Nonlinear-Dynamical Attention Allocation via Information Geometry

Matthew Ikle'[1], Ben Goertzel[2]

[1] Adams State College, Alamosa CO
[2] Novamente LLC, Rockville MD

Abstract. Inspired by a broader perspective viewing intelligent system dynamics in terms of the geometry of "cognitive spaces," we conduct a preliminary investigation of the application of information-geometry based learning to ECAN (Economic Attention Networks), the component of the integrative OpenCog AGI system concerned with attention allocation and credit assignment. We generalize Amari's "natural gradient" algorithm for network learning to encompass ECAN and other recurrent networks, and apply it to small example cases of ECAN, demonstrating a dramatic improvement in the effectiveness of attention allocation compared to prior (Hebbian learning like) ECAN methods. Scaling up the method to deal with realistically-sized ECAN networks as used in OpenCog remains for the future, but should be achievable using sparse matrix methods on GPUs.

Keywords: information geometry, recurrent networks, economic attention allocation, ECAN, OpenCog

1 Introduction

The AGI field currently lacks any broadly useful, powerful, practical theoretical and mathematical framework. Many theoretical and mathematical tools have been important in guiding the design of various aspects of various AGI systems; and there is a general mathematical theory of AGI [17], which has inspired some practical work [18] [22], but has not yet been connected with complex AGI architectures in any nontrivial way. But it is fair to say that AGI is in deep need of unifying ideas.

One possibility in this regard it *information geometry* [3], the theory of the geometric structure of spaces of probability distributions. Given the recent rise of probabilistic methods in AI and the success of geometric methods in other disciplines such as physics, this seems a natural avenue to explore. A companion paper [10] outlines some very broad ideas in this regard; here we present some more concrete and detailed experiments in the same direction. Continuing our prior work with the OpenCog [16] integrative AGI architecture, we model OpenCog's Economic Attention Networks (ECAN) component using information geometric language, and then use this model to propose a novel information geometric method of updating ECAN networks (based on an extension of Amari's ANGL

algorithm). Tests on small networks suggest that information-geometric methods have the potential to vastly improve ECAN's capability to shift attention from current preoccupations to desired preoccupations. However, there is a high computational cost associated with the simplest implementations of these methods, which has prevented us from carrying out large-scale experiments so far. We are exploring the possibility of circumventing these issues via using sparse matrix algorithms on GPUs.

2 Brief Review of OpenCog

Now we briefly describe the OCP (OCP) AGI architecture, implemented within the open-source OpenCog AI framework. OCP provides the general context for the very specific novel algorithmic research presented here.

Conceptually founded on the "patternist" systems theory of intelligence outlined in [12], OCP combines multiple AI paradigms such as uncertain logic, computational linguistics, evolutionary program learning and connectionist attention allocation in a unified architecture. Cognitive processes embodying these different paradigms interoperate together on a common neural-symbolic knowledge store called the Atomspace. The interaction of these processes is designed to encourage the self-organizing emergence of high-level network structures in the Atomspace, including superposed hierarchical and heterarchical knowledge networks, and a self-model network enabling meta-knowledge and meta-learning.

The OpenCog software (incorporating elements of the OCP architecture) has been used for commercial applications in the area of natural language processing and data mining [14], and for the control of virtual agents in virtual worlds [13] (see `http://novamente.net/example` for some videos of these virtual dogs in action).

The high-level architecture of OCP involves the use of multiple cognitive processes associated with multiple types of memory to enable an intelligent agent to execute the procedures that it believes have the best probability of working toward its goals in its current context. OCP handles low-level perception and action via an extension called OpenCogBot, which integrates a hierarchical temporal memory system, DeSTIN [4].

OCP's memory types are the declarative, procedural, sensory, and episodic memory types that are widely discussed in cognitive neuroscience [23], plus – most relevantly for the current paper – attentional memory for allocating system resources generically, and intentional memory for allocating system resources in a goal-directed way. Table 1 overviews these memory types, giving key references and indicating the corresponding cognitive processes, and also indicating which of the generic patternist cognitive dynamics each cognitive process corresponds to (pattern creation, association, etc.). The essence of the OCP design lies in the way the structures and processes associated with each type of memory are designed to work together in a closely coupled way, the operative hypothesis being that this will yield cooperative intelligence ("cognitive synergy") going

beyond what could be achieved by an architecture merely containing the same structures and processes in separate "black boxes."

Memory Type	Specific Cognitive Processes	General Cognitive Functions
Declarative	Probabilistic Logic Networks (PLN) [11]; concept blending [7]	pattern creation
Procedural	MOSES (a novel probabilistic evolutionary program learning algorithm) [20]	pattern creation
Episodic	internal simulation engine [13]	association, pattern creation
Attentional	Economic Attention Networks (ECAN) [15]	association, credit assignment
Intentional	probabilistic goal hierarchy refined by PLN and ECAN, structured according to Psi	credit assignment, pattern creation
Sensory	Supplied by DeSTIN integration	association, attention allocation, pattern creation, credit assignment

Table 1. Memory Types and Cognitive Processes in OpenCog Prime. The third column indicates the general cognitive function that each specific cognitive process carries out, according to the patternist theory of cognition.

Declarative knowledge representation is handled by a weighted labeled hypergraph called the Atomspace, which consists of multiple types of nodes and links, generally weighted with probabilistic truth values and also attention values (ShortTermImportance (STI) and LongTermImportance values, regulating processor and memory use).

OCP's dynamics has both goal-oriented and "spontaneous" aspects. The basic goal-oriented dynamic , is driven by "cognitive schematics", which take the form

$$Context \wedge Procedure \rightarrow Goal < p >$$

(summarized $C \wedge P \rightarrow G$), roughly interpretable as "If the context C appears to hold currently, then if I enact the procedure P, I can expect to achieve the goal G with certainty p."

On the other hand, the spontaneous dynamic is driven by the ECAN component (the subject of the present paper), which propagates STI values in a manner reminiscent of an attractor neural network; cognitive processes or knowledge items that get more importance spread to them are then used to trigger action or cognition or to guide perception. Goal-oriented dynamics also utilizes STI, in that the system's top-level goals are given STI to spend on nominating procedures for execution or to pass to subgoals.

3 Brief Review of Economic Attention Networks

Now we review the essential ideas underlying Economic Attention Networks (ECAN), which is the central process controlling attention allocation and credit assignment within OpenCog. ECAN is a specific approach to resource allocation and associative memory and may be considered a nonlinear dynamical system in roughly the same family as attractor neural networks such as Hopfield nets. As we describe in detail in [19] ECAN is a graph, consisting of generically-typed nodes and links (which may have any of OpenCog's node or link types, but the point is that the type semantics is irrelevant to ECAN even though it may be relevant to other OpenCog modules), and also links that may be typed either HebbianLink or InverseHebbianLink. Each Hebbian or InverseHebbian link is weighted with a probability value.

Each node or link in an ECAN is also weighted with two numbers, representing short-term importance (STI) and long-term importance (LTI). STI values represent the immediate importance of an Atom to ECAN at a particular instant in time, while LTI values represent the value of retaining atoms in memory. The ECAN equations dynamically update these values using an economic metaphor in which both STI and LTI can be viewed as artificial currencies.

The ECAN equations also contain the essential notion of an AttentionalFocus (AF), consisting of those Atoms in the ECAN with the highest STI values. The probability value of a HebbianLink from A to B is the odds that if A is in the AF, so is B; and correspondingly, the InverseHebbianLink from A to B is weighted with the odds that if A is in the AF, then B is not. The main concept here is the following: Suppose there is a high HebbianLink probability between A and B and that A is in the AF. Then A can be viewed as trying to "pull" B into the AF. There is an obvious corresponding but opposite reaction if the nodes share instead a high InverseHebbianLink.

As an associative memory, the ECAN process involves both training and retrieval processes. The entire ECAN training dynamics can be described as a nonlinear function $H : [0,\ 1]^L \longrightarrow \mathcal{R}^M$, where L is the number of nodes, and $M = L^2$, mapping a given set of binary patterns into a connection matrix C of Hebbian weights. The specific ECAN Hebbian updating equations are somewhat complex, and are described in detail in [10]. What is important in our current context, is this view of the process as a nonlinear function on the space of input patterns into the space of weight parameters.

4 Brief Review of Information Geometry

"Information geometry" is a branch of applied mathematics concerned with the application of differential geometry to spaces of probability distributions. In [10] we have suggested some extensions to traditional information geometry aimed at allowing it to better model general intelligence. However for the concrete technical work in the present paper, the traditional formulation of information geometry will suffice.

One of the core mathematical constructs underlying information geometry, is the Fisher Information, a statistical quantity which has a a variety of applications ranging far beyond statistical data analysis, including physics [8], psychology and AI [3]. Put simply, FI is a formal way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ upon which the probability of X depends. FI forms the basis of the Fisher-Rao metric, which has been proved the only Riemannian metric on the space of probability distributions satisfying certain natural properties regarding invariance with respect to coordinate transformations. Typically θ in the FI is considered to be a real multidimensional vector; however, [6] has presented a FI variant that imposes basically no restrictions on the form of θ. Here the multidimensional FI will suffice, but the more general version is needed if one wishes to apply FI to AGI more broadly, e.g. to declarative and procedural as well as attentional knowledge.

In the set-up underlying the definition of the ordinary finite-dimensional Fisher information, the probability function for X, which is also the likelihood function for $\theta \in R^n$, is a function $f(X;\theta)$; it is the probability mass (or probability density) of the random variable X conditional on the value of θ. The partial derivative with respect to θ_i of the log of the likelihood function is called the *score* with respect to θ_i. Under certain regularity conditions, it can be shown that the first moment of the score is 0. The second moment is the Fisher information:

$$\mathcal{I}(\theta)_i = \mathcal{I}_X(\theta)_i = E\left[\left(\left(\frac{\partial}{\partial\theta_i}\ln f(X;\theta)\right)^2\right)|\theta\right]$$

where, for any given value of θ_i, the expression $E[..|\theta]$ denotes the conditional expectation over values for X with respect to the probability function $f(X;\theta)$ given θ. Note that $0 \leq \mathcal{I}(\theta)_i < \infty$. Also note that, in the usual case where the expectation of the score is zero, the Fisher information is also the variance of the score.

One can also look at the whole Fisher information matrix

$$\mathcal{I}(\theta)_{i,j} = E\left[\left(\frac{\partial ln f(X,\theta)}{\partial\theta_i}\frac{\partial ln f(X,\theta)}{\partial\theta_j}\right)|\theta\right]$$

which may be interpreted as a metric g_{ij}, that provably is the only "intrinsic" metric on probability distribution space. In this notation we have $\mathcal{I}(\theta)_i = \mathcal{I}(\theta)_{i,i}$.

Dabak [6] has shown that the geodesic between two parameter vectors θ and θ' is given by the exponential weighted curve $(\gamma(t))(x) = \frac{f(x,\theta)^{1-t}f(x,\theta')^t}{\int f(y,\theta)^{1-t}f(y,\theta')^t dy}$, under the weak condition that the log-likelihood ratios with respect to $f(X,\theta)$ and $f(X,\theta')$ are finite. Also, along this sort of curve, the sum of the Kullback-Leibler distances between θ and θ', known as the J-divergence, equals the integral of the Fisher information along the geodesic connecting θ and θ'.

This suggests that if one is attempting to learn a certain parameter vector based on data, and one has a certain other parameter vector as an initial value, it may make sense to use algorithms that try to follow the Fisher-Rao geodesic

between the initial condition and the desired conclusion. This is what Amari [1] [3] calls "natural gradient" based learning, a conceptually powerful approach which subtly accounts for dependencies between the components of θ.

5 From Information Geometry to Mind Geometry

While here we will formally require only traditional ideas from information geometry, it is worth noting that the present paper was inspired by a companion paper [10] in which information geometry is extended in various ways and conjecturally applied to yield a broad conceptual model of cognitive systems. A family of alternative metrics based on algorithmic information theory is proposed, to complement the FisherRao metric – very roughly speaking, the algorithmic distance between two entities represents the amount of computational effort required to transform between the two. Multi-modular memory systems like OpenCog are then modeled in terms of multiple "mindspaces": each memory system, and the composite system, are geometrized using both Fisher-Rao and algorithmic metrics. Three hypotheses are then proposed:

1. a *syntax-semantics correlation* principle, stating that in a successful AGI system, these two metrics should be roughly correlated
2. a *cognitive geometrodynamics* principle, stating that on the whole intelligent minds tend to follow geodesics in mindspace
3. a *cognitive synergy* principle, stating that shorter paths may be found through the composite mindspace formed by considering multiple memory types together, than by following the geodesics in the mindspaces corresponding to individual memory types.

The results presented in this paper do not depend on any of these broader notions, however they fit in with them naturally. In this context, the present paper is viewed as an exploration of how to make ECAN best exploit the Fisher-Rao geometric structure of OpenCog's "attentional mindspace."

6 Information-Geometric Learning for Recurrent Networks: Extending the ANGL Algorithm

Now we move on to discuss the practicalities foinformation-geometric learning within OpenCog's ECAN component. As noted above, Amari [1, 3] introduced the natural gradient as a generalization of the direction of steepest descent on the space of loss functions of the parameter space. Issues with the original implementation include the requirement of calculating both the Fisher information matrix and its inverse. To resolve these and other practical considerations, Amari [2] proposed an adaptive version of the algorithm, the Adaptive Natural Gradient Learning (ANGL) algorithm. Park, Amari, and Fukumizu [21] extended ANGL to a variety of stochastic models including stochastic neural networks, multi-dimensional regression, and classification problems.

In particular, they showed that, assuming a particular form of stochastic feedforward neural network and under a specific set of assumptions concerning the form of the probability distributions involved, a version of the Fisher information matrix can be written as

$$G(\theta) = E_\xi \left[\left(\frac{r'}{r} \right)^2 \right] E_x \left[\nabla H \left(\nabla H \right)^T \right].$$

Although Park et al considered only feedforward neural networks, their result also holds for more general neural networks, including the ECAN network. What is important is the decomposition of the probability distribution as

$$p\left(\mathbf{y}|\mathbf{x};\theta\right) = \prod_{i=1}^{L} r_i \left(y_i - H_i \left(\mathbf{x}, \theta \right) \right)$$

where

$$\mathbf{y} = \mathbf{H}(\mathbf{x};\theta) + \xi,\ \mathbf{y} = (y_1, \cdots, y_L)^T,\ \mathbf{H} = (H_1, \cdots, H_L)^T,\ \xi = (\xi_1, \cdots, \xi_L)^T,$$

where ξ is added noise. If we assume further that each r_i has the same form as a Gaussian distribution with zero mean and standard deviation σ, then the Fisher information matrix simplifies further to

$$G(\theta) = \frac{1}{\sigma^2} E_x \left[\nabla H \left(\nabla H \right)^T \right].$$

The adaptive estimate for $\hat{G}_{t+1}^{-1}$ is given by

$$\hat{G}_{t+1}^{-1} = (1 + \epsilon_t)\hat{G}_t^{-1} - \epsilon_t (\hat{G}_t^{-1} \nabla H)(\hat{G}_t^{-1} \nabla H)^T.$$

and the loss function for our model takes the form

$$l(\mathbf{x}, \mathbf{y}; \theta) = -\sum_{\mathbf{i=1}}^{\mathbf{L}} \log \mathbf{r}(\mathbf{y_i} - \mathbf{H_i}(\mathbf{x}, \theta)).$$

The learning algorithm for our connection matrix weights θ is then given by

$$\theta_{t+1} = \theta_t - \eta_t \hat{G}_t^{-1} \nabla l(\theta_t).$$

7 Information Geometry for Economic Attention Allocation: A Detailed Example

We now present the results of a series of small-scale, exploratory experiments comparing the original ECAN process running alone with the ECAN process coupled with ANGL. We are interested in determining which of these two lines of processing result in focusing attention more accurately.

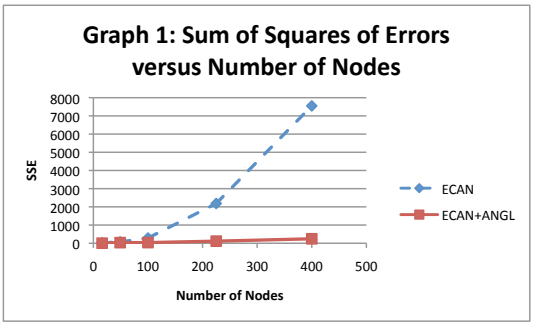

Fig. 1. Results from Experiment 1

The experiment started with base patterns of various sizes to be determined by the two algorithms. In the training stage, noise was added, generating a number of instances of noisy base patterns. The learning goal is to identify the underlying base patterns from the noisy patterns as this will identify how well the different algorithms can focus attention on relevant versus irrelevant nodes.

Next, the ECAN process was run, resulting in the determination of the connection matrix C. In order to apply the ANGL algorithm, we need the gradient, ∇H, of the ECAN training process, with respect to the input $\mathbf{x}$. While calculating the connection matrix C, we used Monte Carlo simulation to simultaneously calculate an approximation to ∇H.

After ECAN training was completed, we bifurcated the experiment. In one branch, we ran fuzzed cue patterns through the retrieval process. In the other, we first applied the ANGL algorithm, optimizing the weights in the connection matrix, prior to running the retrieval process on the same fuzzed cue patterns. At a constant value of $\sigma = 0.8$ we ran several samples through each branch with pattern sizes of 4×4, 7×7, 10×10, 15×15, and 20×20. The results are shown in Figure 1. We also ran several

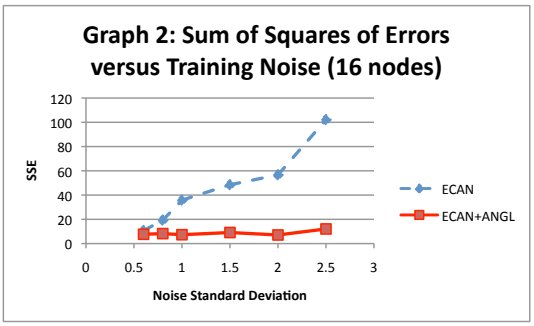

Fig. 2. Results from Experiment 2

experiments comparing the sum of squares of the errors to the input training noise as measured by the value of σ.; see Figures 2 and 3.

These results suggest two major advantages of the ECAN+ANGL combination compared to ECAN alone. Not only was the performance of the combination better in every trial, save for one involving a small number of nodes and little noise, but the combination clearly scales significantly better both as the number of nodes increases, and as the training noise increases.

8 Conclusion

Inspired by a broader geometric conception of general intelligence, we have explored a relatively simple concrete application of information-geometric ideas to the ECAN component of the OpenCog integrative AGI system. Roughly speaking, the idea explored is to have OpenCog shift its attention from current preoccupations toward desired preoccupations, based on following geodesic paths in the Fisher-Rao space of the space of "attentional probability distributions".

The results presented here are highly successful but also quite preliminary, involving small numbers of nodes in isolation rather than integrated into an entire AGI system. We still have much work ahead to determine whether the dramatic improvements reported here continue to scale with millions of nodes in a complete integrative system. Nonetheless, the results from our experiment tantalizingly suggest that incorporating ANGL into the ECAN process can lead to vastly more accurate results, especially as system size and noise increases. The main open question is whether this improvement, can be achieved for large ECAN networks without dramatically increased processing time. To address this problem, we plan to experiment with implementing ECAN+ANGL on many-core GPU machines, using optimized sparse matrix algorithms [9, 5].

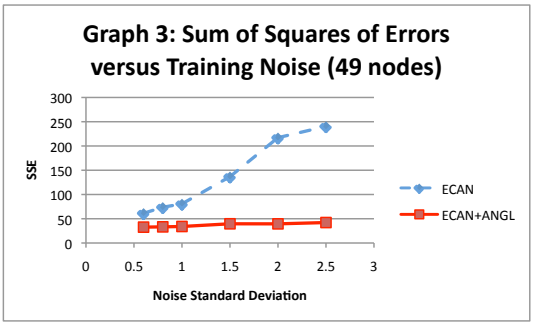

Fig. 3. Results from Experiment 3

We also plan to pursue similar approaches to improving the learning capability of other OpenCog components. For instance, OpenCog's PLN inference framework utilizes a statistically-guided inference control mechanism, which could benefit from information-geometric ideas. And OpenCog's MOSES system for probabilistic program induction (procedure learning) could potentially

be modified to more closely follow geodesics in program space. There is no lack of fertile ground for further, related experimentation.

References

1. Amari, S.: Differential-geometrical methods in statistics. Lecture notes in statistics (1985)
2. Amari, S.: Natural gradient works efficiently in learning. Neural Computing 10, 251–276 (1998)
3. Amari, S.i., Nagaoka, H.: Methods of information geometry. AMS (2000)
4. Arel, I., Rose, D., Coop, R.: Destin: A scalable deep learning architecture with application to high-dimensional robust pattern recognition. Proc. AAAI Workshop on Biologically Inspired Cognitive Architectures (2009)
5. Baskaran, M., Bordawekar, R.: Optimizing Sparse Matrix-Vector Multiplication on GPUs. IBM Research Report (2008)
6. Dabak, A.: A Geometry for Detection Theory. PhD Thesis, Rice U. (1999)
7. Fauconnier, G., Turner, M.: The Way We Think: Conceptual Blending and the Mind's Hidden Complexities. Basic (2002)
8. Frieden, R.: Physics from Fisher Information. Cambridge U. Press (1998)
9. Garland, M.: Sparse matrix computations on manycore gpu's. pp. 2–6. 45th annual Design Automation Conference:08
10. Goertzel, B., Iklé, M.: Steps toward a geometry of mind. In: Schmidhuber, J., Thorisson, K. (eds.) Subm.to AGI-11. Springer (2011)
11. Goertzel, B., M. Ikl, I.G., Heljakka, A.: Probabilistic Logic Networks. Springer (2008)
12. Goertzel, B.: The Hidden Pattern. Brown Walker (2006)
13. Goertzel, B., Et Al, C.P.: An integrative methodology for teaching embodied non-linguistic agents, applied to virtual animals in second life. In: Proc.of the First Conf. on AGI. IOS Press (2008)
14. Goertzel, B., Pinto, H., Pennachin, C., Goertzel, I.F.: Using dependency parsing and probabilistic inference to extract relationships between genes, proteins and malignancies implicit among multiple biomedical research abstracts. In: Proc. of Bio-NLP 2006 (2006)
15. Goertzel, B., Pitt, J., Ikle, M., Pennachin, C., Liu, R.: Glocal memory: a design principle for artificial brains and minds. Neurocomputing (Apr 2010)
16. Goertzel, B.e.a.: Opencogbot: An integrative architecture for embodied agi. Proc. of ICAI-10, Beijing (2010)
17. Hutter, M.: Universal AI. Springer (2005)
18. Hutter, M.: Feature dynamic bayesian networks. In: Proc. of the Second Conf. on AGI. Atlantis Press (2009)
19. Ikle, M., Pitt, J., Goertzel, B., Sellman, G.: Economic attention networks: Associative memory and resource allocation for general intelligence. Proceedings of AGI 2009
20. Looks, M.: Competent Program Evolution. PhD Thesis, Computer Science Department, Washington University (2006)
21. Park, H., Amari, S., Fukumizu, K.: Adaptive natural gradient learning algorithms for various stochastic models. Neural Computing 13, 755–764 (2000)
22. Schaul, T., Schmidhuber, J.: Towards practical universal search. In: Proc. of the 3rd Conf. on AGI. Atlantis Press (2010)
23. Tulving, E., Craik, R.: The Oxford Handbook of Memory. Oxford U. Press (2005)