CHANNEL EQUALIZATION AND THE BAYES POINT MACHINE

Edward Harrington, Jyrki Kivinen and Robert C. Williamson

Research School of Information Sciences and Engineering
The Australian National University
Canberra, ACT 0200

ABSTRACT

Equalizers trained with a large margin have an ability to better handle noise in unseen data and drift in the target solution. We present a method of approximating the Bayes optimal strategy which provides a large margin equalizer, the Bayes point equalizer. The method we use to estimate the Bayes point is to average N equalizers that are run on independently chosen subsets of the data. To better estimate the Bayes point we investigated two methods to create diversity amongst the N equalizers. We show experimentally that the Bayes point equalizer for appropriately large step sizes offers improvement on LMS and LMA in the presence of channel noise and training sequence errors. This allows for shorter training sequences albeit with higher computational requirements.

1. INTRODUCTION

A standard technique for correcting Inter-Symbol Interference (ISI) caused by the communications channel is to apply an equalizer at the receiver [6]. We consider the case of training the equalizer with a known sequence, where the received signal is corrupted by channel noise.

In stochastic gradient methods like Least Mean Squared (LMS) randomness in the equalizer weights $\mathbf{w}_t$ can make the estimate of the gradient based on a single sample prone to updating $\mathbf{w}_t$ in the wrong direction. By making the step size η small the error is reduced but the convergence is slow. Gardner [3] showed that averaging the gradients over B data points reduced the effect of the noise but there was a trade off with convergence rate. Another method known to reduce the effect of noise is to apply a low pass filter to the gradient estimate [6] which also reduces the effect of noise by adding momentum to the weights.

A lot of recent research in pattern classification has focused on producing classifier solutions with a large margin (essentially the minimum distance of a data point to the decision boundary) [7]. It is well known that a large margin on a training sequence provides immunity to drift in the target solution and to noise in unseen data [2]. In this paper we examine the large margin methods of the Bayes Point Machine (BPM) [5] and the Online Bayes Point Machine [4] and apply them to the problem of adaptive channel equalization.

This work was supported by the Australian Research Council. Edward's research was funded by the Defence Science and Technology Organisation, Australia.

1.1. System model

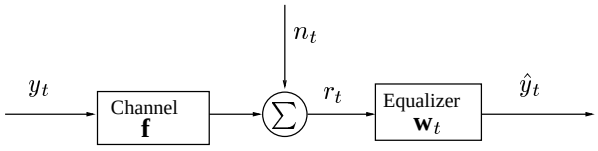

Fig. 1. Communications system discrete time model with channel equalization applied at the receiver.

The basic signal model at time t consists of a transmitted binary signal $y_t \in \{-1, +1\}$ (We only consider the binary case in this paper; this could be extended to complex or multi-class, suitable for quadrature modulation types). As shown by Figure 1, y_t is transmitted over a communications channel $\mathbf{f} = (f_0, \ldots, f_{L-1})$, resulting in the signal at the receiver r_t,

$$r_t = \sum_{l=0}^{L-1} y_{t-l} f_l + n_t, \tag{1}$$

where n_t is Additive Gaussian White Noise (AGWN) which is assumed to have zero mean and variance $\sigma_{n_t}^2$. We consider a linear equalizer $\mathbf{w}_t$ of length $2K+1$. The equalizer produces an estimate $\hat{y}_t$ of the transmitted signal y_t via

$$\hat{y}_t = \mathbf{w}_t \cdot \mathbf{x}_t, \tag{2}$$

where the instances $\mathbf{x}_t = [r_{t-K}, \ldots, r_t, \ldots, r_{t+K}]'$.

1.2. Stochastic gradient methods

The standard stochastic gradient method which approaches the optimal equalizer is to make steps scaled by $\eta \in \mathbb{R}$ in the direction of the gradient of the cost function J,

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \frac{\partial \mathrm{J}\left(\mathbf{w}_t\right)}{\partial \mathbf{w}_t}. \tag{3}$$

We consider two different families of cost functions. The first family consists of the regression methods using the cost functions of Least Mean Squared (LMS) [6] and Least Mean Absolute (LMA) which is known to be more robust to outliers than LMS [1]. The LMS cost function is given by,

$$\mathrm{J}_{\mathrm{LMS}}\left(\mathbf{w}_t\right) = \left(y_t - \mathbf{w}_t \cdot \mathbf{x}_t\right)^2. \tag{4}$$

The LMA cost function is given by,

$$\mathrm{J}_{\mathrm{LMA}}\left(\mathbf{w}_t\right) = \left|y_t - \mathbf{w}_t \cdot \mathbf{x}_t\right|. \tag{5}$$

The second family consists of the methods of the marginalised Perceptron [2] and relaxation with margin [2]. The marginalised Perceptron's cost function is

$$\mathrm{J}_{\mathrm{Perc}}\left(\mathbf{w}_t\right) = \left(\sigma\left(\mathbf{w}_t\right)\left(\rho - y_t \mathbf{w}_t \cdot \mathbf{x}_t\right)\right), \tag{6}$$

with $\sigma\left(\mathbf{w}_t\right) = 1$ when $y_t \mathbf{w}_t \cdot \mathbf{x}_t \leq \rho$ and zero otherwise, and ρ is the induced margin ($\rho = 0$ is the mistake driven standard Perceptron). The cost function of relaxation with margin is

$$\mathrm{J}_{\mathrm{relax}}\left(\mathbf{w}_t\right) = \left(\sigma\left(\mathbf{w}_t\right)\left(\rho - y_t \mathbf{w}_t \cdot \mathbf{x}_t\right)^2\right). \tag{7}$$

1.3. The Bayes point

Consider a fixed class $\mathcal{H}$ of classifiers and a sequence of T training examples $\mathbf{z} = \left(\left(\mathbf{x}_1, y_1\right), \ldots, \left(\mathbf{x}_T, y_T\right)\right)$. We would like to find a classifier from $\mathcal{H}$ which correctly classifies future examples drawn from the same distribution as $\mathbf{z}$. The Bayes optimal classifier chooses the label that minimises the probability of error, given the data $\mathbf{z}$. In general, the Bayes optimal classifier itself is not in $\mathcal{H}$ and may be very difficult to evaluate even if all the probability distributions are known. The *Bayes point* is the single hypothesis from $\mathcal{H}$ that achieves the minimum probability of error [5]. For linear classification the Bayes point is thus given by the weight vector that minimises the probability of a classification error. This is still quite difficult to find, motivating the use of approximations.

2. ESTIMATING THE BAYES POINT

Algorithms which estimate the Bayes point $\mathbf{w}_{\mathrm{BP}}$ are referred to as Bayes Point Machines (BPMs). The Bayes Point is approximated by $\mathbf{w}_{\mathrm{cm}}$ the average of $\mathbf{w}$ drawn from $\mathcal{H}$ according to the posterior distribution [5]. The BPM estimate of $\mathbf{w}_{\mathrm{cm}}$ for the linear classifier is

$$\tilde{\mathbf{w}}_t = \frac{1}{N}\sum_{i=1}^{N} \mathbf{w}_{t_i}, \tag{8}$$

where $\mathbf{w}_1, \ldots, \mathbf{w}_N$ are N different linear classifier solutions to $\mathbf{z}$.

As an illustration of the concepts related to estimating the Bayes point, consider the example of Figure 2 of a two dimensional space i.e. $\mathbf{w} \in \mathbb{R}^2$. The training set $\mathbf{z}$ consists of five examples, $T = 5$, each example $(\mathbf{x}, y)$ defines a half space $\{\mathbf{w} : y\mathbf{w} \cdot \mathbf{x} \geq 0\}$. The weight space defined by the intersection of half spaces is referred to as the version space $V(\mathbf{z})$, hence $V(\mathbf{z})$ defines the space of all weight vectors which correctly classify the five examples.

Now consider the significance of the margin ρ with respect to unseen examples. If the half space for the five examples now includes ρ i.e. $\{\mathbf{w} : y\mathbf{w} \cdot \mathbf{x} > \rho\}$ (as is the case in (6) and (7)) then we get the dotted region of Figure 2. The $\mathbf{w}_{\mathrm{cm}}$ is in the centre of this dotted region with a margin greater than ρ, providing an ability for $\mathbf{w}_{\mathrm{cm}}$ to handle noise in unseen examples. As the number of examples $T \to \infty$ then $\mathbf{w}_{\mathrm{cm}}$ approaches $\mathbf{w}_{\mathrm{BP}}$ [5], where one can imagine the version space defined by the curved line.

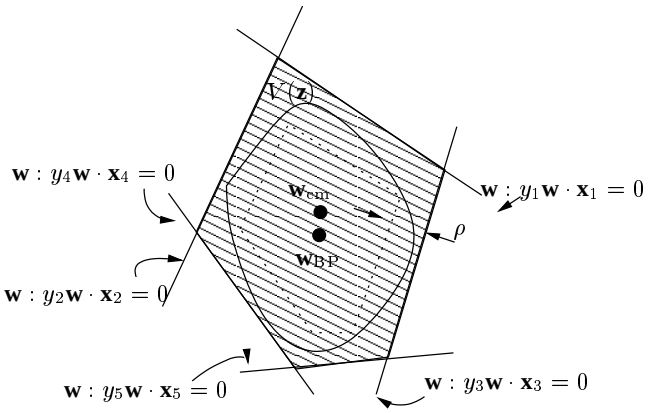

Fig. 2. Illustration of the Bayes point for a two dimensional feature space.

To estimate the $\mathbf{w}_{\mathrm{BP}}$ we endeavour to create diversity amongst N parallel and independently run equalizers, giving N solutions $\mathbf{w}_1, \ldots, \mathbf{w}_N$. We now discuss two different approaches to create the diversity.

2.1. Buffered approach

Consider running N equalizers in parallel where each equalizer $j = 1 \ldots, N$ sees a sequence formed from a permutation of B examples randomly selected from $\mathbf{z}$ without replacement. Hence for each equalizer j we associate a sequence $\pi_j(1), \ldots, \pi_j(B)$ of integers in $\{1, \ldots, T\}$ such that $\forall r, s \in \{1, \ldots, B\}$ with $r \neq s$ we have $\pi_j(r) \neq \pi_j(s)$. The sequence of examples seen by equalizer j is then $z_{t+\pi_j(1)}, \ldots, z_{t+\pi_j(B)}$; see Figure 3. The computational cost varies with B, where the best approximation to the Bayes point is $B = T$ has the highest computational cost. The number of arithmetic operations required for the buffer approach is $O(NKB)$ compared to $O(KT)$ for LMS.

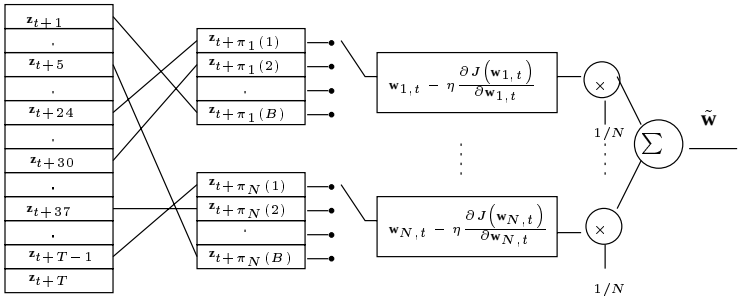

Fig. 3. Architecture of the buffer method.

2.2. Subsampling approach

The buffer approach suffers a latency of T samples. An alternative method which avoids latency is the Online Bayes Point Machine (OBPM) [4]. Given a training example $z_t = (\mathbf{x}_t, y_t)$, we run N equalizers "in parallel" and ensure diversity of their solutions by randomly choosing to present z_t to each equalizer j only if $b_{jt} = 1$, where b_{jt}, $j = 1, \ldots, N$, are independent Bernoulli random variables with $\Pr(b_{jt} = 1) = \tau$; see Figure 4. This process results in a subsample which has an average sample size τT and requires $O(\tau NKT)$ arithmetic operations. This approach exploits the fact that in general some examples are redundant to learning the best hypothesis.

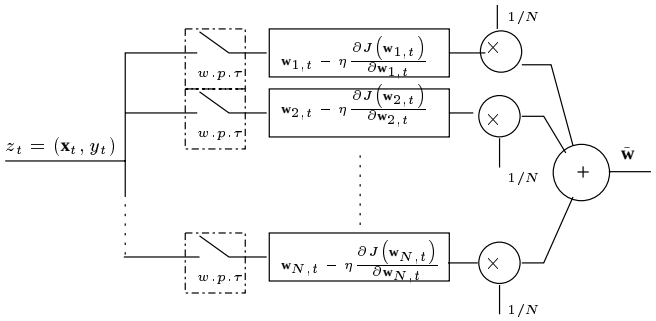

Fig. 4. Architecture of the subsampling method.

The cost functions one can use with OBPM are the second family of stochastic gradient methods; i.e. the classifiers (6) and (7). Unfortunately there is no gain by applying the subsampling of OBPM to regression methods. To see this, consider the example of LMS which is updated according to the cost function (4) in (3). If we set $\mathbf{w}_{i,0} = 0$ for all $i = 1, \ldots, N$ then equation (3) and (8) imply

$$\begin{aligned} \tilde{\mathbf{w}}_{t+1} &= \tilde{\mathbf{w}}_t + \eta y_t \mathbf{x}_t \sum_{i=1}^{N} \frac{b_{i,t}}{N} \\ &- \eta \left(\left(\frac{1}{N} \sum_{i=1}^{N} b_{i,t} \mathbf{w}_{i,t} \right) \cdot \mathbf{x}_t \right) \mathbf{x}_t. \end{aligned} \tag{9}$$

From (9) in the limit as $N \to \infty$ in the second term $\sum_{i=1}^{N} \frac{b_{i,t}}{N} \to \tau$ and in the third term $\frac{1}{N} \sum_{i=1}^{N} b_{i,t} \mathbf{w}_{i,t} \to \tau \tilde{\mathbf{w}}$. Therefore the subsampling for LMS simply scales η by τ.

3. EXPERIMENTS

To demonstrate the effectiveness of OBPM and BPM equalizers we considered two channels from [6, pages 631 and 686] in the experiments: channel A, $\mathbf{f}_A = (0.04, -0.05, 0.07, -0.21, 0.72, 0.36, 0, 0.21, 0.03, 0.07)$ and channel B, $\mathbf{f}_B = (0.26, 0.96, 0.26)$. The experimental results for the OBPM and BPM were all produced with the number of equalizers $N = 100$, and 200 Monte-Carlo trials. The training size in the BPM experiments was set at 300 binary labeled examples as that sufficed for convergence for the step sizes tried, η ranging from 0.01 to 0.09 in intervals of 0.01, and the step sizes reported are those which showed LMS results in MSE and probability of error in the best light. The test set was drawn independently to the training set and consisted of 10000 binary labeled examples.

3.1. Buffered experiments

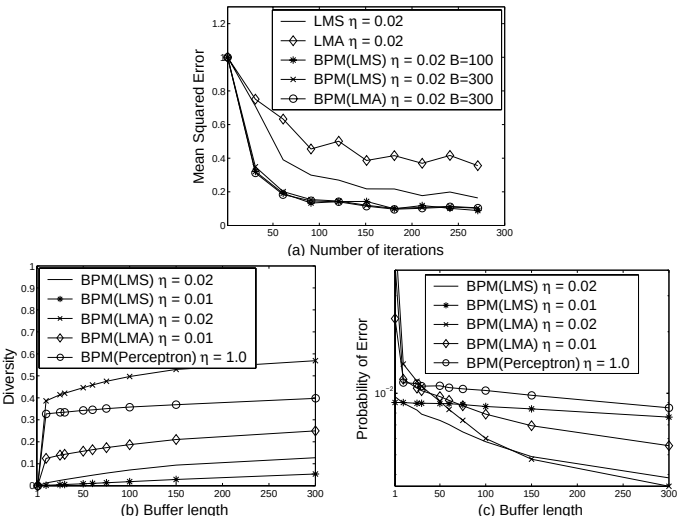

Fig. 5. BPM equalizer results for Channel A.

We investigated the performance of a 31 tap BPM equalizer on channel A with an SNR=10dB, results are in Figure 5. From Figure 5 (a) we see that the LMS and LMA have converged to a larger MSE compared to the BPM. The BPM MSE results for buffer sizes of 100 and 300 as shown in Figure 5 (a) were close. The relationship between the diversity of $\mathbf{w}$ given by $\frac{1}{N} \sum_{i=1}^{N} (\mathbf{w}_i - \tilde{\mathbf{w}})^2$ and the buffer length used by BPM is shown in Figure 5 (b). We see from Figure 5 (c) that there is an improvement in the probability of error after 300 training examples for channel A when using the

BPM(LMS) approach compared to LMS (buffer length of 1) for larger step sizes, η.

To simulate the possibility of phase tracker errors in the receiver we randomly flipped the labels so on average every tenth label was flipped. Label errors in the training sequence present a particularly difficult problem for schemes that try to maximize a margin. The label flipping experiment was performed on channel B with an SNR of 30dB and an 11 tap equalizer. The MSE of Figure 6 (a) shows that BPM(LMS) and BPM(LMA) were more stable for $\eta = 0.05$ compared to LMS and LMA (this was true over the range of η used). This indicates that this scheme is robust to label noise. Figure 6 (b) showed that the BPM(LMA) performed better than BPM(LMS).

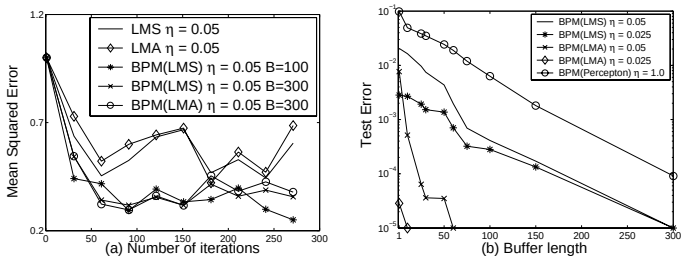

Fig. 6. Label noise results for Channel B using BPM equalizer, where only the training examples had label flipping, test examples had no label flipping.

3.2. Subsampled experiments

For the subsampled experiment the number of training examples went up to 20000 and the use of the Perceptron and relaxation with margin were investigated (i.e. (3) with (6) and (7)) with an SNR=10dB and $\rho = 0$ (we did not want to study the effect of ρ on producing large margin but τ). Due to limited space we only show the Perceptron results in Figure 7. The Perceptron was chosen instead of the relaxation algorithm since the probability of error for the choices of τ tried, ranging from 0.01 to 0.9 increments of 0.01 were better over the range. From Figure 7 we see that when τ was not equal to one (one being the standard Perceptron) the diversity amongst the N equalizers increased, with an improved probability of error. The Perceptron had a slower convergence in these experiments compared to the regression methods by an order of magnitude; taking 3000 training examples rather than 300.

4. CONCLUSIONS

We presented two methods which create diversity amongst N equalizer solutions run independently and in parallel. By taking the average of the weight vectors obtained by N equalizers we estimate the Bayes point which improves immunity

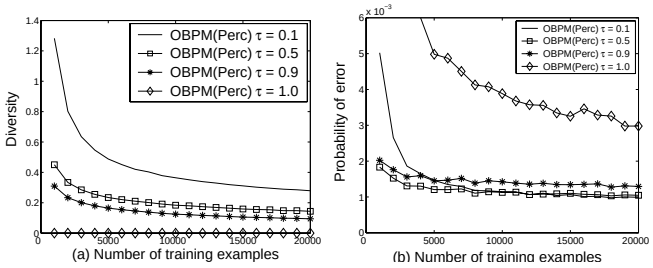

Fig. 7. OBPM equalizer results for Channel B.

to noise in the channel. We showed experimentally that, for appropriately large step sizes, the Bayes point equalizer was an improvement on LMS and LMA in the presence of channel noise and training sequence errors. The use of a Bayes point equalizer allows the training sequence to be made shorter, although at the cost of higher computational demands. Shorter training sequences are desirable increasing channel throughput. An area for further research is the incorporation of phase tracking with this equalizer to make a more practical system.

5. REFERENCES

[1] C. M. Bishop, (1995) *Neural Networks for Pattern Recognition.* Oxford University Press.

[2] R.O. Duda, P.E. Hart and D.G. Stork, (2000) *Pattern Classification And Scene Analysis, 2nd Edition.* John Wiley.

[3] W.A. Gardner, (1984) Learning Characteristics of Stochastic-Gradient Descent Algorithms: A General Study, Analysis, and Critique, *Signal Processing* vol. 6, pp. 113-133.

[4] E. Harrington, J. Kivinen, R. C. Williamson, R. Herbrich and J. Platt, (2002) Online Bayes Point Machine. *in preparation*.

[5] R. Herbrich, T. Graepel and C. Campbell, (2001) Bayes Point Machines. *Journal of Machine Learning Research*, **1**:245-279.

[6] J.G. Proakis, (2001) *Digital Communications, 4th Edition.* McGraw-Hill.

[7] A. Smola, P.L. Bartlett, B. Schölkopf and D. Schuurmans (Eds), (2000) *Advances in Large Margin Classifiers*, MIT Press.