Potential Energy and Particle Interaction Approach for Learning in Adaptive Systems

Deniz Erdogmus[1], Jose C. Principe[1], Luis Vielva[2], David Luengo[2]

[1] CNEL, NEB 454, University of Florida, Gainesville, Florida 32611, USA
{deniz, principe}@cnel.ufl.edu
[2] GTAS-DICOM, Universidad de Cantabria, Santander, Spain
luis@dicom.unican.es, david@gtas.dicom.unican.es

Abstract. Adaptive systems research is mainly concentrated around optimizing cost functions suitable to problems. Recently, Principe et al. proposed a particle interaction model for information theoretical learning. In this paper, inspired by this idea, we propose a generalization to the particle interaction model for learning and system adaptation. In addition, for the special case of supervised multi-layer perceptron (MLP) training we propose the interaction force backpropagation algorithm, which is a generalization of the standard error backpropagation algorithm for MLPs.

1. Introduction

Adaptive system training algorithms research has long been driven by pre-defined cost functions deemed suitable for the application. For instance, mean-square-error (MSE) has been extensively utilized as the criterion in supervised learning and adaptation, although alternatives have been proposed and investigated relatively less frequently [1]. Second order statistics, by definition, have also been the cost function for principal component analysis [2]. Other higher order statistics, including higher order cumulants like the kurtosis, high order polyspectra, etc., and information theoretic cost functions have mainly been studied in the context of blind signal processing with applications to independent component analysis (ICA), blind source separation (BSS), and blind deconvolution [3-5]. The commonality of all the research on these is that the analyses are mainly motivated by the corresponding selected adaptation criterion.

Working on the same problems, Principe et al. have utilized Renyi's quadratic entropy definition and introduced the term *information theoretical learning* to the adaptive systems literature [6]. Their nonparametric estimator for Renyi's quadratic entropy, which is based on Parzen windowing with Gaussian kernels, incited the idea of particle interactions in adaptation. Specifically considering the blind source separation problem, they have defined and demonstrated the *quadratic information forces* and the *quadratic information potential* at work in this context. Their insight on the adaptation process as an *interaction between information particles* deserves further investigation. Erdogmus and Principe have recently extended the entropy estimator to any entropy order and kernel function in Parzen windowing [7]. This

generalization of the entropy estimator also led to the extensions of the definitions of *information potential* and *force*. Successful applications of this entropy estimator in supervised and unsupervised learning scenarios have increased confidence and interest on *information theoretic learning* [8,9].

Inspired by the above-mentioned *information-particle interaction model for learning* proposed in [6], we investigate in this communication the possibility of generalizing the concept of particle interaction learning. Our aim is to determine a unifying model to describe the learning process as an interaction between *particles*, where for some special case these may be the *information particles* or for some other special case, we may end up with the commonly utilized second order statistics of the data. The formulations to be presented in the sequel will achieve these objectives and we will call this general approach the *potential energy extremization learning* (PEEL). Also, specifically applied to supervised learning, we will obtain the *minimum energy learning* (MEL). In addition, we will propose a generalized backpropagation algorithm to train MLPs under MEL principle. For the specific choice of the *potential field* (to be defined later) that reduces the minimum energy criterion to MSE, we will observe that the generalized backpropagation algorithm reduces to the standard backpropagation algorithm.

2. Adaptation by Particle Interactions

Traditionally, the adaptation process is regarded as an optimization process, where a suitable pre-defined performance criterion is maximized or minimized. In this alternative view, we will treat each sample of the training data set as a particle and let these particles interact with each other according to the interactions laws that we define. The parameters of the adaptive system will then be modified in accordance with the interactions between the particles.

2.1 Particle Interaction Model

Suppose we have the samples $\{z_1,\ldots,z_N\}$ generated by some adaptive system. For simplicity, assume we are dealing with single dimensional random variables; however, note that extensions to multi-dimensional situations are trivial. In the particle interaction model, we assume that each sample is a particle and a potential field is emanated from it. Suppose z_i generates a potential energy field. If the potential field that is generated by each particle is $v(x)$, we require this function to be continuous and differentiable, and to satisfy the even symmetry condition $v(x) = v(-x)$. Notice that due to the even symmetry and differentiability, the gradient of the potential function at the origin is zero. With these definitions, we observe that the potential energy of particle z_j due to particle z_i is $V(z_j|z_i) = v(z_j - z_i)$. The total potential energy of z_j due to all the particles in the training set is then given by

$$V(z_j) = \sum_{i=1}^{N} V(z_j \mid z_i) = \sum_{i=1}^{N} v(z_j - z_i)\,. \tag{1}$$

Defining the interaction force between these particles, in analogy to physics, as

$$F(z_j \mid z_i) \stackrel{\Delta}{=} \partial V(z_j \mid z_i)/\partial z_j = \partial v(x)/\partial x\big|_{x=(z_j - z_i)} = v'(z_j - z_i)\,. \tag{2}$$

we obtain the total force acting on particle z_j

$$F(z_j) = \sum_{i=1}^{N} F(z_j \mid z_i) = \sum_{i=1}^{N} v'(z_j - z_i)\,. \tag{3}$$

Notice that the force applied to a particle by itself is $F(z_j \mid z_j) = v'(0) = 0$. Finally, the total potential energy of the sample set is the sum (possibly weighted) of the individual potentials of each particle. Assuming that each particle is weighted by a factor $g(z_j)$ that may depend on the particle's value, which may as well be independent from the value of the particle, but different for each particle, the total energy of the system of particles is found to be

$$V(z) = \sum_{j=1}^{N} g\ (z_j) \sum_{i=1}^{N} v(z_j - z_i)\,. \tag{4}$$

Assuming that $g(z_j)=1$ for all samples, we can determine the sensitivity of the overall potential of the particle system with respect to the position of a specific particle z_j. This is given by

$$\frac{\partial V(z)}{\partial z_k} = \frac{\partial}{\partial z_k} \sum_{j=1}^{N} \sum_{i=1}^{N} v(z_j - z_i) = \ldots = 2F(z_k)\,. \tag{5}$$

In the adaptation context, since the samples are generated by a parametric adaptive system, the sensitivity of the total potential with respect to the weights of the system is also of interest. This sensitivity is directly related to the interaction forces between the samples as follows

$$\frac{\partial V}{\partial w} = \frac{\partial}{\partial w} \sum_{j=1}^{N} \sum_{i=1}^{N} v(z_j - z_i) = \ldots = \sum_{j=1}^{N} \sum_{i=1}^{N} F(z_j \mid z_i) \left(\frac{\partial z_j}{\partial w} - \frac{\partial z_i}{\partial w} \right). \tag{6}$$

2.2 Some Special Cases

Consider for example the potential function choice of $v(x) = x^2/(2N^2)$ and weighting function choice of $g\ (z_j) = 1$ (i.e. unweighted) for all samples. Then upon direct substitution of these values in (4), we obtain $V(z)$ equals the biased sample variance, i.e. minimization of this potential energy will yield the minimum variance solution for the weights of the adaptive system. In general, if we select potential functions of the form $v(x) = |x^{\,p}|$, where $p>1$, with no weighting of the particles we obtain cost functions of the form

$$V(z) = \sum_{j=1}^{N} \sum_{i=1}^{N} \left| (z_j - z_i)^{\,p} \right|. \tag{7}$$

which are directly related to the absolute central moments of the random variable Z, for which z_j's are samples. Each value of p corresponds to a different choice of the

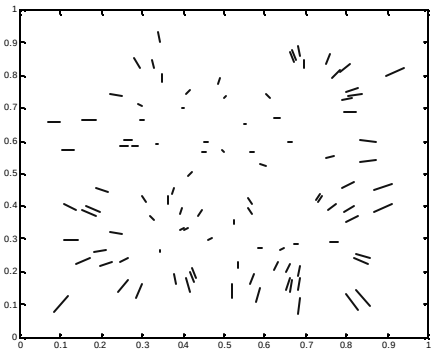

Fig. 1. A snapshot of the information particles (output vector samples) and the instantaneous information forces acting on these particles in two-dimensional BSS.

distance metric between the particles from the family of Minkowski norms.

The information potential estimators of [6] and [7] also fall into this same category of cost energy functions. The quadratic information potential (based on Renyi's quadratic entropy) estimator in [6], which uses Gaussian kernels $G_s(.)$ with standard deviation s (named the kernel size), is

$$V_2(z) = \frac{1}{N^2} \sum_{j=1}^{N} \sum_{i=1}^{N} G_s(z_j - z_i) . \tag{8}$$

The generalized information potential estimator in [7], on the other hand is

$$V_a(z) = \frac{1}{N^a} \sum_{j=1}^{N} \left(\sum_{i=1}^{N} k_s(z_j - z_i) \right)^{a-1} . \tag{9}$$

In (9), a is the entropy order for Renyi's definition and $k_s(.)$ is the kernel function, which must be a valid pdf. Notice that for the potential function choice $v(x) = G_s(x) / N^2$ and $g(z_j) = 1$ in (4), we obtain the quadratic information potential of (8). Additionally, for $v(x) = k_s(x) / N^2$ and $g(z_j) = \hat{p}(z_j)^{a-2}$, we obtain (9) from (4). In the latter, $\hat{p}(z_j)$ is the Parzen window estimate, using kernel $k_s(.)$, of the probability density of particle z_j [10].

2.3 Illustration of Information Forces in Independent Component Analysis

As an example consider the quadratic information forces acting on the samples in a two-dimensional ICA/BSS scenario where the topology is a square matrix of weights followed by nonlinearities matched to the cumulative densities of the sources as described in [6]. Renyi's quadratic joint entropy of the outputs of the nonlinearities is to be maximized to obtain two independent sources. It is shown in [6] that

maximizing Renyi's quadratic entropy is equivalent to minimizing the quadratic information potential given in (8). In this expression, a circular two-dimensional Gaussian kernel is employed as the potential field emanating from each particle and this is used to evaluate the information forces between particles. Under these circumstances, a snapshot of the particles and the instantaneous quadratic information forces, which can be calculated from (3), acting on these particles are shown in Fig. 1. Since the optimal solution is obtained when the joint entropy is maximized, these forces are repulsive and as clearly seen in the figure, the particles repel each other to arrive at a uniform distribution in the unit square in the two-dimensional output space.

3. BackPropagation of Interaction Forces in MLPs

In this section, we will derive the backpropagation algorithm for an MLP trained supervised under the MEL principle. This extended algorithm backpropagates the interaction forces between the particles through the layers instead of the error, as is the case in the standard MSE criterion case. For simplicity, consider the unweighted potential of the error as the cost function. For multi-output situations, we simply sum the potentials of the error signals from each output. Assume the MLP has l layers with m_o processing elements (PE) in the o^{th} layer. We denote the input vector with layer index zero. Let w_{ji}^{o} be the weight connecting the i^{th} input to the j^{th} output in the o^{th} layer. Let $v_j^o(s)$ be the synapse potential of the j^{th} PE at o^{th} layer corresponding to the input sample $x(s)$, where s is the sample index. Let j (.) be the sigmoidal nonlinearity of the MLP, same for all PEs, including the output layer. Assume v(.) is the potential function of choice and we have N training samples. The total energy of the error particles, where $e_k(t)$ is the error at the k^{th} output for training sample t is then

$$V = \sum_{s=1}^{N}\sum_{t=1}^{N}\sum_{k=1}^{m_l} v(e_k(s) - e_k(t)) \sum_{s=1}^{N}\sum_{t=1}^{N} e(s,t). \tag{10}$$

In order to save space, we skip the derivation and present the algorithm. It suffices to tell that the derivation of the algorithm follows the same lines as the derivation of the standard backpropagation, which can be found in numerous textbooks on neural networks [1]. In the algorithm below, h is the learning rate and $j'(.)$ is the derivative of the MLP's sigmoid function.

Algorithm. Let the interaction force acting on sample s due to the potential field of sample t be $F(e_j(s) \mid e_j(t)) = v'(e_j(s) - e_j(t))$ in the j^{th} output node of the MLP. These interactions will minimize the energy function in (10).

1. Evaluate local gradients for the output layer for $s,t=1,\ldots,N$ and $j=1,\ldots,m_l$ using

$$d_j^l(s \mid t) = -F(e_j(s) \mid e_j(t)) \cdot j'(v_j^l(s)), \quad d_j^l(t \mid s) = -F(e_j(t) \mid e_j(s)) \cdot j'(v_j^l(t))$$

2. For layer index o going down from l-1 to 1 evaluate the local gradients

$$d_j^o(s \mid t) = j'(v_j^o(s)) \sum_{k=1}^{m_{o+1}} d_k^{o+1}(s \mid t) w_{kj}^{o+1}, \quad d_j^o(t \mid s) = j'(v_j^o(t)) \sum_{k=1}^{m_{o+1}} d_k^{o+1}(t \mid s) w_{kj}^{o+1}$$

3. For each layer index o from 1 to l evaluate the weight updates (to minimize V)

$$\Delta w_{ji}^{o} = -\eta \left(d_j^{o}(s \mid t) y_i^{o-1}(s) + d_j^{o}(t \mid s) y_i^{o-1}(t) \right)$$

Notice that for the squared error criterion with $v(x) = x^2$, the interaction force becomes $F(e_j(s) \mid e_j(t)) = 2(e_j(s) - e_j(t))$ and the algorithm reduces to the backpropagation of error values.

4. Discussion

Adaptive systems research is traditionally motivated by the optimization of suitable cost functions and is centered on the investigation of learning algorithms that achieve the desired optimal solution. In this paper, inspired by the idea of *information theoretic learning through particle interactions*, we have proposed an alternative approach to adaptation and learning. This new approach allows us to regard this process in analogy with interacting particles in a force field in physics. Besides the intellectual appeal of this viewpoint provides us for further theoretical study on learning, it may be promising in designing real systems that utilize physical forces to change its state and eventually adapt to its environment to need. It might also fascilitate self-organization in distributed systems, through pairwise interactions.

References

1. Haykin, S.: Neural Networks: A Comprehensive Foundation. 2nd edn. Prentice Hall, New Jersey (1999)
2. Oja, E.: Subspace Methods for Pattern Recognition. Wiley, New York (1983)
3. Haykin, S. (ed.): Unsupervised Adaptive Filtering, Vol. 1: Blind Source Separation. Wiley, New York (2000)
4. Haykin, S. (ed.): Unsupervised Adaptive Filtering, Vol. 2: Blind Deconvolution. Wiley, New York (2000)
5. Hyvarinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley, New York (2001)
6. Principe, J.C., Xu, D., Fisher, J.W.: Information Theoretic Learning. In: Haykin, S. (ed.): Unsupervised Adaptive Filtering, Vol. 1: Blind Source Separation. Wiley, New York (2000)
7. Erdogmus, D. Principe, J.C.: Generalized Information Potential Criterion for Adaptive System Training. To appear in IEEE Trans. in Neural Networks (2002)
8. Santamaria, I., Erdogmus, D., Principe, J.C.: Entropy Minimization for Digital Communications Channel Equalization. To appear in IEEE Trans. on Signal Processing (2002)
9. Torkkola, K., Campbell, W.M.: Mutual Information in Learning Feature Transformations. In: Proceedings of the International Conference on Machine Learning. Stanford (2000)
10. Parzen, E.: On Estimation of a Probability Density Function and Mode. In: Time Series Analysis Papers. Holden-Day, California (1967)