Critical Branching Neural Computation, Neural Avalanches, and 1/f Scaling

Christopher T. Kello (ckello@ucmerced.edu)
Bryan Kerster (bkerster@ucmerced.edu)
Eric Johnson (ejohnson5@ucmerced.edu)
Cognitive and Information Sciences, 5200 North Lake Rd., Merced, CA 95343 USA

Abstract

It is now well-established that intrinsic fluctuations in human behavior tend to exhibit long-range correlations in the form of 1/f scaling. Their meaning is an ongoing matter of debate, and some researchers argue they reflect the tendency for neural and bodily systems to poise themselves near critical states. A spiking neural network model is presented that self-tunes to a critical point in terms of its spike branching ratio (i.e. critical branching). The model is shown to exhibit 1/f scaling near critical branching, as well neural avalanches, and critical branching is associated with maximal computational capacity when assessed in terms of reservoir computing. The model provides a basis for connecting neural and behavioral activity and function via criticality.

Keywords: Critical branching, 1/f scaling, neural avalanche, criticality, metastability, reservoir computing.

Introduction

Variability is the essence of neural and behavioral activity, and this variability is what theories of cognition must ultimately account for. Some of this variability can be ascribed to effects of sensory stimulation, but much of it is intrinsic in nature (Fox, Snyder, Vincent, & Raichle, 2007). Like all biological systems, neural and behavioral systems exhibit activities that can neither be attributed to extrinsic factors, nor controlled by them. These systems are constantly at work to maintain themselves, and this work results in intrinsic variations in activities. The nature of intrinsic variability provides basic information about how components of these systems work together.

Intrinsic variability is observed when experimental manipulations are minimized, e.g. when spontaneous neural activity is measured in cortical slice preparations (Beggs & Plenz, 2003), or in brain images during the wakeful resting state (Bullmore et al., 2001), or when behavioral acts are repeated with minimal variation in intentions and measurement conditions (Kello, Anderson, Holden, & Van Orden, 2008). What should one expect from system activity when components are in this "relaxed", default state?

A reasonable hypothesis is that component activities (e.g. neurons, cortical columns, brain areas, muscle groups, etc.) decouple to become relatively independent, and effectively random. If fluctuations in system activities reflect component sums in intrinsic measurement conditions, then activities should tend towards "white noise", i.e. random samples drawn from a normal distribution. In fact this is the basic assumption of linear models with Gaussian error terms. However, numerous studies of intrinsic variability do not bear out this assumption.

Scaling Laws in Neural and Behavioral Activity

In many different studies of neural and behavioral activity, intrinsic variations have been reported to follow scaling laws across a wide range of scales. Scaling laws generally relate one variable as function of another raised to a power, $f(X) \sim X^a$, where typically $a < 0$. Well-known examples from psychology and cognitive science include Steven's law, Zipf's law, scale-free semantic networks, and power laws of learning and forgetting (for review see Kello et al., 2010).

Here we focus on two different scaling laws that have attracted a great deal of attention in recent years. One is a power law distribution in neural activity referred to as a "neural avalanche" (Beggs & Plenz, 2003), and the other is long-range correlated fluctuations in behavioral and neural activity, known as 1/f scaling (Kello et al., 2008).

The term "neural avalanche" originally referred to bursts of neural spiking activity found in local field potentials recorded from slice preparations that are designed for observing intrinsic variations. Probability distributions of burst sizes S were found to go as $P(S) \sim 1/S^{\beta}$, where $\beta \sim 3/2$ over a moderate range of scales. Analogous burst size distributions have also been found in EEG, MEG, and fMRI recordings (see Poil, van Ooyen, & Linkenkaer-Hansen, 2008).

1/f scaling refers to autocorrelations in time series of repeated measurements, in our case taken from neural or behavioral systems. Each measurement is correlated with previous ones, and correlations decay slowly as an inverse power of lag between measurements. In the frequency domain, this scaling relation holds between spectral power S and frequency f as $S(f) \sim 1/f^{\alpha}$, where $\alpha \sim 1$ over a moderate to wide range of scales. This scaling law has been observed in local field potentials, EEG, fMRI, and a wide variety of behavioral measures of intrinsic variation, including tapping, walking, reaction times, interval estimates, and the acoustics of spoken word repetitions (see Kello et al., 2008).

Criticality and Computation

What do neural avalanches and 1/f scaling tell us about neural and behavioral systems? One possibility is suggested by the particular exponent values observed, because they are both predicted to occur in the intrinsic variations of systems near their *critical points*. Critical points occur at the transitions between phases (i.e. modes) of component interactions, and many different kinds of complex systems have been hypothesized to self-organize towards their critical points (Bak, 1996). Theoretical work has shown that

systems near their critical points universally exhibit scaling laws in their intrinsic dynamics (Sornette, 2004). 1/f scaling with $\alpha \sim 1$ has been shown to hold for a wide range of model systems poised near transitions between ordered versus disordered states, whereas neural avalanches with $\beta \sim 3/2$ hold for systems poised near transitions between diminishing versus expanding branching processes.

While the predicted scaling exponents lend credence to the idea that neural and behavioral systems tend to be poised near critical points, one is led to ask, why would this be so? One possible reason is that both kinds of phase transitions have been associated with adaptive cognitive properties (Kello et al., 2010). Here we focus on the maximization of information transmission and memory capacity in critical branching networks (more on this in the Conclusion).

Any given spiking neural network can be viewed as a branching process whereby a given spike occurring at time *t* may subsequently "branch" into some number of spikes at time $t + \Delta t$ over the neurons connected via its axonal synapses. Let us call the former an "ancestor" presynaptic spike, and the latter are "descendant" postsynaptic spikes. The expected number of descendants for each given ancestor is the *branching ratio* of a spiking network, $\sigma = \mathrm{E}(N_{post} / N_{pre})$, where E() is expected value.

If σ is less than one, then spikes diminish over time, and information transmission through the network is inhibited in terms of dampened propagation of spiking activity. If σ is greater than one, then spikes grow over time and eventually come to saturate the network, which also inhibits information transmission. $\sigma = 1$ is the critical branching point at which spikes are conserved over time, and so propagate without dying out or running rampant.

An analogous critical point between convergent and divergent dynamics (i.e. Lyapunov exponents near one) has been shown to maximize memory capacity in a recurrent network of threshold gating units known as a *liquid state machine* (Bertschinger & Natschlager, 2004). Weights on connections between units were set to be near the critical point, and intrinsic gate dynamics (switches between 1 and -1) were perturbed by external inputs. There were two arbitrary input patterns (i.e. one "bit" of information), and one of the patterns was chosen randomly for input on each time step.

Past inputs may have effects on gate dynamics that carry forward in time. The *memory capacity* of the network was defined by two factors: The distance in time over which information about past inputs was carried forward in current gate values, and the degree to which different patterns of past inputs were distinguishable in current gate values. Memory capacity was assessed by using a linear regression "readout" function to classify patterns of gate values over units according to nonlinear functions of past input bits (i.e. XOR and parity). Linear readout can only succeed if the effects of past inputs carry forward to current gate values, and only if gate dynamics take nonlinearly separable inputs and make them linearly separable. Results showed that memory capacity was maximal when weights were set near the critical point between convergent and divergent dynamics.

Critical Branching Model

The studies reviewed thus far leave us with two gaps: 1) a single model has not been shown to exhibit both neural avalanches and 1/f scaling, and 2) a biologically plausible neural network algorithm has not been formulated to drive synapses towards a critical branching point. Kello and Mayberry (2010) made progress towards filling these gaps by presenting a spiking neural network model with a critical branching self-tuning algorithm. The model exhibited neural avalanches and maximal memory capacity near its critical point, but 1/f scaling was not demonstrated, and the model was not biologically realistic.

Here we present a more realistic, spiking neural network model that self-tunes to a critical branching point, and in doing so exhibits both neural avalanches and 1/f scaling. Moreover, deviations from both scaling laws are exhibited as the model moves either toward subcritical or supercritical phases, and memory capacity is maximized near the critical branching point. Memory capacity is measured by applying *reservoir computing* functions to spike dynamics, as done in liquid state machines. Our work provides a basis for spiking neural network models that connect neural and cognitive functions via the principle of criticality.

Model Variables and Update Equations. A basic kind of model spiking neuron is the leaky integrate-and-fire (LIF) unit. LIF units generally have the following variables (Roman letters) and parameters (Greek letters): A membrane potential V_i for each neuron *i*, a membrane threshold θ_i and membrane leak λ_i, and a level of potentiation w_j for each axonal synapse *j*, where $w_j >= 0$ for excitatory neurons and $w_j <= 0$ for inhibitory neurons. Models may also include variable synaptic delays τ_j, as well as parameters governing the time course of action potentials and postsynaptic potentials (e.g. membrane resistance).

Our model included all of the above, except that action potentials and postsynaptic potentials were instantaneous for the sake of simplicity. The model was biologically realistic in that 1) variable updates were local in time and local with respect to immediately connected synapses and neurons (numerical values were not transmitted over connections among neurons, as they are in e.g. backpropagation), and 2) synaptic and neuronal updates were asynchronous and event-based (i.e. time was not discretized, it was coded with arbitrary precision). The latter criterion helped ensure the plausibility of our critical branching tuning algorithm.

Each update event in the model begins when a given neuron receives as input a postsynaptic potential I_j at time *t*, which may either come from another neuron within the model, or to from an external input source (i.e. neurons outside the model or sensory stimulation):

$$V_i \leftarrow V_i e^{-\lambda_i (t-t')} + I_j, \qquad [1]$$

where $\leftarrow$ denotes the instantaneous update of a variable, and t' is the previous time that V_i was updated. Thus the

model included continuous exponential leak, applied each time a given neuron received an input. Immediately after each V_i update, if $V_i > \theta_i$, then $V_i \leftarrow 0$, and a postsynaptic potential I_j was generated for each axonal synapse of i. Each $I_j = w_j$, and was applied at time $t + \tau_j$.

In a typical connectionist model, w_j can be any real-valued number, possibly bounded by some minima and maxima. However, recent neurophysiological evidence suggests that synapses may be similar to noisy binary switches with only two levels of potentiation (e.g. O'Connor, Wittenberg, & Wang, 2005), and it has been argued that this limitation has little effect on the computational capacity of synapses (Baldassi, Braunstein, Brunel, & Zecchina, 2007). Therefore we used discrete-valued synapses in order to limit the number of activated synapses ($w_j \neq 0$), and to enable a stochastic tuning algorithm. In particular, we used synapses with only two possible levels of potentiation, 0 or φ_j.

Each LIF model neuron has two free parameters, λ_i and θ_i, and each synapse has two free parameters, τ_j and φ_j. In some spiking network models, parameters are set according to empirical data on particular kinds of neurons (e.g. pyramidal cells; ref). However, perhaps the most basic and overarching fact about neurons is their heterogeneity: Parameters vary across different kinds of neurons, and across different neurons of a given kind. To reflect the general fact of heterogeneity, values for all four free parameters were sampled randomly from uniform distributions whose ranges were set to reasonable default values. In particular, values were real numbers in the ranges $1 < \theta_i < 2$, $0.5 < \lambda_i < 1$, $1 < \tau_j < 1.5$, $1 < \varphi_j < 2$ for excitatory units, and $-1 < \varphi_j < -0.1$ for inhibitory units.

The set of membrane potentials V and postsynaptic potentials I comprise the dynamics of neurons in our LIF model. These variables are governed by event-based updates (Eq 1, plus threshold dynamics) that may occur asynchronously across neurons, at any point in continuous time (simulated with arbitrary precision, no need to choose a time discretization). The set of synaptic weights w comprise the dynamics of synapses, and are governed by the critical branching algorithm described next.

Self-Tuning Algorithm. The objective of the self-tuning algorithm is to activate and de-activate synapses so that each ancestor spike is followed by one descendant spike on average. A local estimate for σ is computed over the interspike interval (ISI) for each model neuron i. This means that only $N_{post,i}$ need be estimated, because $N_{pre,i}=1$ by definition with respect to a given neuron's ISI. Thus, to achieve critical branching, $N_{post,i}$ should sum to one.

When a given neuron spikes, its local estimate of σ is reset, $N_{post,i} \leftarrow 0$. For each axonal synapse's *first* spike occurring at time t, $N_{post,i}$ was incremented by $s_i = e^{-\lambda_i (t-t')}$. For each increment, each descendant spike was weighted as a decaying function of the time interval between pre- and postsynaptic spikes, with maximal weighting when the former was immediately followed by the latter.

The sum of time-weighted descendants is used (before it is reset to zero) each time the neuron spikes to update weights on its axonal synapses. In particular, if $N_{post,I} < 1$, then perform the update $w_j \leftarrow \varphi_j$ for each synapse j with probability

$$\eta f(s_i) \left| N_{post,i} - 1 \right| / U , \qquad [2]$$

where η is a global tuning rate parameter (fixed at 0.1), and U is the number of synapses available for potentiation. $f(s_i) = 1 - e^{-\lambda_i (t-t')}$ if neuron i was excitatory, and $f(s_i) = e^{-\lambda_i (t-t')}$ if inhibitory. If $N_{post,I} > 1$, then perform the update $w_j \leftarrow 0$ with probability set according to Eq 3, except U is the number of synapses available for de-potentiation, and the assignment of $f(s_i)$ is switched for excitatory versus inhibitory neurons.

In essence, the critical branching algorithm activates synapses when too few descendant spikes occur, and de-activates when too many occur. Spikes are time-weighted because effects of ancestor spikes on descendant neurons diminish according to their leak rates. Critical branching weight updates increase in likelihood as local branching ratio estimates diverge from one, and depend on spike timing. With regard to spike timing, excitatory synapses are more likely to be potentiated when postsynaptic neurons have not fired recently (and vice versa), which helped to spread spikes across neurons. The same principle leads to the opposite rule for inhibitory neurons.

Model Architecture. The model consisted of 200 inputs units and 1000 reservoir units. All input units were excitatory, and reservoir units were excitatory or inhibitory with probability 0.5. Input units were connected to each reservoir unit with probability 0.2, and reservoir units were connected to each other with probability 0.2. All synaptic weights were initialized to zero.

Simulation 1: Scaling Laws

We first examine intrinsic variations exhibited by the model under two different random noise input conditions. In the *high input* condition, exactly one half of the input units were induced to spike at a random time within the first half of each unit time interval. In the *low input* condition, only five input units were induced to spike per unit time interval. High input caused a steady fluctuation in spikes, whereas low input caused "bursts" of activation above baseline.

In Figure 1, two time series are shown for the first 8000 time intervals of an example run in the high input condition. The top series is the mean branching ratio estimate per unit interval, and the bottom series is the number of reservoir units that spiked per unit time interval. The top series shows the ability of the critical branching self-tuning algorithm to reach and maintain $\sigma \sim 1$, starting from zero potentiated synapses. The bottom series shows variations around a mean of ~210 spikes per time interval. These variations are largely intrinsic to the model, because there was no variation in the number of input spikes per time interval.

In Figure 2, a spectral analysis is shown for the last 4096 data points in Figure 1, in log-log coordinates. Fluctuations in numbers of spikes are shown to closely follow a 1/f scaling relation in the lower frequencies (ideal 1/f shown by the dashed line), which represents the vast majority of variation in the time series because power is on a logarithmic scale. The small amount of remaining variation (despite appearances in the figure) in the higher frequencies is uncorrelated noise (slope near zero). This general pattern is common to nearly all empirical observations of 1/f scaling, including those in behavioral and neural activity. At least some of this variation comes from the temporal dispersion of inputs within each time interval, and variations in neuron and synapse parameters.

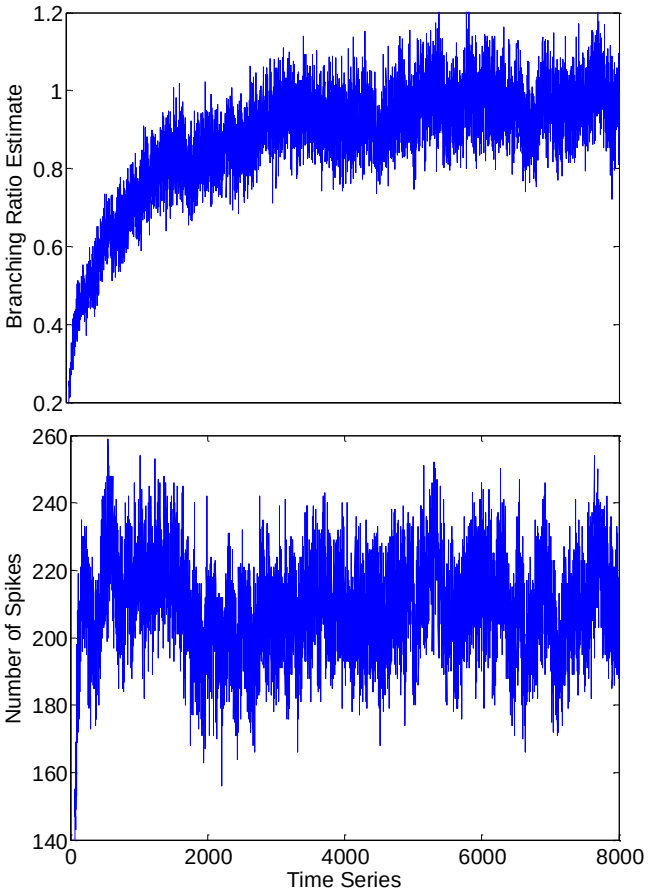

Figure 1: Branching Ratio and Spike Variations in the High Input Condition

In Figure 3, a portion of the spike series from the low input condition is shown. Only a small stretch is shown in order to highlight the "burst-like" nature of the time series. As in the high input condition, these bursts are intrinsic variations because the rate of inputs spikes was held constant at 5 per time interval. The size of each avalanche (i.e. burst) was defined as the sum of spikes over contiguous points with 10 or more spikes (red arrows show three example avalanches).

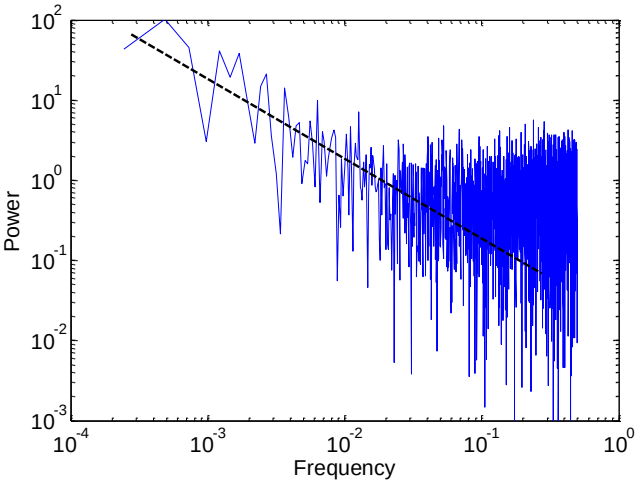

Figure 2: Spectral Analysis of Spike Time Series in Fig 1

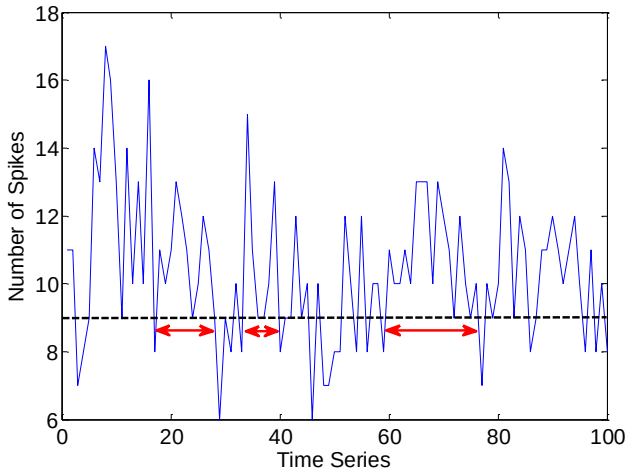

Figure 3: Spike Variations and Avalanche Time Series in the Low Input Condition

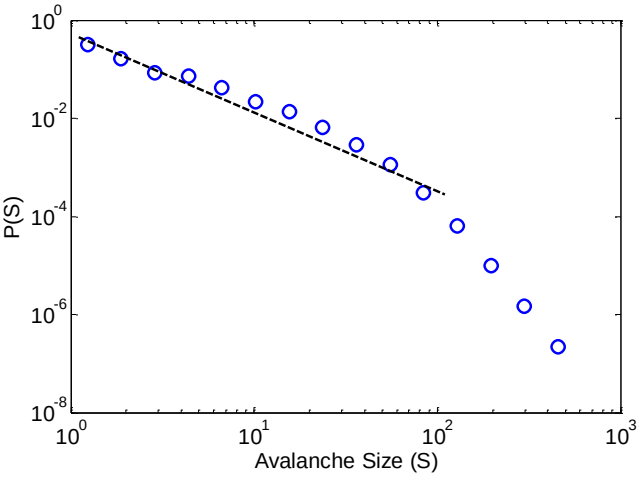

Figure 4: Avalanche Probability Distribution (log bins)

In Figure 4, a histogram of avalanche sizes is plotted in log-log coordinates for a run of 200,000 time steps in the low input condition (after first tuning to critical branching). The ideal 3/2 power law is shown by the dashed line. The

fall off in larger avalanche sizes is also characteristic of real neural avalanche data, and is likely due to the limited size of the model (a similar fall off in local field potential data is apparently due to limited numbers of electrodes).

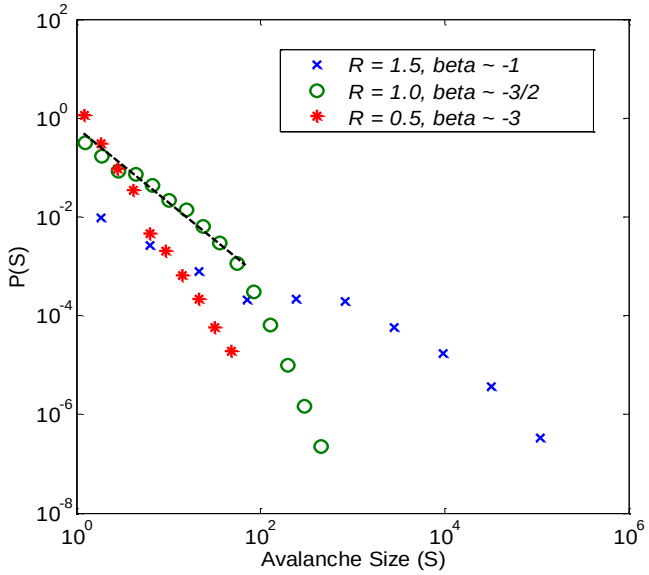

Figure 5: Aggregate Avalanche Histograms for Target Branching Ratios of 0.5, 1.0, and 1.5

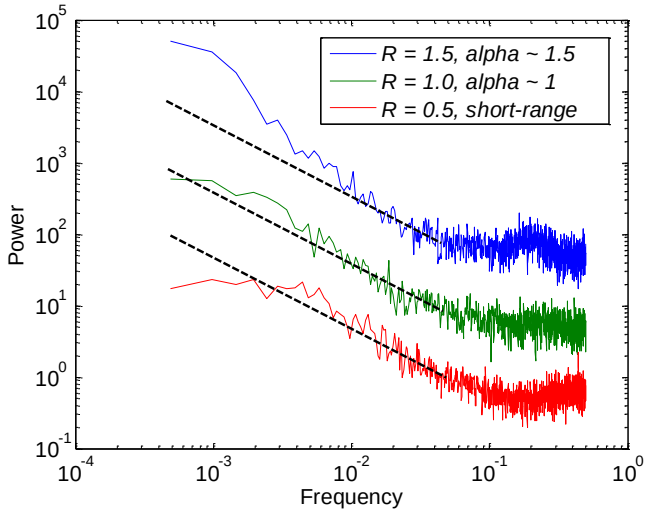

Figure 6: Means Spectral Plots for Target Branching Ratios of 0.5, 1.0, and 1.5

Results thus far show that intrinsic variations in critical branching spiking activity exhibit near ideal neural avalanches and 1/f scaling under low and high input conditions, respectively. These results alone do not distinguish whether the scaling laws are associated with critical branching, or something more general about how the algorithm works.

To test whether critical branching is important for simulating these scaling laws, the critical branching algorithm was generalized to target branching ratios other than one. In particular, the tuning algorithm was generalized to target a given branching ratio R by replacing the $\left|N_{post,i}-1\right|$ term with $\left|N_{post,i}-R\right|/R$ in Eq 3. In Figures 5 and 6, results are shown at three different targeted branching ratios, i.e. σ = 0.5, σ = 1.0, and σ = 1.5. Spike burst size distributions are shown to diverge from ideal neural avalanches when the branching ratio diverges from one, and summed spike fluctuations are shown to diverge from 1/f scaling. For avalanches, the power law tail of the distribution either becomes too light (subcritical, R = 0.5) or too heavy (supercritical, R = 1.5). For spectra, either fluctuations lose their long-range correlations as seen in a flattening of the spectrum in the lower frequencies (R = 0.5), or fluctuations deviate towards Brownian motion (R = 1.5).

Simulation 2: Memory Capacity

The memory capacity of spiking dynamics was assessed as a liquid state machine (Maass, Natschlager, & Markram, 2002). The only change to the critical branching model was in the inputs. Half of the input units were assigned to represent one bit value (0), and the other half were assigned to the other (1). For each time interval, one of the bit values was chosen at random, and all of its corresponding input units were induced to spike. The resulting sequence of bit inputs caused reservoir units to spike, and the critical branching tuning algorithm was engaged as in Simulation 1.

Once tuning asymptoted, a group of 15 "readout" units was used to test the XOR function on adjacent input bits that occurred from 1 to 15 time intervals in the past. Readout units used logistic outputs (instead of spiking), and were only for assessing spike dynamics (not part of critical branching). Weights on connections into readout units were real-valued and initialized in the range [-0.1, 0.1].

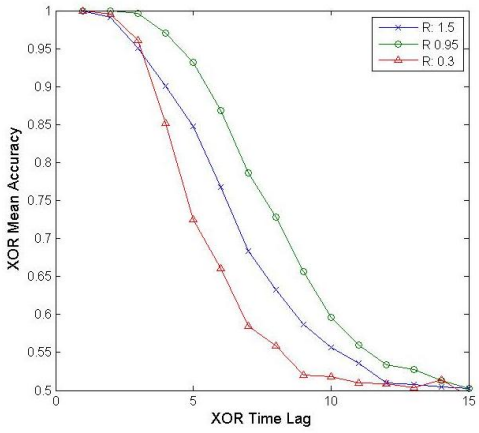

Figure 7. XOR classification accuracy as a function of time lag for three different targeted branching ratios

For each unit time interval, reservoir spikes resulted in output activations over readout units. For 10000 trials of training, readout units received XOR targets based on past input bits, and targets were compared with outputs using sum squared error. The resulting error signal was used to update connection weights using the delta learning rule (momentum = 0.5, learning rate = 0.00005). At testing, net

inputs to each readout unit had to be on the correct side of 0 to be considered correct. It is important to note that net inputs were a *linear* function of their weights, which meant that XOR performance relied on the memory and representational capacity of reservoir spiking dynamics.

The model was tested at 11 different branching ratios from 0.3 to 1.5 (run 10 times each per ratio). In Figure 7, XOR performance is shown to be greatest for the most recent time lags, and falls off to chance (0.5) by lag 15 (for replication, see Bertschinger & Natschlager, 2004). Performance was best when the targeted branching ratio was 0.95; ideal critical branching is at 1, and mean performance was slightly less for this target ratio (see Figure 8). This slight shift in the predicted peak at 1 is due to the use of a model with no output units where spikes could "exit" the system (see Kello & Mayberry, 2010).

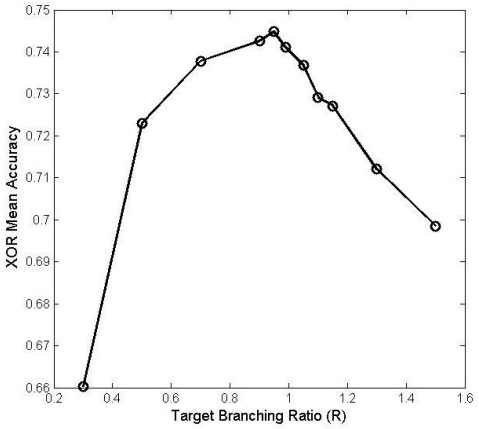

Figure 8. XOR classification accuracy, averaged over time delays, and plotted as function of targeted branching ratio

Conclusion

An LIF spiking network model was tuned to its critical branching point, which yielded 1/f scaling and neural avalanches, as well as maximal memory capacity. The model's basis in criticality provided the connection between these two heretofore unconnected scaling laws, and between the scaling laws and functional, cognitive properties of neural networks. The ability to address both neural and behavioral phenomena was facilitated by modeling at the level of spikes. In future simulations, we will further leverage the model by examining temporal autocorrelations and mutual dependencies among spike trains, interspike interval and spike rate distributions, and pervasive 1/f scaling in intrinsic fluctuations of behavioral activity.

If the model continues to account for basic facts about intrinsic variations in neural and behavioral activity, and grows to integrate the readout function, then it may provide a theoretical framework with broad empirical support that enables spiking dynamics models of cognitive function.

Acknowledgments

This work was funded by a DARPA sub-contract on an award to IBM.

References

Bak, P. (1996). *How nature works*. New York: Springer-Verlag.

Baldassi, C., Braunstein, A., Brunel, N., & Zecchina, R. (2007). Efficient supervised learning in networks with binary synapses (Vol. 104, pp. 11079-11084).

Beggs, J. M., & Plenz, D. (2003). Neuronal avalanches in neocortical circuits. *The Journal of Neuroscience, 23*, 11167-11177.

Bertschinger, N., & Natschlager, T. (2004). Real-time computation at the edge of chaos in recurrent neural networks. *Neural Computation, 16*(7), 1413-1436.

Bullmore, E., Long, C., Suckling, J., Fadili, J., Calvert, G., Zelaya, F., et al. (2001). Colored noise and computational inference in neurophysiological (fMRI) time series analysis: Resampling methods in time and wavelet domains. *Human Brain Mapping, 12*(2), 61-78.

Fox, M. D., Snyder, A. Z., Vincent, J. L., & Raichle, M. E. (2007). Intrinsic Fluctuations within Cortical Systems Account for Intertrial Variability in Human Behavior. *Neuron, 56*(1), 171-184.

Kello, C. T., Anderson, G., Holden, J. G., & Van Orden, G. C. (2008). The pervasiveness of 1/f scaling in speech reflects the metastable basis of cognition. *Cognitive Science, 32*, 1217-1231.

Kello, C. T., Brown, G. D. A., Ferrer-i-Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., et al. (2010). Scaling laws in cognitive sciences. *Trends in Cognitive Sciences, 14*(5), 223-232.

Kello, C. T., & Mayberry, M. R. (2010). Critical branching neural computation. In *International Joint Conference on Neural Networks* (pp. 1475-1481). Barcelona, Spain: IEEE.

Maass, W., Natschlager, T., & Markram, H. (2002). Real-Time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations. *Neural Computation, 14*(11), 2531-2560.

O'Connor, D. H., Wittenberg, G. M., & Wang, S. S. H. (2005). Graded bidirectional synaptic plasticity is composed of switch-like unitary events (Vol. 102, pp. 9679-9684).

Poil, S.-S., van Ooyen, A., & Linkenkaer-Hansen, K. (2008). Avalanche dynamics of human brain oscillations: Relation to critical branching processes and temporal correlations. *Human Brain Mapping, 29*(7), 770-777.

Sornette, D. (2004). *Critical phenomena in natural sciences : chaos, fractals, self-organization, and disorder : concepts and tools* (2nd ed.). Berlin ; New York: Springer.