The Weighted EM Learning and Monitoring Structure

6 G－4

Yasuo MATSUYAMA

† Department of Electrical, Electronics & Computer Engineering,
Waseda University, Tokyo, 169 Japan

I. INTRODUCTION

Computing an expectation and maximization is a set of powerful tools in statistical data processing. Dempster et al. [DEM78] collected examples from diverse areas, built a unified theory and coined the name "EM algorithm." Then, Jordan and Jacobs [JOR94] connected their leaning strategy on hierarchical mixtures of experts with this EM algorithm. These algorithms heavily depend on the logarithm and the nonnegativity of the Kullback-Leibler's divergence. Yet, there is a wider class for such an information measure; the divergence of order α. This paper uses the generalized measure in order to derive a probability weighted EM algorithm and learning strategies. Extended versions of statistics methods such as the Fisher's measure of information and the Cramér-Rao's bound appear in the execution of learning. Finally, usage of the generalized EM algorithm as a building block is discussed.

II. PRELIMINARIES

2.1 Rényi's divergence and α-divergence

Rényi [REN60] is the first who presented the order α divergence. We replace his α by $(1+\alpha)/2$ and normalize the total amount.

$$D_R^{(\alpha)}(p\|q) = -4/(1-\alpha^2)\log\{\textstyle\sum_i p_i(q_i/p_i)^{(1+\alpha)/2}\}.$$

If p and q are continuous, the summation is replaced by an integration.

The α-divergence [HAV67], [AMA93] has the same kernel of the summand.

$$D^{(\alpha)}(p\|q) = 4/(1-\alpha^2)\{1-\textstyle\sum_i p_i(q_i/p_i)^{(1+\alpha)/2}\}$$

There is a monotonic relationship between the above two measures.

2.2 Extended Logarithm

Csiszár [CSI72] (its early version is by Rényi [REN60]) presented a general divergence measure:

$$D_C(p\|q) = \textstyle\sum_i p_i f(q_i/p_i) = \sum_i q_i g(p_i/q_i),$$

where f and g are twice differentiable convex function with $f(1) = g(1) = 0$. Thus, the α-divergence $D^{(\alpha)}(p\|q)$ is the case of $f(x) = 4\{x - x^{(1+\alpha)/2}\}/(1-\alpha^2)$. If $\alpha = -1$, then $D^{(\alpha)}(p\|q)$ is reduced to the well-known Kullback-Leibler divergence. Thus, $\log x$ corresponds to $\{2/(1+\alpha)\}\{x^{(1+\alpha)/2} + c(x,\alpha)\}$. The term $c(x,\alpha)$ is not unique as long as it satisfy a couple of necessary conditions. We select $c(x,\alpha) = -1$. Therefore,

$$L^{(\alpha)}(x) = \tfrac{2}{1+\alpha}(x^{\frac{1+\alpha}{2}} - 1)$$

is selected as an extension of $\log x$. Note that the constant "-1" is often cancelled out when $L^{(\alpha)}(x)$ is differetiated by a parameter or compared with other $L^{(\alpha)}(y)$.

III. WEIGHTED EM FROM α-DIVERGENCE

Let $p_{Y|\phi}(y|\phi)$ be the probability density of observed data y, Let x be the complete data which contains unknown parts for the observer. ϕ and ψ denote structures defining the probability densities. The simplest case is that ϕ and ψ are parameters. Thus,

$$p_{Y|\phi}(y|\phi) = \int_{\mathcal{X}(y)} p_{X|\phi}(x|\phi)dx$$

is the relationship describing the observation. Let

$$R_Y^{(\alpha)}(\psi|\phi) = \{p_{Y|\psi}(y|\psi)/p_{Y|\phi}(y|\phi)\}^{(1+\alpha)/2},$$

and

$$L_Y^{(\alpha)}(\psi|\phi) = 2\{R_Y^{(\alpha)}(\psi|\phi) - 1\}/(1+\alpha).$$

Let the conditional probability be

$$p_{X|Y,\phi}(x|y,\phi) = p_{X|\phi}(x|\phi)/p_{Y|\phi}(y|\phi).$$

Then, we use the following convention:

$$\phi \leftrightarrow p_{X|Y,\phi}(x|y,\phi); \qquad \psi \leftrightarrow p_{X|Y,\psi}(x|y,\psi).$$

Then, $D^{(\alpha)}(\phi\|\psi) \geq 0$ gives the following equations.

$$S_{Y|X}^{(\alpha)}(\psi|\phi) \stackrel{\text{def}}{=} \int_{\mathcal{X}(y)} p_{X|Y,\phi}(x|y,\phi)\left\{\frac{p_{Y|X,\psi}(y|x,\psi)}{p_{Y|X,\phi}p(y|x,\phi)}\right\}^{\frac{1+\alpha}{2}} dx,$$

$$Q_{Y|X}^{(\alpha)}(\psi|\phi) = 2\{S_{Y|X}^{(\alpha)}(\psi|\phi) - 1\}/(1+\alpha),$$

$$\frac{2}{1-\alpha}L_Y^{(\alpha)}(\psi|\phi) \geq \frac{2}{1-\alpha}Q_{Y|X}^{(\alpha)}(\psi|\phi).$$

The following theorem and corollary are obtained.

[Theorem 3.1] The weighted EM algorithm is a series of applications of E-step and M-step.

E-step:

Compute $Q_{Y|X}^{(\alpha)}(\psi|\phi)$.

M-step:

Compute $\psi^* = \arg\max_\psi Q_{Y|X}^{(\alpha)}(\psi|\phi)$.

[Corollary 3.2] The weighted EM algorithm is classified into the following three cases depending on the number α.

E-step:

Compute $S_{Y|X}^{(\alpha)}(\psi|\phi)$.

M-step:

Compute the followings.

1. $\alpha < -1$: $\psi^* = \arg\min_\psi S_{Y|X}^{(\alpha)}(\psi|\phi)$,
2. $\alpha = -1$: $\psi^* = \arg\max_\psi E_{p_{X|Y,\phi}}\left[\log p_{Y|X,\psi}\right]$,
3. $\alpha > -1$: $\psi^* = \arg\max_\psi S_{Y|X}^{(\alpha)}(\psi|\phi)$.

重み付きEM学習とモニタ構造
松山泰男
早稲田大学理工学部電気電子情報工学科
〒169 東京都新宿区大久保3-4-1

IV. WEIGHTED EM FOR NEURAL NETWORKS OF HIERARCHICAL EXPERTS

4.1 WEM for Hierarchical Experts NN

In the neural network of hierarchical experts [JOR94], random variables have the following correspondences.

$$X \leftrightarrow X; \qquad Y \leftrightarrow (Y, Z).$$

The random variable X stands for an input, Y is the teacher, and Z gives a path of the hierarchy. Then,

$$S^{(\alpha)}_{YZ|X}(\psi|\phi) = E_{p_{Z|XY,\phi}}\left[\left\{\frac{p_{YZ|X,\psi}(y,z|x,\psi)}{p_{YZ|X,\phi}(y,z|x,\phi)}\right\}^{\frac{1+\alpha}{2}}\right],$$

$$Q^{(\alpha)}_{YZ|X}(\psi|\phi) = \frac{2}{1+\alpha}\left\{S^{(\alpha)}_{YZ|X}(\psi|\phi) - 1\right\},$$

and

$$\psi^* = \arg\max_{\psi} Q^{(\alpha)}_{YZ|X}(\psi|\phi).$$

4.2 Gradient Ascent Learning Based on WEM

Let $U^{(\alpha)}_{ij}(\phi)$ be a neural weight connecting element j to i in the probability world of ϕ. Let ψ be a post-learning world. Then $\psi = \phi + \Delta\phi$ corresponds to

$$\begin{cases} U^{(\alpha)}_{ij}(\psi) = U^{(\alpha)}_{ij}(\phi) + \Delta U^{(\alpha)}_{ij}(\phi), \\ \Delta U^{(\alpha)}_{ij}(\psi) = \rho\frac{\partial}{\partial U_{ij}}\left[\frac{2}{(1+\alpha)}\left\{p_{Y|X,\phi}(y|x,\phi)\right\}^{\frac{1+\alpha}{2}}\right] \end{cases}$$

where ρ is a small constant. First, we consider a successive version of the learning. The increment is as follows.

$$\Delta U^{(\alpha)}_{ij}(\phi) = \rho p_{Y|X,\phi}(y|x,\phi)^{\frac{-1+\alpha}{2}} \frac{\partial}{\partial U_{ij}}\left\{p_{Y|X,\phi}(y|x,\phi)\right\}$$

The case of $\alpha = -1$ is the traditional "log" version.

$$\Delta U^{(\alpha)}_{ij}(\phi) = p_{Y|X,\phi}(y|x,\phi)^{\frac{1+\alpha}{2}}\Delta U^{(-1)}_{ij}(\phi)$$

holds. Therefore, a large α emphasizes learning at a high probability density. The batch learning version uses $\prod_t p_{Y|X,\phi}(y(t)|x(t),\phi)^{\frac{1+\alpha}{2}}$.

V. VARIOUS STATISTICS MEASURES AND EXPECTATION LEARNING FOR THE WEM

5.1 Statistics Measures and Their α-Versions

There can be many α-information measure corresponding to the Fisher's information measure. We list up the following two versions for $M^{(\alpha)}$.

1. Exponential expectation: $-E_{\exp(L)}\left[\frac{\partial^2 L^{(\alpha)}}{\partial\phi\partial\phi^T}\right]$
2. Plain expectation: $-E\left[p^{-\frac{1+\alpha}{2}}\left(\frac{\partial^2 L^{(\alpha)}}{\partial\phi\partial\phi^T}\right)\right]$

Related to the second $M^{(\alpha)}(\phi)$ is the Cramér-Rao's bound for parameter estimation:

$$V(\hat{\phi} - \phi) \geq 1/V(\partial\ell/\partial\phi) = 1/M^{(-1)}(\phi).$$

Here, V is the variance, and $\hat{\phi}$ is an estimate of ϕ statisfying $E[\hat{\phi}] = \phi$. This Cramér-Rao bound can be derived from the α-efficent score:

$$V(\hat{\phi} - \phi) \geq 1/V(p^{-(1+\alpha)/2}\partial L^{(\alpha)}/\partial\phi).$$

The righthand side is reduced to

$$1/V(p^{-(1+\alpha)/2}\partial L^{(\alpha)}/\partial\phi) = 1/V(\partial\ell/\partial\phi).$$

Therefore, one obtains

$$m \stackrel{\text{def}}{=} M^{(\alpha)}(\phi)/M^{(-1)}(\phi) = (1-\alpha)/2.$$

Thus, this number m reflects the speed that the learning system acquires knowlegde from the training. This m can be called the aptitude number.

5.2 Newton-Raphson Learning and α-Information Measure

Let the τ-th iteration value of the extended logarithm of p_τ be

$$L^{(\alpha)}_\tau(\phi) = 2/(1+\alpha)\{p_\tau^{\frac{1+\alpha}{2}} - 1\}.$$

An optimizaton method using a Hessian matrix is as follows.

$$\phi_{\tau+1} = \phi_\tau - \left[E\left\{\partial^2 L^{(\alpha)}_\tau/\partial\phi\partial\phi^T\right\}\right]^{-1}\partial L^{(\alpha)}_\tau/\partial\phi.$$

There are many variants.

VI. SYSTOLIC AND MONITORING WEM

The WEM can be used as a building block. This idea comes from the fact that the WEM (EM) is too monolithic to model complex systems such as brains. Fig. 6.1 illustrates such an example. Each block can bifurcate. One branch can be regarded as a monitor.

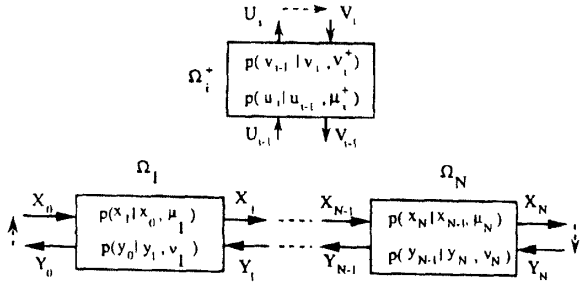

Fig. 6.1 A systolic layer of the WEM as an example.

VII. CONCLUDING REMARKS

In this paper, the wighted EM algorithm (WEM) was presented first. Then, extended versions of the Fisher measure and scoring, and Cramér-Rao bound were used to discuss computational aspects. The block monitoring idea presented in Section 6 will create many profitable structures.

REFERENCES

[REN60] A. Rényi, Proc. 4th Berkeley Symp., **1**, 547-561, 1960.

[HAV67] J.H. Havrda and F. Chavat, Kybernetica, **3**, 30-35, 1967.

[CSI72] I. Csiszár, Period. Math. Hungarica, **2**, 191-213, 1972.

[DEM78] A.P. Dempster, N.M. Laird and D.B. Rubin, J. R. Stat., B, **39**, 1-38, 1978.

[AMA93] S. Amari and H. Nagaoka, ***Methods of Information Geometry*** (in Japanese), Iwanami, 1993.

[JOR94] M.J. Jordan and R.A. Jacobs, Neural Computation, **6**, 181-214, 1994.