NATURAL ELASTIC NETS FOR FAITHFUL REPRESENTATIONS

Jiann-Ming Wu and Zheng-Han Lin

Department of Applied Mathematics,

National Donghwa University, Hualien, Taiwan, R.O.C.

Email: *jmwu*@server.am.ndhu.edu.tw

ABSTRACT

In this work, we derive self-organization by constructing a generative mode, which is composed of piecewise multivariate Gaussian distributions for characterizing the parameter space. The fitness of this generative model to all parameters provides a smoothness criterion to the essential exploration of the clustering topology within the parameter space. Combined with the minimal wiring criterion proposed by Durbin and Willshaw, the new criterion is able to produce faithful representations of self-organization. We apply a hybrid of the mean field annealing and the gradient descent method to the optimization of mathematical framework in treatment of discrete combinatorial variables and continuous geometrical variables, and obtain three sets of interactive dynamics to which the corresponding unsupervised learning process is termed as natural elastic net algorithm. If the covariance matrix of each piecewise multivariate Gaussian distribution is fixed as an identity matrix, the interactive dynamics describe a learning process for the elastic net of Durbin and Willshaw.

I. INTRODUCTION

A self-organizing algorithm aims to form a coherent mapping from a parameter space R^d to a cortex like space, such as a $M \times M$ two dimensional lattice structure. In the Kohonen algorithm [5] and the elastic net algorithm [2], each node on the lattice is attached with a cortical point and all cortical points collectively constitute internal representations for the parameter space. These cortical points $\{y_k \in R^d, 1 \le k \le M^2\}$ form a nonoverlapping Voronoi partition $\{\Omega_k(y)\}$ into the parameter space, where $\Omega_k(y) = \{x \mid \arg\min_j \|x - y_j\| = k, x \in R^d\}$, with $\|x-y\|$ as the Euclidean distance and the property of a nonoverlapping partition $\bigcup_k \Omega_k(y) = R^d$ and $\Omega_k(y) \bigcap \Omega_h(y) = \Phi$, for all $k \neq h$. These two algorithms aim to find cortical points by learning samples from the parameter space subject to prior criteria. By Voronoi partition, each point in the parameter space is mapped to one and only one node on the lattice, and this mapping is expected to produce a dimensional reduction mapping from d to 2 with topology preserving properties. Such a dimensional-reduction mapping has been used to explore the embedded self-organization of ocular dominance bands and the orientation module of the visual cortex [1].

Proper prior criteria for a self-organizing algorithm are reviewed to include metric multidimensional scaling, minimal wiring, minimal path length and minimal distortion [4], of which all are based on the Euclidean distance. In practical applications, parameter samples may be generated in a stochastic process that defines a statistical dependence among the components of them. This causes non-faithful representations for the parameter space in using the Euclidean distance for the measure of similarity measure [8]. This work explores weighted measures for similarity to release the assumption that components of samples are statistical independent and aims to achieve faithful representations [7][8] for the parameter space. Here the Mahalanobis distance measures the distance between two samples.

The natural elastic net developed in this work possesses three sets of interactive dynamics for the competitive mechanism of underlying data clustering and independent component analysis. The elastic net algorithm of Durbin and Willshaw is proved to be a special case of the natural elastic net in this work. The natural elastic net is expected to provide an effective competitive mechanism to exploring the formation of ocular dominance and orientation structure in primary visual cortex [10][11]. We start our derivation at a generative model, consisting of piece-wise multivariate Gaussian distributions, characterizing the parameter space. Then we consider a log likelihood function as the measure of the fitness of this generative model to all training samples. Combined with the criterion of the minimal wiring principle, this measure forms a mathematical framework, including objectives and a set of constraints, for an essential coherent mapping. A hybrid of the mean field annealing and the gradient descent method is applied to the optimization of this mathematical framework. As a result, three sets of interactive dynamics are obtained for the unsupervised learning process of self-organization, of which an artificial temperature similar to the process of physical annealing modulates the evolution.

This article is organized as follows. A mathematical framework for the natural elastic net is developed in section II. Three sets of interactive dynamics for the natural elastic net are derived in section III. And we conclude our work in the last section.

II. A MATHEMATICAL FRAMWORK FOR NATURAL ELASTIC NET

Based on the Mahalanobis distance, a modified Voronoi partition of kernels $\{y_k\}$ into parameter space contains a set of non-overlapping internal regions $\{\Omega_k\}$ with $\Omega_k(y)=\left\{x \mid \arg\min_j \|x-y_j\|_A = k, x\in R^d\right\}$, $1\le k\le K$, where $\|x\|_A = x^t Ax$. Each region Ω_k is associated with a local generative model in a multivariate Gaussian distribution $P_k(x)$ centered at y_k with a common covariance matrix A like

$$P_k(x)=\frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|A^{-1}|}}\exp\left[-\frac{(x-y_k)^t A(x-y_k)}{2}\right] \quad (1)$$

Where $|A|$ denotes the determinant of a matrix A.

N membership vectors $\{\delta_i\}$ denote the mapping from N training samples to internal regions. Each vector $\delta_i=\{\delta_{i1},\cdots,\delta_{iK}\}$ belongs the set $\{e^k, 1\le k\le K\}$, where e^k is a standard unitary vector with the *kth* element one and the others zero. That $\{\delta_i\}$ is equal to e^k represents the *ith* training sample is mapped to the *kth* region. For consistency of the internal representation, the log of the local likelihood function l_k measures the fitness of the local generative model P_k to all training samples in the *kth* region, where

$$l_k=\log\left(\prod_{x_i\in\Omega_k} p_k(x_i)\right)=\sum_{x_i\in\Omega_k}\log p_k(x_i) \quad (2)$$

By summing up all l_k, we have the following log likelihood function

$$l=\sum_k l_k \quad (3)$$

$$=\sum_k\sum_{x_i\in\Omega_k}\log p_k(x_i)$$

$$=\sum_i\sum_k \delta_{ik}\log p_k(x_i)$$

$$=\sum_i\sum_k \delta_{ik}\left[-\frac{(x_i-y_k)^t A(x_i-y_k)}{2}-\frac{1}{2}\log|A^{-1}|-\frac{d}{2}\log(2\pi)\right]$$

$$=-\frac{1}{2}\sum_i\sum_k \delta_{ik}(x_i-y_k)^t A(x_i-y_k)-\frac{N}{2}\log|A^{-1}|-\frac{Nd}{2}\log(2\pi)$$

By neglecting the last constant term, reversing the sign and using the fact $|A^{-1}|=|A|^{-1}$, we obtain the first objective for the natural elastic net as follows

$$E_1=\frac{1}{2}\sum_i\sum_k \delta_{ik}(x_i-y_k)^t A(x_i-y_k)-\frac{N}{2}\log|A| \quad (4)$$

The maximization of l turns to be the minimization of E_1. The transition from the second line to the third line in the above derivation uses the fact $\sum_k \delta_{ik}=1$ *and* $\bigcap_k \Omega_k=\Phi$. A coherent mapping insists on that nearby points in the parameter space are ordered as smooth as possible on the cortex-like map, and the objective function costs between neighboring cortical points are as smooth as possible. The resulting dimensional-reduction mapping possesses a high reliable topology preserving property. The objective E_1 circumspective sets up a smooth generative model for the parameter space. The minimal wiring principle used by Durbin and Willshaw proposes another criterion

$$E_2=\frac{1}{2}\sum_k\sum_{j\in NB(k)}\|y_j-y_k\|_A \quad (5)$$

where $NB(k)$ denotes all neighboring nodes of the *kth* node on the lattice.

A weighted combination of minimizing E_1 and E_2, leads to the following mathematical framework for the natural elastic net, minimizing

$$E=E_1+CE_2$$

$$=\frac{1}{2}\sum_i\sum_k \delta_{ik}(x_i-y_k)^t A(x_i-y_k)$$

$$-\frac{N}{2}\log|A|+\frac{C}{2}\sum_k\sum_{j\in NB(k)}\|y_j-y_k\| \quad (6)$$

Subject to

$$\sum_k \delta_{ik}=1 \quad , \quad \text{for every i} \quad (7)$$

III. DYNAMICS FOR THE NATURAL ELASTIC NET

The above mathematical framework involves the optimization of discrete combinatorial variables of $\{\delta_i\}$ and continuous geometrical variables of $\{y_k\}$ and the matrix A, exactly forming a mixed integer and linear programming. Since the energy function that has discrete variables is not differentiable with respect to these discrete variables, the gradient descent method is not applicable. To overcome the computational difficulty, we relate each membership vector δ_i to a Potts neuron and use the mean field equation to find its mean activation at each temperature. The mean field equation can be derived by the following free energy,

which is similar to that proposed by Peterson and Söderberg [9].

$$\psi(A,Y,\langle\delta\rangle,u)=E(A,Y,\langle\delta\rangle)+\sum_i\sum_k\langle\delta_{ik}\rangle u_{ik}$$
$$-\frac{1}{\beta}\sum_i\ln\left(\sum_k\exp(\beta u_{ik})\right) \quad (8)$$

where $Y,\langle\delta\rangle$ and u denote the set $\{y_k\}$, $\{\langle\delta_i\rangle\}$ and $\{u_i\}$ respectively, β is the inverse of an artificial temperature, and each u_i is an auxiliary vector.

By setting

$$\frac{\partial\psi}{\partial\langle\delta_{ik}\rangle}=0 \quad \text{for all i, k} \quad (9)$$

$$\frac{\partial\psi}{\partial u_{ik}}=0 \quad \text{for all i, k} \quad (10)$$

We have the following mean field equation for evaluating mean activations of discrete neural variables

$$u_{ik}=-\frac{\partial E}{\partial\langle\delta_{ik}\rangle}=-\frac{1}{2}(x_i-y_k)^t A(x_i-y_k) \quad (11)$$

$$\langle\delta_{ik}\rangle=\frac{\exp(\beta u_{ik})}{\sum_k\exp(\beta u_{ik})} \quad (12)$$

Based on the mean configuration, we can apply the gradient descent method to the adaption of each y_k. That is

$$\Delta y_k\propto-\frac{\partial E}{\partial y_k}$$
$$=\frac{1}{2}\sum_i\langle\delta_{ik}\rangle(A+A^t)(x_i-y_k)$$
$$+\frac{C}{2}\sum_{j\in NB(k)}(A+A^t)(y_j-y_k) \quad (13)$$

To an zero gradient $\Delta y_k=0$, we have

$$y_k=\frac{\left(\sum_i\langle\delta_{ik}\rangle x_i+C\sum_{j\in NB(k)}y_j\right)}{\left(\sum_i\langle\delta_{ik}\rangle+CN_k\right)} \quad (14)$$

Where N_k denotes the number of nodes in the set $NB(k)$.

The updating method of each element A_{ab} in the covariance matrix is derived as follows

$$\Delta A_{ab}\propto-\frac{\partial E}{\partial A_{ab}}$$
$$=-\frac{1}{2}\sum_i\sum_k\langle\delta_{ik}\rangle(x_{ia}-y_{ka})(x_{ib}-y_{kb})+\frac{N}{2}\left[(A^t)^{-1}\right]_{ab}$$
$$-\frac{C}{2}\sum_k\sum_{j\in NB(k)}(y_{ja}-y_{ka})(y_{jb}-y_{kb}) \quad (15)$$

Again, when $\Delta A_{ab}=0$, we have

$$A=(W^{-1})^t \quad (16)$$

Where

$$W=\frac{1}{N}\sum_i\sum_k\langle\delta_{ik}\rangle(x_{ia}-y_{ka})(x_{ib}-y_{kb})$$
$$+\frac{C}{N}\sum_k\sum_{j\in NB(k)}(y_{ja}-y_{ka})(y_{jb}-y_{kb}) \quad (17)$$

The following step-by-step statement describes the natural elastic net algorithm for finding the minimum of the objective function (6).

1. Initialize β as a sufficiently small value, $A=0.01\times I$ (identity matrix), $y_k\approx\frac{1}{N}\sum_i x_i$, $\langle\delta_{ik}\rangle\approx\frac{1}{K}$

2. Update $\langle\delta_{ik}\rangle$ by equations (11) and (12).

3. Update y_k by equation (14).

4. Update *A* by equations (17) and (16).

5. If $\sum_i\sum_k\langle\delta_{ik}\rangle^2>\theta$ then halt, else

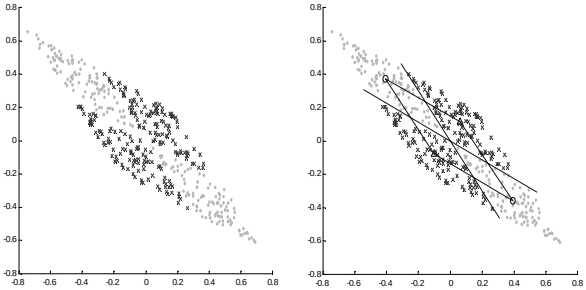
$\beta=\beta\times\frac{1}{0.95}$ Go to step2

Where θ is a threshold, ex. $\theta=0.95\times N$.

IV. NUMERICAL SIMULATIONS AND CONCLUSIONS

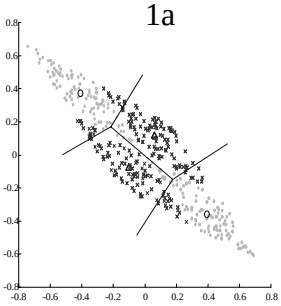
Consider the 400 training parameters in figure *1a*, which are generated by a linear mixture of two independent uniform distributions. Two components of these input parameters are not more statistical independent. We applied 2×2 natural elastic net to learn this training set and obtained an equivalent partition into the primitive cell as shown in figure *1b*. When fixing A as an identity matrix, we obtained a Voronoi partition as shown in figure *1c*, a result of the elastic net proposed by Durbin and Willshaw. When we multiplied all input parameters and kernels by a

de-mixing matrix B, which could be obtained by solving $B^t B = A$, by the input space, a new system with independent components.



1a

1b



1c

1a. Training parameters

1b. A equivalent partition obtained by the natural elastic net

1c. Voronoi partition obtained by the elastic net.

We have used piecewise multivariate Gaussian distributions to construct a generative model for input parameters in developing the natural elastic net. The fitness of this generative model to all training samples combined with the minimal wiring objective constitutes an optimization framework of the novel unsupervised learning. We have shown that a hybrid of mean field annealing and the gradient descent method is applicable to the development of interactive dynamics of the natural elastic net. As a special case of fixing the covariance matrix in the generative model as an identity matrix, the elastic net of Durbin and Willshaw fails to produce faithful representations when facing a problem with statistical dependent components. This difficulty is properly overcome by the natural elastic net. Applying the natural elastic net to blind source separation and artificial visible systems is our urgent future work.

V. REFERENCES

[1] Blasdel, G. and Salama, G.(1986). Voltage-sensitive dyes reveal a modular organization in monkey striate cortex. J. Neurosci, 12(8), 3139-3161.

[2] Durbin, R. and D. Willshaw.(1987). An Analogue Approach to the Traveling Salesman Problem Using an Elastic Net Method. Nature 326, 689-691.

[3] Durbin R., Mitchison G.(1990). A dimension reduction framework for cortical maps, Nature 343, 644-647.

[4] Goodhill G.J., Finchs., Sejnowski T.J.(1996). Optimizing Cortical Mappings, in Touretzky D.S., et al.(eds.), Advances in Neural Information Processing Systems 8, MIT Press, Cambridge/Boston/London, 330-336.

[5] Kohonen, T.(1982). Self-Organized Formation of Topologically Correct Feature Maps. Biological Cybernetics 43, 59-69.

[6] Liou C. Y., Wu J. M.(1996), Self-Organization using Potts models, Neural networks vol. 9, no. 4, 671-684.

[7] Lin J.K., Cowan J.D. and Grier D.G.(1997). Source Separation and Density Estimation by Faithful Equivalent SOM, in Mozer M.C., etal.(eds.), Advances in Neural Information Processing Systems 9, MIT Press/Bradford Books, Cambridge/London, 536-542.

[8] Lin J.K., Cowan J.D. and Grier D.G.(1997). Faithful Representation of Separable Distributions, Neural computation 9, 1305-1320.

[9] Peterson C. and Söderberg B.(1989), A new method for mapping optimization problems onto neural network, Int. J. Neural Syst. 1,3.

[10] Piepenbrock C., Ritter H., Obermayer K.(1997), The Joint Development of Orientation and Ocular Dominance: Role of Constraints, Neural Computation, Vol 9, No 5.

[11] Piepenbrock C., Ritter H., Obermayer K.(1998), Effects of lateral competition in the primary visual cortex on the development of topographic projections and ocular dominance maps.

[12] Yullle A. L., Kolodny J. A. Lee C. W.(1996), Dimension Reduction, Generalized Deformable models and the Development of ocularity and orientation, Neural networks, vol. 9, no. 2, 309-319.