Adaptive estimation of parameters using partial information of desired outputs

Joby Joseph and K V S Hari
ECE Department, IISc, Bangalore, 560012
India

A general framework for forming an adaptive algorithm for problems where only partial information about the desired output is available, is proposed. Based on preliminary analysis it can be shown that this framework can be used to efficiently choose deep, narrow minima when there are many local minima. For problems like separation of instantaneous mixtures (Independent Component Analysis, ICA) and separation of convolutive mixtures when cast in the proposed framework is shown to give the same efficient algorithms as those available in the literature.

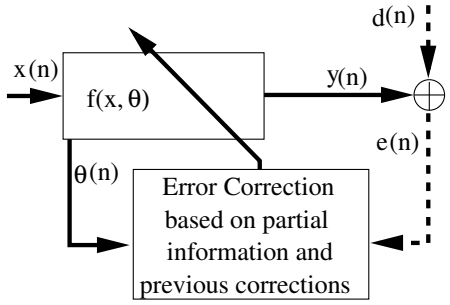

Figure 1: Flow graph of adaptive algorithms for estimation of parameters $\boldsymbol{\theta}$of the system $\mathbf{F}(\mathbf{x}(n), \boldsymbol{\theta})$

1 Introduction

Consider the flow-graph of adaptive algorithms for parameters shown in figure(1). Given realizations[1] $\mathbf{x}(n) \in \mathcal{X}$, a system model $\mathbf{y}(n) = \mathbf{f}(\mathbf{x}(n), \boldsymbol{\theta}(n))$ where $\mathbf{f}(\mathbf{x}(n), \boldsymbol{\theta}(n))$ corresponds to bounded input bounded output transformation parameterized by $\boldsymbol{\theta}(n) \in \Theta$, and desired response $\mathbf{d}(n) \in \mathcal{D}$**,** we are required to estimate the desired parameters $\boldsymbol{\theta}(n) = [\theta_1(n), \theta_2(n), \ldots \theta_M(n)]^T$ adaptively. The standard approach is to propagate the error $\mathbf{e}(n)$ back to the parameters $\boldsymbol{\theta}$(n) and correct them at a certain adaptation rate so as to decrease the instantaneous error. In this paper, we present a generalization, where the proposed adaptive algorithm can be used when there is partial information about $\mathbf{d}(n)$. These problems can also be classified as semiparametric problems [1, 2, 3] because the incompleteness of information about desired outputs can be looked upon as due to the effect of nuisance parameters. It is to be noted that additive noise is not considered as a nuisance parameter. It is the systematic errors in $\mathbf{d}(n)$ that is characterized by the set of nuisance parameters, φ[3]. Then we show that problems like ICA, separation of convolutive mixtures can be modeled using this paradigm. Further, this approach leads to the same set of efficient adaptive algorithms available in literature.

[1] **Notation:** Lower case bold face represents a vector. Upper case bold face represents a matrix. $\| \cdot \|$ denotes norm of a vector. $E_{\mathcal{X}}\{g(\mathbf{x})\} = \int_{\mathcal{X}} g(\mathbf{x}) r(\mathbf{x}) d\mathbf{x}$ denotes expectation over the set $\mathcal{X}$.

2 The adaptive algorithm framework

In the normal cases of adaptive algorithms where $\mathbf{d}(n)$ is known, except for the additive noise, $\boldsymbol{\theta}(n)$ is adapted to decrease the magnitude of the instantaneous error $\mathbf{e}(n) = \mathbf{d}(n) - \mathbf{y}(n)$. This error measure may be any positive function $\mathcal{E}(\mathbf{x}, \boldsymbol{\theta})$, of $\mathbf{e}(n)$, $\mathcal{E}(\mathbf{x}, \boldsymbol{\theta}) = \| \mathbf{e}(n) \|$ being an example suitable for some problems. Here it is assumed that the error $\mathcal{E}(\mathbf{x}, \boldsymbol{\theta})$ is a smooth function of $\boldsymbol{\theta}$. The correction term for elements of $\boldsymbol{\theta}(n)$ is calculated by propagating the error to the parameters $\boldsymbol{\theta}(n)$ of $\mathbf{f}(\mathbf{x}(n), \boldsymbol{\theta}(n))$ [4] as.

$$\boldsymbol{\theta}(n+1) = \boldsymbol{\theta}(n) - \mu(n) \frac{\partial \mathcal{E}}{\partial \boldsymbol{\theta}} \mid_{\mathbf{x}(n), \theta(n)} \tag{1}$$

Here $\mu(n)$ is the adaptation rate, $\frac{\partial \mathcal{E}}{\partial \boldsymbol{\theta}}$ is $[\frac{\partial \mathcal{E}}{\partial \theta_1} \ \frac{\partial \mathcal{E}}{\partial \theta_2} \ \dots \frac{\partial \mathcal{E}}{\partial \theta_N}]^T$. Equation (1) minimizes an approximation to the cost-function $\mathbf{C}_{\mathcal{X}}(\boldsymbol{\theta}) = E_{\mathcal{X}}\{\mathcal{E}(\mathbf{x}|\boldsymbol{\theta})\}$ with respect to $\boldsymbol{\theta}$. Given $\boldsymbol{\theta}$, the function $\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x})$ characterizes the errors that can occur for various $\mathbf{x}$. At the optimum $\boldsymbol{\theta} = \boldsymbol{\theta}_{opt}$, $\mathbf{C}(\boldsymbol{\theta}) = 0$ and $\mathbf{J} = cov(\mathcal{E}(\mathbf{x}|\boldsymbol{\theta}))$ has finite eigenvalues.

Consider the case where $\mathbf{d}(n)$ is not available, for then equation(1) cannot be used. In some cases, partial information about $\mathbf{d}(n)$ is available, like, for example, $\mathbf{d}(n) \in \mathcal{R}^m$, $\mathbf{d}(n)$ is known for some n and unknown otherwise, elements of $\mathbf{d}(n)$ are independent, elements of $\mathbf{d}(n)$ are distributed as some $r(.)$. Using such partial information, $\frac{\partial \mathcal{E}}{\partial \boldsymbol{\theta}} \mid_{\mathbf{x}(n), \theta(n)}$ can be approximated. Define the set $\mathcal{X}_e(\boldsymbol{\theta}) = \{\mathbf{x} : \mathbf{y}$ *such that is possibly correct*$\}$. Let $\mathcal{X}'_e(\boldsymbol{\theta})$ be the complement of the set of $\mathcal{X}_e(\boldsymbol{\theta})$, ie. set of $\mathbf{x}$ such that $\mathbf{y}$ is definitely in error. If adaptation framework in (1) is used then on convergence it would have minimized an approximation to the cost $\mathbf{C}_{\mathcal{X}'_e}(\boldsymbol{\theta}) = E_{\mathcal{X}'_e}\{\mathcal{E}(\mathbf{x}|\boldsymbol{\theta})\}$ rather than $\mathbf{C}_{\mathcal{X}}(\boldsymbol{\theta})$. This is because equation (1) prevents the freedom to explore of the elements of the parameter set Θ around the converged value after convergence. To facilitate this exploration we can add a $\Delta\boldsymbol{\theta}$ factor to equation (1). Because $\mathbf{C}_{\mathcal{X}}(\boldsymbol{\theta})$ is assumed to be smooth it is more probable that average direction of the corrections so far, $\boldsymbol{\theta}$, represents the right direction for this term $\Delta\boldsymbol{\theta}$. Therefore $\Delta\boldsymbol{\theta} = \alpha\Delta\boldsymbol{\theta}$ the average correction direction at the n^{th} instant. Thus under availability of partial information the stochastic gradient descent adaptive scheme is modified to *do the corrections when ever and where ever information is available about* $\mathbf{d}(n)$ *or else assume that the previous corrections were in the right direction and do a weighted correction in those directions.* This modified adaptive framework is as follows:

$$\boldsymbol{\theta}(n+1) = \boldsymbol{\theta}(n) - \mu(n) \frac{\partial \mathcal{E}}{\partial \boldsymbol{\theta}} \mid_{\mathbf{x}(n), \boldsymbol{\theta}(n)} + \alpha(n)\boldsymbol{\theta}(n) \tag{2}$$

Here $\mu(n)$ is the adaptation rate[2] of the instantaneous error, and $\alpha(n)$ is that of the term due to all previous adaptations(ie. memory). Let $\beta(n) = \mu(n)/\alpha(n)$. Then equation(2) can be written as

$$\boldsymbol{\theta}(n+1) = \boldsymbol{\theta}(n) - \beta(n)\alpha(n) \frac{\partial \mathcal{E}}{\partial \boldsymbol{\theta}} \mid_{\mathbf{x}(n), \boldsymbol{\theta}(n)} + \alpha(n)\boldsymbol{\theta}(n) \tag{3}$$

Remarks:

- We can interpret $\alpha(n)\boldsymbol{\theta}(n)$ as the memory term and $-\alpha(n)\beta(n)\frac{\partial \mathcal{E}}{\partial \boldsymbol{\theta}} \mid_{\mathbf{x}(n), \boldsymbol{\theta}(n)}$ as the instantaneous error correction term.
- Denoting the j^{th} element of $\mathcal{E}$, as $\mathcal{E}_j$, we propose that $\frac{\partial \mathcal{E}_j}{\partial \theta_i}$ can be set to zero when there is no information available to compute it.
- For convergence in mean it is required that $E\{\boldsymbol{\theta}(n+1) - \boldsymbol{\theta}(n)\} = 0$. For this to be satisfied

  $$\lim_{n \to \infty} E\{-\alpha(n)\beta(n) \frac{\partial \mathcal{E}}{\partial \boldsymbol{\theta}} \mid_{\mathbf{x}(n), \boldsymbol{\theta}(n)} + \alpha(n)\boldsymbol{\theta}(n)\} = 0$$

- For stability, variance of the correction term, $J(n) = var\{-\alpha(n)\beta(n)\frac{\partial \mathcal{E}}{\partial \boldsymbol{\theta}} \mid_{\mathbf{x}(n), \boldsymbol{\theta}(n)} + \alpha(n)\boldsymbol{\theta}(n)\}$ should converge to a constant finite value $J(\infty)$. Thus the correction term

[2] If y is a scalar and $\mathbf{x}$ is a vector then $\frac{\partial y}{\partial \mathbf{x}} = [\frac{\partial y}{\partial x_1} \ \frac{\partial y}{\partial x_2} \ \dots \frac{\partial y}{\partial x_N}]^T$. If $\mathbf{y}$ and $\mathbf{x}$ are vectors then ijth element of the matrix $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ is $\frac{\partial y_j}{\partial x_i}$. $E_{\mathcal{X}}\{g(\mathbf{x})\} = \int_{\mathcal{X}} g(\mathbf{x}) r(\mathbf{x}) d\mathbf{x}$ denotes expectation over the set $\mathcal{X}$.

can be interpreted as *an estimating function*[5].

- For the case where full information about $\mathbf{d}(n)$is available, to make the algorithm in equation (3) reduce to the conventional stochastic gradient algorithm we have to make $\alpha(n) \rightarrow 0$ and $\beta(n) \rightarrow \infty$ so that $\beta(n)\alpha(n)$ is finite. $\beta(n)$ has to be adjusted to make the algorithm stable and can be even data depended, as in the case of learning parameter of Normalized Least Mean Squared (NLMS) algorithm. It is not possible to say about the choice of $\alpha(n)$ and $\beta(n)$ nor the stability of the adaptation algorithm without knowing structure of $\mathbf{f}(\mathbf{x}(n), \boldsymbol{\theta})$ which is specific to an application.

- You may notice that this is similar to but not identical to the momentum method proposed by Rumelhart et al.[6], see equation(4), *in a context other than with only partial information.*

$$\boldsymbol{\theta}(n+1) = \boldsymbol{\theta}(n) - \beta(n)\alpha(n)\frac{\partial \mathcal{E}}{\partial \boldsymbol{\theta}} \mid_{\mathbf{x}(n),\boldsymbol{\theta}(n)} + \alpha(n)(\boldsymbol{\theta}(n) - \boldsymbol{\theta}(n-1)) \tag{4}$$

In equation(2) if we replace $\alpha(n)\boldsymbol{\theta}(n)$ with $\alpha(n)(\boldsymbol{\theta}(n) - \boldsymbol{\theta}(0))$ we can compare it with the momentum method and see the difference. While in momentum method only the immediate past adaptation is used, proposed method uses the entire past history of θ to do the present adaptation and thus can capture the average direction of descend when only partial error information is available for correction.

3 Properties of the proposed framework

Convergence behavior

Substituting recursively for $\boldsymbol{\theta}(n)$ in equation (3), and assuming $\alpha(n) = \alpha_0$, $\beta(n) = \beta_0$ constants,

$$\boldsymbol{\theta}(n) = (1+\alpha_0)^n \boldsymbol{\theta}(0) - \sum_{i=0}^{n-1} \alpha_0\beta_0 \frac{\partial \mathcal{E}}{\partial \boldsymbol{\theta}} \mid_{\mathbf{x}(n),\boldsymbol{\theta}(n)} \tag{5}$$

$$\delta\boldsymbol{\theta}(n) = \boldsymbol{\theta}(n) - \boldsymbol{\theta}(n-1) = (1+\alpha_0)^{n-1}\alpha_0\boldsymbol{\theta}(0) - \sum_{i=0}^{n-2} \alpha_0^2 (1+\alpha_0)^i \beta_0 \frac{\partial \mathcal{E}}{\partial \boldsymbol{\theta}} \mid_{\mathbf{x}(n-i),\boldsymbol{\theta}(n-i)} - \alpha_0\beta_0 \frac{\partial \mathcal{E}}{\partial \boldsymbol{\theta}} \mid_{\mathbf{x}(n),\boldsymbol{\theta}(n)} \tag{6}$$

Thus the adaptation term at any time n is having three terms.

1. $-\alpha_0\beta_0 \frac{\partial \mathcal{E}}{\partial \boldsymbol{\theta}} \mid_{\mathbf{x}(n),\boldsymbol{\theta}(n)}$ corresponds to the instantaneous adaptation term.

2. $-\sum_{i=0}^{n-2} \alpha_0^2 (1+\alpha_0)^i \beta_0 \frac{\partial \mathcal{E}}{\partial \boldsymbol{\theta}} \mid_{\mathbf{x}(n-i),\boldsymbol{\theta}(n-i)}$ corresponds to the memory term. It is the global direction inferred from all the past corrections based on the errors available. For all the past corrections to be weighted equally, $\alpha_0 \rightarrow 0$ and simultaneously $\beta_0 \rightarrow \infty$ for the learning rate $\alpha_0\beta_0$ to be finite.

3. $(1+\alpha_0)^{n-1}\alpha_0\boldsymbol{\theta}(0)$ corresponds to the initialization. This would produce a biased estimate. For this factor to vanish $\theta(0)$ has to be set to 0. But some problems would require that $\theta(0)$ be nonzero for the iteration to start moving.

Behavior near the optimum, $\boldsymbol{\theta}_{opt}$

Assume $\mathcal{E}$ to be quadratic in $\mathbf{e}$ near $\boldsymbol{\theta}_{opt}$, which implies that on convergence to $\boldsymbol{\theta}_{conv}$,

$$\begin{aligned} \frac{\partial \mathcal{E}}{\partial \mathbf{e}}|_{\mathbf{x}(n),\boldsymbol{\theta}_{conv}} &= \kappa\mathbf{e}(n) = \kappa\left(\mathbf{f}(\mathbf{x}(n),\boldsymbol{\theta}_{opt}) - \mathbf{f}(\mathbf{x}(n),\boldsymbol{\theta}_{conv})\right) \\ &= d\mathbf{f} = \left(\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}}\right)^T d\boldsymbol{\theta} = \left(\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}}\right)^T (\boldsymbol{\theta}_{opt} - \boldsymbol{\theta}_{conv}) \end{aligned} \tag{7}$$

where κ is some constant. This implies

$$\begin{aligned} \frac{\partial \mathcal{E}}{\partial \boldsymbol{\theta}}|_{\mathbf{x}(n),\boldsymbol{\theta}_{conv}} &= \frac{\partial \mathbf{e}}{\partial \boldsymbol{\theta}}\frac{\partial \mathcal{E}}{\partial \mathbf{e}} = \left(\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}}\right)^T \frac{\partial \mathcal{E}}{\partial \mathbf{e}} = \left(\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}}\right)^T d\boldsymbol{\theta} \\ &= \hat{\mathbf{R}}_{\mathbf{f}}(\boldsymbol{\theta}(n),\mathbf{x}(n))\,(\boldsymbol{\theta}_{conv} - \boldsymbol{\theta}_{opt}) \end{aligned} \tag{8}$$

where $\hat{\mathbf{R}}_{\mathbf{f}}(\boldsymbol{\theta}(n),\mathbf{x}(n)) = \left(\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}}\right)^T$. Therefore when $\boldsymbol{\theta}(n)$is in the neighborhood of $\boldsymbol{\theta}_{opt}$,

$$\boldsymbol{\theta}_{conv} = \boldsymbol{\theta}_{conv} + \alpha(n)\boldsymbol{\theta}_{conv} - \beta(n)\alpha(n)\hat{\mathbf{R}}_{\mathbf{f}}(\boldsymbol{\theta}_{conv},\mathbf{x}(n))(\boldsymbol{\theta}_{conv} - \boldsymbol{\theta})_{opt}) \tag{9}$$

On convergence $\boldsymbol{\theta}(n+1) = \boldsymbol{\theta}(n) = \boldsymbol{\theta}_{conv}$ which is possible if

$$\begin{aligned} &E\left\{\alpha(n)\boldsymbol{\theta}_{conv} + \boldsymbol{\theta}_{conv}\beta(n)\alpha(n)\hat{\mathbf{R}}_{\mathbf{f}}(\boldsymbol{\theta}_{conv},\mathbf{x}(n))\right\} \\ &\quad = E\left\{\boldsymbol{\theta}_{opt}\beta(n)\alpha(n)\hat{\mathbf{R}}_{\mathbf{f}}(\boldsymbol{\theta}_{conv},\mathbf{x}(n))\right\} \end{aligned} \tag{10}$$

Therefore on convergence

$$\begin{aligned}\boldsymbol{\theta}_{conv} &= \boldsymbol{\theta}_{opt} E\left\{\beta(n)\hat{\mathbf{R}}_{\mathbf{f}}(\boldsymbol{\theta}_{conv}, \mathbf{x}(n))\right\} \\ &\quad \left(\mathbf{I} + E\left\{\beta(n)\hat{\mathbf{R}}_{\mathbf{f}}(\boldsymbol{\theta}_{conv}, \mathbf{x}(n))\right\}\right)^{-1} \\ &= \boldsymbol{\theta}_{opt}\left(E\left\{\beta(n)\hat{\mathbf{R}}_{\mathbf{f}}(\boldsymbol{\theta}_{conv}, \mathbf{x}(n))\right\}^{-1} + \mathbf{I}\right)^{-1} \quad (11)\end{aligned}$$

The adaptive framework make $\boldsymbol{\theta}$ converge to a value near to $\boldsymbol{\theta}_{opt}$ if $E\left\{\beta(n)\hat{\mathbf{R}}_{\mathbf{f}}(\boldsymbol{\theta}(n), \mathbf{x}(n))\right\}$ is full rank and has large norm. *This implies that this approach can be used to choose the minima with such a property, ie. deep narrow minima, from among all the local minima. Also we can infer that for the case where the proposed method reverts to the standard gradient descent, which happens when $\beta \to \infty$ and $\alpha \to 0$, for $\boldsymbol{\theta}$ to converge to $\boldsymbol{\theta}_{opt}$ we need $\hat{\mathbf{R}}_{\mathbf{f}}(\boldsymbol{\theta}_{conv}, \mathbf{x}(n))$ to be full rank* . The proposed framework will be studied further with the help of a few applications in the following sections.

4 Application of the algorithm to various problems

4.1 Blind separation of instantaneous linear mixtures

Here the measurements available are $\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n)$, $\mathbf{x}(n) = [x_1(n)\ x_2(n)\ \ldots\ x_M(n)]^T$, $\mathbf{s}(n) = [s_1(n)\ s_2(n)\ \ldots\ s_M(n)]^T$ whose elements s_i are distributed with unknown non-gaussian density function $r(s)$ are independent [7, 8] and $\mathbf{A}$ is an $M \times M$ matrix whose elements are A_{ij}. The problem is to estimate $\mathbf{A}$ such that $\mathbf{y}(n) = \mathbf{A}^{-1}\mathbf{x}(n) = \hat{\mathbf{s}}(n)$, an estimate of $\mathbf{d}(n) = \mathbf{s}(n)$. This problem is also known as Independent Component Analysis (ICA). The information we know about desired output $\mathbf{d}(n)$ is that its elements are iid non-gaussian random variables. Since nothing else is known about the random variables it is proposed to assume as distributed uniformly with in the interval $-\gamma$ to γ. Using equation (2) it is possible to write down the adaptation algorithm for this case.

$$\mathbf{W}(n+1) = \mathbf{W}(n) + \alpha\mathbf{W}(n) + \beta\alpha\frac{\partial\mathcal{E}}{\partial\mathbf{W}}\mid_{\mathbf{x}(n),\mathbf{W}(n)} \quad (12)$$

where $\alpha$and β have been chosen to be constants. The fact that s_i are independent implies that corrections to each parameter affecting s_i can be done independently. Define $e_i(n)$ as

$$e_i(n) = \begin{cases} -\hat{s}_i(n)/|\hat{s}_i(n)| & \text{if} \quad |\hat{s}_i(n)| > \gamma \\ 0 \quad \text{otherwise} \end{cases}$$

Form the vector $\mathbf{e}(n) = [e_1(n)\ e_2(n) \ldots e_M(n)]^T$. The gradient $\partial y_i/\partial W_{ij} = x_j$, and therefore $\partial\mathcal{E}/\partial W_{ij} = e_i\partial y_i/\partial W_{ij} = e_i x_j$, assuming quadratic $\mathcal{E}$. Therefore $\frac{\partial\mathcal{E}}{\partial\mathbf{W}}\mid_{\mathbf{x}(n),\mathbf{W}(n)} = \mathbf{e}(n)\mathbf{x}^T(n)$. Using natural gradient to orthogonalize the corrections[9], the correction factor becomes $\mathbf{e}(n)\mathbf{x}^T(n)\mathbf{W}^T(n)\mathbf{W}(n) = \mathbf{e}(n)\mathbf{y}^T(n)\mathbf{W}(n)$. Therefore equation (12) becomes

$$\begin{aligned}\mathbf{W}(n+1) &= \mathbf{W}(n) + \alpha\mathbf{W}(n) + \beta\alpha\mathbf{e}(n)\mathbf{y}^T(n)\mathbf{W}(n) \\ &= \mathbf{W}(n) + \alpha(\mathbf{I} + \beta\mathbf{e}(n)\mathbf{y}^T(n))\mathbf{W}(n)\end{aligned}$$

which is the adaptation algorithm arrive at in [7, 10, 11, 12, 13]. The work [13] shows the above error function which is the replacement of nonlinearity in [13] is stable for any continuous, symmetric, differentiable, non-gaussian $r(.)$.

Remark: This example demonstrates that using equation (3) we can obtain the adaptation rule which was originally derived starting a maximum likelihood(ML) cost function[7, 10, 11, 12, 13].

4.2 Blind separation of convolutive mixtures

Here the measurements available are

$$\mathbf{x}(n) = \sum_{p=0}^{P}\mathbf{B}(p)\mathbf{s}(n-p) + \sum_{q=1}^{Q}\mathbf{A}(p)\mathbf{x}(n-p) = \mathbf{B}(z)\mathbf{A}^{-1}(z)\mathbf{s}(n)$$

where $\mathbf{x}(n)$ and $\mathbf{s}(n)$ are as defined previously and $\mathbf{A}(z)$ and $\mathbf{B}(z)$ are matrix polynomials. It is required to find $\mathbf{W}(z)$ such that $\mathbf{y}(n) = \mathbf{W}(z)\mathbf{x}(n) = \mathbf{s}(n)$ have independent elements. The adaptation algorithm would be in the form

$$\mathbf{W}_p(n+1) = \mathbf{W}_p(n) + \alpha\mathbf{W}_p(n) + \beta\alpha\frac{\partial\mathcal{E}}{\partial\mathbf{W}_p}\mid_{\mathbf{x}(n),\mathbf{W}_p(n)} \quad (13)$$

where $\alpha$and β are fixed constants. Since the information available about the source density is the same as in the instantaneous mixture case

the error function is the same. Following the same line of argument as in the instantaneous mixture case, $\frac{\partial \mathcal{E}}{\partial \mathbf{W}_p} = \mathbf{e}(n)\mathbf{x}^T(n-p)$. Using the fact that the elements of $\mathbf{s}(n)$ are independent and by applying natural gradient to the error correction

$$\mathbf{W}_p(n+1) = \begin{cases} \mathbf{W}_p(n) + \alpha \mathbf{W}_p(n) + \\ \beta\alpha \mathbf{e}(n)\mathbf{x}^T(n-p)\mathbf{W}^T(z^{-1})\mathbf{W}(z) \end{cases}$$

Remark: Using equation (3) we could arrive at the same algorithm derived starting from a ML cost function in [2, 10, 11].

5 Discussion

In the above examples application of the general principle (2) to problems with partial information on desired output led to identical algorithms as derived using various approaches in the existing literature in two cases of source separation. This approach gives a clue as to how possibly the available information about the nuisance parameter can be used to form adaptive estimators. But once having formulated the algorithm it can be explained in various ways following other works in the literature. For example in the case of ICA because the error function $e()$ is nonlinear it can be expanded in terms of powers of s_i and hence the adaptation algorithm minimizes the cross cumulants of the recovered signals. Another way to interpret is that the algorithm minimizes marginal mismatch since the information about $r(s_i)$ used is that about the marginal densities. Since adaptation term is formed assuming independence of sources we can say that it corrects for deviation from independence.

6 Conclusion

The current work shows a practical framework for forming adaptive algorithms for problems where some partial information is available about the desired output. In two cases where in the literature algorithms have been formed using different approaches like estimating function and maximum likelihood method, the adaptive algorithms constructed using the principle in this paper turns out to be the same. The method gives a way of using the known information efficiently unlike in other cases where this is left to be in the form of an unknown nonlinearity. This approach can be used to choose the minima with such a property, ie. deep narrow minima, from among all the local minima. We are working on finding, for what class of $\mathbf{f}(\mathbf{x}, \boldsymbol{\theta})$ is equation (3) applicable and how to choose α and β.

References

[1] Shun-Ichi Amari and Jean-Francois Cardoso, "Blind source separation — semiparametric statistical approach," *IEEE Transaction on SP*, pp. –, 1997.

[2] L.-Q. Zhang, S. Amari, and Cichocki, *Advances in Neural Information Processing 12*, MIT Press, 2000.

[3] Shun ichi Amari and Motoaki Kawanabe, "Information geometry of estimating functions in semiparametric statistical models," *Bernouli*, vol. 2, no. 3, pp. –.

[4] Simon Haykins, *Adaptive Filter Theory*, Prentice-Hall International, Inc, first edition, 1996.

[5] S. Amari and M. Kawanabe, "Estimating functions in semiparametric statistical model," *Selected Proceedings of the Symposium on Estimating Functions, IMS Lecture Notes–Monograph Series*, vol. 32, pp. 65–81, 1997.

[6] D. Rumelhart, G. Hinton, and R. Williams, *Parallel Distributed Processing*, MIT Press, 1986.

[7] Jean-Francois Cardoso, C. N. R. S., and E. N. S. T., "Blind signal separation: Statistical principles," *Proceeedings of IEEE*, vol. 86, no. 10, pp. 2009–2025, October 1998.

[8] Pierre Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, pp. 287–314, 1994.

[9] Shun ichi Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.

[10] Shun ichi Amari, Scott C. Douglas, Andrzej Cichocki, and Howar H Yang, "Novel online adaptive learning algorithm for blind deconvoution using the natural gradient approach," pp. –.

[11] Scott C. Douglas, Andrzej, and Shun ichi Amari, "Multichannel blind separation and deconvolution of sources with arbitrary distributions," *IEEE workshop on Neural Networks or Signal Processing, Almelia Island Plantation*, pp. 436–445, Sept 1997.

[12] Pierre Common, Chritian Jutten, and Jeanny Herault, "Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, pp. 1–10, 1991.

[13] Heinz Mathis an Scott C. Douglas, "On the existance of universal nonlinearities for blind source separation," *IEEE Transaction on SP*, vol. 50, no. 5, pp. 1007–1016, May 2002.