? Surely the most we can assert is the present observed value. This constitutes a network martingale assumption:

$$E(O_i^t \mid A_s) = O_i^s \quad \text{if } s \leq t . \quad (10)$$

Since the product $O_i O_j$ occurs in the q term of solution (9), we must further assume that the two observation processes are conditionally independent martingale processes: $E(O_i^t O_j^t \mid A_s) = O_i^s O_j^s$ for $s \leq t$. (Technically we also assume the processes are suitably integrable and filtration measurable.)

So is the edge process a martingale if the observation processes are conditionally independent martingales? It turns out that the edge process is a submartingale process. (The proof, in Kosko[14], uses Jensen's inequality and the convexity of the exponentials in the modified (9).) The expected future value is at least as big as the present value when conditioned on all the information available up to the present time. Hence $E(e_{ij}^t \mid A_s) \geq e_{ij}^s$ for $s \leq t$. The proof is facilitated by interpreting the integrals in solution (9) as conditional expectations (since stochastic integrals are always martingales). For instance, the first integral then takes the form $E(p_t \mid A_t)$. This interpretation keeps account of the acquired information and outputs functions not constants.

Hence neural or causal networks can be expected to become more connected on average as time passes. The causal consequences of this may seem anti-entropic, perhaps even a violation of the second law of thermodynamics. But in fact total connectivity is the maximum entropy case (where entropy is intuitively interpreted). Order is established by a contrast between edge connections and edge disconnections, as in a military hierarchy or a bureaucratic dictatorship rather than in a voting democracy. Better, in terms of dynamic communication connections, note how a monitored debate tends unidirectionally to a free-for-all discussion.

Another consequence is that edge processes can be decomposed into a sum of a martingale and an increasing process:

$$e_{ij}^t = M_{ij}^t + N_{ij}^t, \tag{11}$$

where M is the martingale process and N is either positive or zero. M represents what stays the same in the connection process. Hence N represents what is new or novel, and, in the spirit of Kohonen[16], we call it the novelty process. A consequence of the Doob-Meyer decomposition (11) is that N, and hence its residual, can be written as follows:

$$N_{ij}^t = \sum_{k=1}^{t} E(e_{ij}^k - e_{ij}^{k-1} \mid A_{k-1}), \tag{12}$$

$$N_{ij}^t - N_{ij}^{t-1} = E(e_{ij}^t - e_{ij}^{t-1} \mid A_{t-1}) . \tag{13}$$

Of the many things that can be said of this novelty process, perhaps the most significant concerns what it, and hence the edge process, encodes. (13) makes the contribution to differential Hebbian learning. What is significant is that a conditioned average is encoded even though not experienced. The expected edge residual is conditioned on exactly what it should be--all the information accumulated up to the present.

REFERENCES

1. D. Hume, An Inquiry Concerning Human Understanding (1748).
2. J. S. Mill, A System of Logic (1843).
3. D. O. Hebb, The Organization of Behavior (1949).
4. S. Grossberg, Psych. Rev., 58, 1 (1980).
5. S. Grossberg, Studies of Mind and Brain (Reidel, Boston, 1982).
6. M. A. Cohen and S. Grossberg, IEEE Trans. SMC, 13, 815 (1983).
7. J. J. Hopfield, Proc. Nat. Acad. Sci, 79, 2554 (1982).
8. A. H. Klopf, Proc. 2nd Conf on Comp. with Neural Net. (1986).
9. R. S. Sutton and A. G. Barto, Psych. Rev., 88, 135 (1981).
10. D. E. Rumelhart, G. E. Hinton, R. J. Williams, in Parallel Distributed Processing, Rumelhart and McClelland Eds. (MIT Press, 1986).
11. Y. S. Abu-Mostafa and J. St. Jacques, IEEE Trans. IT, 31, 461 (1985).
12. J. Von Neumann, Mathematical Foundations of Quantum Mechanics (Springer, Berlin, 1932).
13. S. Watanabe, Pattern Recognition: Human and Mechanical (Wiley, NY, 1985).
14. B. Kosko, "Adaptive Inference," subitted for publication (1986).
15. B. Kosko and J. S. Limm, Proc. SPIE, 579, 104 (1985).
16. T. Kohonen, Self-Organization and Associative Memory (Springer-Verlag, Berlin, 1984).