An on-line Fisher discriminant

Manuel Ortega-Moral, Vanessa Gómez-Verdejo,
Jerónimo Arenas-García and Aníbal R. Figueiras-Vidal
Department of Signal Theory and Communications

Universidad Carlos III de Madrid
Avda. de la Universidad 30, 28911 Leganés (Madrid) SPAIN.
{*ortegam,vanessa,jarenas,arfv*}*@tsc.uc3m.es*

Abstract. Many applications in signal processing need an adaptive algorithm. Adaptive schemes are useful when the statistics of the problem are unknown or when facing varying environments. Nonetheless, many of these applications deal with classification tasks, and most algorithms are not specifically thought to tackle these kinds of problems. Whereas Fisher's criterion aimed to find the most adequate direction to discriminate classes in a stationary setting, the newly proposed On-line Fisher Linear Discriminant (OFLD) is able to adaptively update its parameters maintaining its discrimination goal. The algorithm has been tested in an equalization problem for several conditions.

1 Introduction

Adaptive filtering is a fundamental tool in many signal processing applications, such as system modeling, noise and echo cancellation, or channel equalization, among many others [4]. Adaptive schemes are suitable in scenarios where the statistics of the filtering problem are not completely known, or, specially, when these statistics are time-varying, given the fact that the proposed solution is refined when more data are available, forgetting also the oldest patterns, that do not longer reflect the current characteristics of the filtering problem.

The adaptation of an adaptive filter is carried out with the objective of minimizing a cost function, usually the quadratic difference between the output of the filter and a reference signal (resulting, for instance, in the well-kwown Least Mean Squares (LMS) [5] or Recursive Least Squares (RLS) methods [2]). Nonetheless, in many applications the real aim of the filter is to discriminate between a finite number of classes (i.e., to predict to which class a new sample belongs to), and in these cases minimizing the quadratic error is only an indirect manner to achieve the desired goal.

Fisher Linear Discriminant (FLD) [1] lies on a cost function which is a measure of class separability. However, up to this time, it has only been used to design classifiers in stationary settings. In this paper, we propose to adapt online the required statistics to provide FLD with adaptive capabilities, obtaining the so-called On-line FLD (OFLD). The resulting scheme is specially suitable for classification tasks where adaptability is a mandatory issue.

The rest of the paper is organized as follows: in the next section we briefly review FLD. Section 3 is dedicated to introduce our approach, explaining how classification accuracy and adaptability can be put together by means of OFLD.

Section 4 is devoted to some experiments in an adaptive channel equalization environment, that show the interest and potential of the new adaptive method. Finally, conclusions and further work are presented in section 5.

2 Fisher's Linear Discriminant

In a binary class problem, FLD linearly projets d-dimesional data onto a one-dimensional space, so that an input vector $\mathbf{x}$ is projected onto

$$y = \mathbf{w}^T \mathbf{x} \tag{1}$$

where $\mathbf{w}$ are the projection weights. FLD seeks an optimal direction for class separability by maximizing a function which represents the difference between the projected class means, normalized by a measure of the within-class scatter along the direction of $\mathbf{w}$. This is known as the Fisher's criterion [1]:

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} \tag{2}$$

m_j and s_j, with $j = 1, 2$, being the mean and the within-class covariance of projected class C_j, respectively. It can be easily shown that a more appropriate expression for Fisher's criterion can be derived by defining the between-class covariance matrix

$$\mathbf{S}_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T \tag{3}$$

where

$$\mathbf{m}_j = \frac{1}{N_j} \sum_{\mathbf{x} \in C_j} \mathbf{x} \tag{4}$$

N_j being the number of samples belonging to class C_j. Furthermore, total within-class covariance matrix can be defined as

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2 = \sum_{\mathbf{x} \in C_1} (\mathbf{x} - \mathbf{m}_1)(\mathbf{x} - \mathbf{m}_1)^T + \sum_{\mathbf{x} \in C_2} (\mathbf{x} - \mathbf{m}_2)(\mathbf{x} - \mathbf{m}_2)^T \tag{5}$$

This way, (2) can be rewritten as

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} \tag{6}$$

Then, by equating to 0 the gradient of (6) with respect to $\mathbf{w}$, we can conclude that this expression is maximized with the FLD

$$\mathbf{w} \propto \mathbf{S}_W^{-1} [\mathbf{m}_2 - \mathbf{m}_1] \tag{7}$$

where, for classification purposes, scalar factors have been dropped because we are only interested in the direction of $\mathbf{w}$ and not in its norm.

3 On-line Fisher Discrimination

Aforementioned FLD formulation (7) is non-adaptive, being unsuitable for problems like channel equalization or pattern recognition in varying environments. In this section we propose an adaptive version of FLD, that lies on an on-line updating of the mean of the within-class covariance matrix.

To estimate adaptively each mean we will calculate the exponential average of the input samples $\mathbf{x}_j(n)$ belonging to class j:

$$\widehat{\mathbf{m}}_j(n) = \frac{1-\lambda_m}{1-\lambda_m^n} \sum_{i=1}^{n} \lambda_m^{n-i} \mathbf{x}_j(i) \tag{8}$$

where λ_m is a weighting factor. The quotient that premultiplies the summation serves to cancel the effect in the total sum of introducting factor λ_m. Taking λ_m less than 1 and considering $n >> 1$ we can approximate

$$\begin{aligned} \widehat{\mathbf{m}}_j(n) &\approx (1-\lambda_m) \sum_{i=1}^{n} \lambda_m^{n-i} \mathbf{x}_j(i) \\ &= (1-\lambda_m) \sum_{i=1}^{n-1} \lambda_m^{n-i} \mathbf{x}_j(i) + (1-\lambda_m)\mathbf{x}_j(n) \\ &= \lambda_m \widehat{\mathbf{m}}_j(n-1) + (1-\lambda_m)\mathbf{x}_j(n) \end{aligned} \tag{9}$$

Similarly, each term of the total within-class covariance matrix (5) is also varying with time and can be calculated recursively whenever $\mathbf{x}(n)$ belongs to class j

$$\widehat{\mathbf{S}}_j(n) = \lambda_S \widehat{\mathbf{S}}_j(n-1) + [\mathbf{x}(n) - \widehat{\mathbf{m}}_j(n)][\mathbf{x}(n) - \widehat{\mathbf{m}}_j(n)]^T \tag{10}$$

where λ_S, slightly smaller than 1, is a weighting factor that ensures $\widehat{\mathbf{S}}_W(n)$ is corrected at every instant paying more attention to recent samples. So, the total within-class covariance matrix is simply given by

$$\widehat{\mathbf{S}}_W(n) = \widehat{\mathbf{S}}_1(n) + \widehat{\mathbf{S}}_2(n) \tag{11}$$

We could likewise update between-class covariance matrix

$$\widehat{\mathbf{S}}_B(n) = [\widehat{\mathbf{m}}_1(n) - \widehat{\mathbf{m}}_2(n)][\widehat{\mathbf{m}}_1(n) - \widehat{\mathbf{m}}_2(n)]^T \tag{12}$$

although this is only necessary to compute the Fisher's cost function but not the discriminant. Therefore, Fisher's criterion becomes time-dependent

$$J(\mathbf{w}, n) = \frac{\mathbf{w}(n)^T \widehat{\mathbf{S}}_B(n) \mathbf{w}(n)}{\mathbf{w}(n)^T \widehat{\mathbf{S}}_W(n) \mathbf{w}(n)} \tag{13}$$

As we actually want to maximize expression (13) with respect to $\mathbf{w}(n)$, equation (7) is easily extended to

$$\mathbf{w}(n) \propto \widehat{\mathbf{S}}_W^{-1}(n)[\widehat{\mathbf{m}}_2(n) - \widehat{\mathbf{m}}_1(n)] \tag{14}$$

which main computational difficulty is to calculate the inverse of $\widehat{\mathbf{S}}_W$. Luckily, it can be recursively updated by means of the matrix inversion lemma [1]. Let $\mathbf{P}(n) = \widehat{\mathbf{S}}_W^{-1}(n)$, then

$$\mathbf{P}(n) = \lambda_S^{-1}\mathbf{P}(n-1) - \lambda_S^{-2}\mathbf{g}(n)(\mathbf{x}(n) - \widehat{\mathbf{m}}_j(n))^T)\mathbf{P}(n-1) \tag{15}$$

where $\mathbf{x}(n)$ is the class j input at time n which mean is $\widehat{\mathbf{m}}_j(n)$ and $\mathbf{g}(n)$ is defined as a gain vector

$$\mathbf{g}(n) = \frac{\mathbf{P}(n-1)(\mathbf{x}(n) - \widehat{\mathbf{m}}_j(n))}{\lambda_S + (\mathbf{x}(n) - \widehat{\mathbf{m}}_j(n))^T\mathbf{P}(n-1)(\mathbf{x}(n) - \widehat{\mathbf{m}}_j(n))} \tag{16}$$

Supposing means are known or estimated by means of (9), the OFLD algorithm is shown in Table 1. It can be easily shown that the computational cost of the algorithm is $O(N^2)$, like RLS.

```
1.-  Initialization:
       $\widehat{\mathbf{m}}_1(0) = 0$, $\widehat{\mathbf{m}}_2(0) = 0$, $\mathbf{P}(0) = \delta^{-1}\mathbf{I}$, $\mathbf{w}_{Fisher}(0) = \mathbf{0}$
2.-  For $n = 1, 2, 3, \ldots$
       $\widehat{\mathbf{m}}_j(n) = \lambda_m\widehat{\mathbf{m}}_j(n-1) + (1-\lambda_m)\mathbf{x}(n)$; supossed $\mathbf{x}(n) \in C_j$
       $\mathbf{\Pi}(n) = \mathbf{P}(n-1)(\mathbf{x}(n) - \widehat{\mathbf{m}}_j)$
       $\mathbf{g}(n) = \frac{\mathbf{\Pi}(n)}{\lambda_S + (\mathbf{x}(n) - \widehat{\mathbf{m}}_j)^T\mathbf{\Pi}(n)}$
       $\mathbf{P}(n) = \lambda_S^{-1}(\mathbf{I} - \lambda_S^{-1}\mathbf{g}(n)(\mathbf{x}(n) - \widehat{\mathbf{m}}_j(n))^T)\mathbf{P}(n-1)$
       $\mathbf{w}_{Fisher}(n) = \mathbf{P}(n)(\widehat{\mathbf{m}}_1(n) - \widehat{\mathbf{m}}_2(n))$
```

Table 1: On-line Fisher Linear Discriminant Pseudocode

4 Experimental Results

Adaptive channel equalization usually has to cope with Intersymbol Interference (ISI) problems, which represents some of the worst obstacles for high speed communications. Equalizers are typically built using an adaptive filter and designed to approximately invert and track time-varying channel distortions [3]. On-line equalizers are updated sample by sample and need a training phase where a training sequence, known by both the transmitter and receiver, is used to design a filter which inverts the effects of the unknown channel. Then, during the decision oriented phase, the error between the equalizer's output and the symbol decision is generally used to adapt the filter, which is able to track changes in the channel as long as these occur at a sufficiently low speed. This is the case of LMS and RLS equalizers that feedback the quadratic error. However, OFLD equalizer works in a different manner, feedbacking its own symbol decision, which is used to decide which estimated mean needs to be updated at each iteration.

[1] $(A + BCD)^{-1} = A^{-1} - A{-}1B(C{-}1 + DA^{-1}B)^{-1}DA^{-1}$

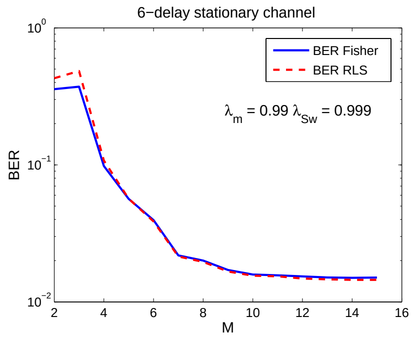

Fig. 1: Bit Error Rate versus equalizer tap length in a steady-state problem

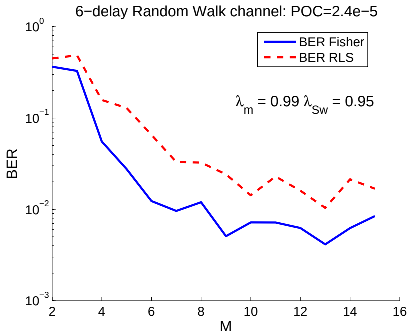

Fig. 2: Bit Error Rate versus equalizer tap length in a non-steady-state problem

Both algorithms have been compared in Figure 1 that presents the Bit Error Rate (BER) versus the receiver's tap length obtained by OFLD and RLS equalizers in a problem with a stationary channel $h_c = 0.2 + 0.5z^{-1} - 0.1z^{-2} + z^{-3} + 0.3z^{-4} + 0.1z^{-5}$, Gaussian noise and $SNR = 9$ dB. Parameters are $\lambda_m = 0.99$ and $\lambda_S = \lambda_{RLS} = 0.999$ while the training sequence has 500 samples and the decision oriented phase works with 10^4 samples. All results are averaged over 100 runs. As it can be seen, no relevant differences exist between both algorithms in this situation. However, when the channel is time-varying, performances can be rather different. For the time-varying problem we have chosen the Random Walk channel $\mathbf{h}_c(n) = \mathbf{h}_c(n-1) + \mathbf{q}(n)$ where the entries of $\mathbf{q}(n)$ are i.i.d. Gaussian values. The power of changes (POC) in the channel is measured by the trace of $E\{\mathbf{q}(n)\mathbf{q}(n)^T\}$, being $E\{\cdot\}$ the expectation operator. For the example in Figure 2 we have selected $POC = 2.4e-5$, $\lambda_m = 0.99$ and $\lambda_S = 0.95$. Note that in this case we have selected a lower value for λ_S as a consequence of the higher adaptation necessity to track the channel. As it can be seen in Figure 2, OFLD outperforms RLS with the difference being more significative for adaptive equalizers with a high number of taps.

5 Conclusions

An adaptive approach to Fisher's criterion has been presented. Its discriminative character makes it specially suitable for problems in which detection and classification tasks must be accomplished. Its capability to adapt to varying environments has been tested by means of an equalization setting, showing how, in the examples, the algorithm performance is less sensitive to gradient noise and can be better in tracking situations without increasing the computational effort. Future work will be addressed to better understand and explore the effects of the relation between free parameters and to test the algorithm in more complex situations and applications.

References

[1] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, New York, NY, 1995.

[2] R. L. Placket. Some theorems in least squares. *Biometrika*, (37):149, 1950.

[3] S. U. H. Qureshi. Adaptive equalization. In *Proceedings of the IEEE*, volume 73, September 1985.

[4] A. H. Sayed. *Fundamentals of Adaptive Filtering.* Wiley, New York, NY, 2003.

[5] B. Widrow and M. E. Hoff. Adaptive switching circuits. In *IRE Wescon Conv. Records*, pages 96–104, 1960.