Decision feedback equaliser design using support vector machines

S.Chen, S.Gunn and C.J.Harris

Abstract: The conventional decision feedback equaliser (DFE) that employs a linear combination of channel observations and past decisions is considered. The design of this class of DFE is to construct a hyperplane that separates the different signal classes. It is well known that the popular minimum mean square error (MMSE) design is generally not the optimal minimum bit error rate (MBER) solution. A strategy is proposed for designing the DFE based on support vector machines (SVMs). The SVM design achieves asymptotically the MBER solution and is superior in performance to the usual MMSE solution. Unlike the exact MBER solution, this SVM solution can be computed very efficiently.

1 Introduction

The equalisation technique plays an ever-increasing role in combating distortion and interference in modern communication links [1, 2] and high-density data storage systems [3, 4]. The conventional DFE, in particular, is widely used as it provides an excellent balance between performance and complexity. The conventional DFE [1] is based on a linear filtering of the channel observations and the past decisions. Research has also investigated other DFE structures that employs nonlinear combinations of the channel observations and the past decisions [5–10]. These nonconventional DFEs provide better performance at a cost of increasing complexity. In this study we revisit the conventional DFE from the viewpoint of the powerful statistical learning theory [11].

Without confusion, we use the term DFE to refer to the conventional linear-combiner DFE. From the viewpoint of classification, a DFE forms a hyperplane in the observation space. This geometrical interpretation provides insights to various DFE designs. The most popular design strategy is the MMSE design. The MMSE DFE has an attractive advantage in that it leads to a simple adaptive implementation in terms of the least mean square (LMS) algorithm with very low complexity. However, the MMSE solution is not the MBER solution, the bit error rate (BER) being the ultimate performance criterion of equalisation. It has been shown that, in the asymptotic case of large signal to noise ratio (SNR), the hyperplane of the MMSE DFE is orthogonal to the last axis of a translated observation space [12], which clearly illustrates why the MMSE solution does not achieve the full performance potential of the DFE structure.

IEE Proceedings online no. 20000360

DOI: 10.1049/ip-vis:20000360

Paper first received 26th April and in revised form 12th October 1999

The authors are with the Department of Electronics and Computer Science, University of Southampton, Highfield, Southampton SO17 1BJ, UK

A better design in terms of performance is to choose the equaliser coefficients to minimise BER directly [12–15]. Although the performance of the resulting MBER DFE is superior to the MMSE solution, adaptive implementation of the MBER solution [12] is computationally much more complex than the simple LMS algorithm. Furthermore, unlike the mean square error surface, which is quadratic and has a unique minimum, the BER surface can be highly irregular and a gradient algorithm cannot generally guarantee to converge to a global minimum.

A relevant development to the MBER DFE is the approximate MBER design for linear equalisers without decision feedback [16, 17]. This approximate MBER solution was derived for the special case of equalisable channels. Equalisability refers to the linear separability of different channel state classes. It is well known that the linear separability cannot always be guaranteed when a linear equaliser is used [18]. This approximate MBER design does have a simple adaptive implementation similar in the form to the LMS algorithm [16, 17]. Unfortunately, this adaptive algorithm requires an extremely long training period owing to the nature of minimising BER.

SVMs are powerful approach for solving various classification and regression problems [11, 19–21]. The formulation of SVMs embodies the structural risk minimisation (SRM) principle [11]. SRM minimises an upper bound on the expected risk, as opposed to the traditional empirical risk minimisation that minimises the error on the training data. This difference equips SVMs with a greater ability to generalise. The SVM approach results in a simple quadratic optimisation problem. For classification problems, generalisation ability is optimised by maximising the margin, and the solution is obtained as a set of sparse support vectors (SVs), which lie on the margin boundary and summarise the information required to separate the data.

In this study we apply SVMs to design the DFE. As the decision feedback always guarantees the linear separability [12], the design of an optimal hyperplane to separate different signal classes using SVMs becomes a simple matter, and the SVM solution is obtained by maximising the margin, which is defined as the minimum distance from the signal states to the separating hyperplane. The SVM design leads to a well-conditioned quadratic programming defined by a few channel states (SVs) that lie on the margin. The SVM solution does not depend on the noise variance and is asymptotically the MBER solution. Unlike the MBER solution, however, the SVM design is guaranteed to be unique and can be computed very efficiently. Simulation results also indicate that the performance of the SVM DFE is virtually indistinguishable from the MBER DFE.

2 Decision feedback equaliser structure

Consider the channel that is modelled as a finite impulse response filter with an additive noise source. Specifically, the received signal at sample k is

$$r(k) = \bar{r}(k) + e(k) = \sum_{i=0}^{n_a-1} a_i s(k-i) + e(k) \qquad (1)$$

where $\bar{r}(k)$ denotes the noiseless channel observation, n_a is the channel length and a_i are the channel tap weights, $\{e(k)\}$ is a Gaussian white noise sequence having zero mean and variance σ_e^2, and the symbol sequence $\{s(k)\}$ is independently identically distributed and is uncorrelated with $e(k)$. The SNR of the system is defined as

$$\text{SNR} = \left(\sum_{i=0}^{n_a-1} a_i^2\right)\sigma_s^2/\sigma_e^2, \qquad (2)$$

where σ_s^2 is the symbol variance. In this study, for the purpose of easy geometrical visualisation, we assume that $s(k)$ is binary taking value from the set $\{\pm 1\}$. For multi-level signalling schemes, such as M-PAM, the current results are still valid [15].

The structure of a generic DFE is depicted in Fig. 1. The equalisation process defined in Fig. 1 uses the information present in the channel observation vector

$$\boldsymbol{r}(k) = [r(k) \cdots r(k-m+1)]^T \qquad (3)$$

and the past detected symbol vector

$$\hat{\boldsymbol{s}}_b(k) = [\hat{s}(k-d-1) \cdots \hat{s}(k-d-n)]^T \qquad (4)$$

to produce an estimate $\hat{s}(k-d)$ of $s(k-d)$, where the integers d, m and n are the decision delay, the feedforward and feedback orders, respectively. In particular, for the class of DFEs considered in this study, the decision is made by quantising the filter output

$$f(\boldsymbol{r}(k), \hat{\boldsymbol{s}}_b(k)) = \boldsymbol{w}^T \boldsymbol{r}(k) + \boldsymbol{b}^T \hat{\boldsymbol{s}}_b(k) \qquad (5)$$

where

$$\left.\begin{aligned} \boldsymbol{w} &= [w_0 \cdots w_{m-1}]^T \\ \boldsymbol{b} &= [b_1 \cdots b_n]^T \end{aligned}\right\} \qquad (6)$$

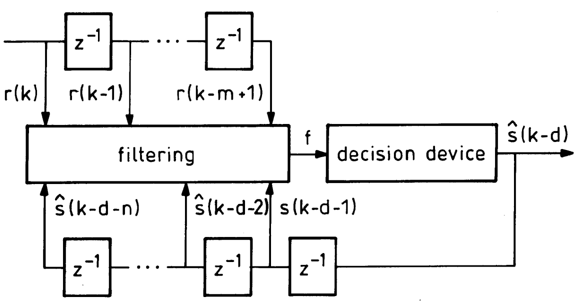

Fig. 1 *Schematic of generic decision feedback equaliser*

are the coefficients of the feedforward and feedback filters, respectively. Without the loss of generality, we choose $d = n_a - 1$, $m = n_a$ and $n = n_a - 1$, as this choice of the DFE structure parameters is sufficient to guarantee the linear separability [12].

To better understand how the DFE works, a space translation [12] can be made to 'remove' the feedback term from eqn. 5. The observation vector (eqn. 3) can be arranged as

$$\boldsymbol{r}(k) = \boldsymbol{F}\boldsymbol{s}(k) + \boldsymbol{e}(k) \qquad (7)$$

where $\boldsymbol{e}(k) = [e(k) \cdots e(k-m+1)]^T$, $\boldsymbol{s}(k) = [\boldsymbol{s}_f^T(k)\boldsymbol{s}_b^T(k)]^T$ with

$$\left.\begin{aligned} \boldsymbol{s}_f(k) &= [s(k) \cdots s(k-d)]^T \\ \boldsymbol{s}_b(k) &= [s(k-d-1) \cdots s(k-d-n)]^T \end{aligned}\right\} \qquad (8)$$

and the $m \times (d+1+n)$ matrix $\boldsymbol{F}$ has the form

$$\boldsymbol{F} = [\boldsymbol{F}_1 \quad \boldsymbol{F}_2] \qquad (9)$$

with the $m \times (d+1)$ matrix $\boldsymbol{F}_1$ and $m \times n$ matrix $\boldsymbol{F}_2$ defined by

$$\boldsymbol{F}_1 = \begin{bmatrix} a_0 & a_1 & \cdots & a_{n_a-1} \\ 0 & a_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_1 \\ 0 & \cdots & 0 & a_0 \end{bmatrix} \qquad (10)$$

and

$$\boldsymbol{F}_2 = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ a_{n_a-1} & 0 & \ddots & \vdots \\ a_{n_a-2} & a_{n_a-1} & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 \\ a_1 & \cdots & a_{n_a-2} & a_{n_a-1} \end{bmatrix} \qquad (11)$$

respectively. Under the assumption of correct decision feedback, that is, $\hat{\boldsymbol{s}}_b(k) = \boldsymbol{s}_b(k)$,

$$\boldsymbol{r}(k) = \boldsymbol{F}_1\boldsymbol{s}_f(k) + \boldsymbol{F}_2\hat{\boldsymbol{s}}_b(k) + \boldsymbol{e}(k) \qquad (12)$$

Thus the decision feedback translates the original space $\boldsymbol{r}(k)$ into a new space $\boldsymbol{r}'(k)$

$$\boldsymbol{r}'(k) \triangleq \boldsymbol{r}(k) - \boldsymbol{F}_2\hat{\boldsymbol{s}}_b(k) \qquad (13)$$

and it can be shown that the linear filter (eqn. 5) is reduced to

$$f'(\boldsymbol{r}'(k)) = \boldsymbol{w}^T\boldsymbol{r}'(k) \qquad (14)$$

in the translated space [12]. Previous work [12, 14] has pointed out that the elements of $\boldsymbol{r}'(k)$ can be computed recursively according to

$$\left.\begin{aligned} r'(k-i) &= z^{-1}r'(k-i+1) - a_{n_a-i}\hat{s}(k-d-1), \\ & \qquad i = m-1, \ldots, 2, 1 \\ r'(k) &= r(k) \end{aligned}\right\} \qquad (15)$$

where z^{-1} is interpreted as the unit delay operator, and the DFE structure of Fig. 1 is equivalent to the equalisation structure of Fig. 2.

By considering the alternative but equivalent DFE structure of Fig. 2, the geometric insight of the equalisation process becomes evident: the DFE design is basically to

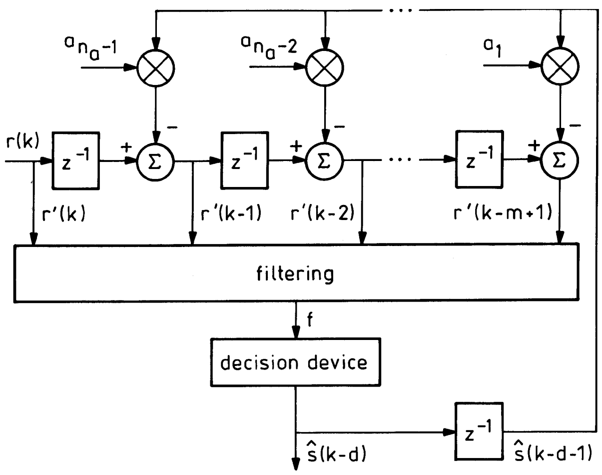

Fig. 2 *Schematic of translated decision feedback equaliser*

choose the weight vector $\boldsymbol{w}$ of the hyperplane $\{\boldsymbol{r}:\boldsymbol{w}^T\boldsymbol{r}=0\}$. For the DFE, a hyperplane can always be constructed to separate the two different classes of signal states, as shown in the following lemma. Let the $N_f=2^{d+1}$ sequences of $\boldsymbol{s}_f(k)$ be $\boldsymbol{s}_{f,j}$, $1\leq j\leq N_f$. The set of the noiseless channel states in the translated space is defined by

$$\mathcal{R}\triangleq\{\boldsymbol{r}_j=\boldsymbol{F}_1\boldsymbol{s}_{f,j},\ 1\leq j\leq N_f\}\tag{16}$$

This set can be partitioned into two subsets conditioned on $s(k-d)=\pm1$

$$\mathcal{R}^{\pm}\triangleq\{\boldsymbol{r}_j\in\mathcal{R}|s(k-d)=\pm1\}\tag{17}$$

Lemma 1: $\mathcal{R}^+$ and $\mathcal{R}^-$ are linearly separable.

Proof: Choose the weight vector of a hyperplane $\boldsymbol{w}^T\boldsymbol{r}=0$ to be:

$$\boldsymbol{w}=\left[0\cdots0\ \frac{1}{a_0}\right]^T$$

For any $\boldsymbol{r}^+\in\mathcal{R}^+$ and $\boldsymbol{r}^-\in\mathcal{R}^-$, we have $\boldsymbol{w}^T\boldsymbol{r}^+=1>0$ and $\boldsymbol{w}^T\boldsymbol{r}^-=-1<0$.

2.1 MMSE decision feedback equaliser

From the discussion so far, it is clear that different DFE designs correspond to different constructions of hyperplanes to separate the two classes of signal states $\mathcal{R}^{\pm}$. The best known design is the MMSE DFE. The MMSE solution obtained by minimising $\mathrm{E}[(\boldsymbol{w}^T\boldsymbol{r}'(k)-s(k-d))^2]$, where $\mathrm{E}[\cdot]$ denotes the expectation operation, is given by

$$\boldsymbol{w}_{\mathrm{MMSE}}=\sigma_s^2\Gamma^{-1}[a_{n_a-1}a_{n_a-2}\cdots a_0]^T,\tag{18}$$

where the matrix Γ is symmetric with elements $\gamma_{j,l}=\gamma_{l,j}$

$$\gamma_{l,j}=\left(\sum_{i=0}^{(n_a-1)-l}a_ia_{i+(l-j)}\right)\sigma_s^2+\sigma_e^2\delta(l-j),\ \ 0\leq j\leq l\leq m-1\tag{19}$$

and $\delta(q)$ is the discrete Dirac delta function defined by

$$\delta(q)=\begin{cases}1, & q=0\\ 0, & q\neq0\end{cases}\tag{20}$$

The nonoptimal nature of the MMSE solution is best illustrated by the asymptotic case of SNR$\to\infty$, given in the following lemma (see [12] for a proof).

Lemma 2

$$\lim_{\mathrm{SNR}\to\infty}\boldsymbol{w}_{\mathrm{MMSE}}=\left[0\ 0\cdots0\ \frac{1}{a_0}\right]^T\tag{21}$$

In the limit case of SNR$\to\infty$, the MMSE hyperplane is always orthogonal to the last axis of the $\boldsymbol{r}'(k)$-space, which cannot be optimal in general. A simple example taken from [12] clearly illustrates this observation. Consider the two-tap channel defined by

$$\text{Channel 1}:\boldsymbol{a}=[a_0\ a_1]^T=[0.5\ 1.0]^T\tag{22}$$

and the DFE given by $d=1$, $m=2$ and $n=1$. The set of the four channel states $\mathcal{R}$ in the translated observation space is shown in Fig. 3, where the asymptotic hyperplane of the MMSE solution for large SNR is also depicted. Common sense would suggest that the "optimal" separating hyperplane for this example should have a slope of $-1(w_0/w_1=1)$, which in fact is the asymptotic MBER solution for large SNR. When the noise is added the hyperplane of the MMSE solution (eqn. 18) will rotate and is no longer orthogonal to the last axis of the translated observation space. Consider the channel 1 of eqn. 22 again. Given SNR = 15 dB, the MMSE hyperplane has a slope of -0.27, which is clearly nonoptimal. In fact, the numerical solution of the MBER DFE in this case gives a slope of -1.02.

2.2 MBER decision feedback equaliser

It can be shown that the BER of the DFE of eqn. 14, taking into account the symmetry of $\mathcal{R}^+$ and $\mathcal{R}^-$ and the equiprobability of states, is given by (e.g. [12])

$$P_\varepsilon(\boldsymbol{w})=\frac{2}{N_f}\sum_{\boldsymbol{r}_j\in\mathcal{R}^+}Q\left(\frac{|\boldsymbol{w}^T\boldsymbol{r}_j|}{\|\boldsymbol{w}\|\sigma_e}\right)\tag{23}$$

where

$$Q(x)=\int_x^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)dx\tag{24}$$

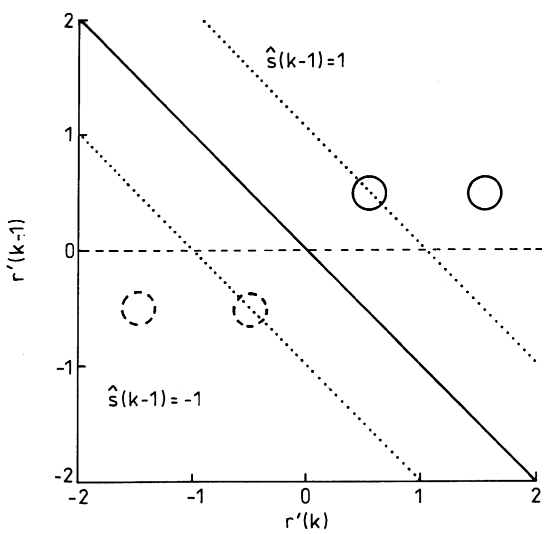

Fig. 3 *Two asymptotic decision hyperplanes corresponding to large SNR for channel* $\boldsymbol{a}=[0.5\ 1.0]^T$

Circles denote two subsets of states $\mathcal{R}^-$ and $\mathcal{R}^+$
—— SVM
---- MMSE
····· margin

The MBER DFE is obtained by minimising $P_\varepsilon(\boldsymbol{w})$

$$\boldsymbol{w}_{\text{MBER}} = \arg\min_{\boldsymbol{w}} P_\varepsilon(\boldsymbol{w}) \tag{25}$$

Notice that the elements of $\boldsymbol{w}$ are linearly dependent. The optimisation in eqn. 25 should ideally be subject to some constraint on $\boldsymbol{w}$, e.g. $\|\boldsymbol{w}\| = 1$.

Obviously, $\boldsymbol{w}_{\text{MBER}}$ is optimal for the linear-combiner DFE structure discussed here. Unlike the MMSE solution, there is no closed-form expression for $\boldsymbol{w}_{\text{MBER}}$ and the solution must be sought via some gradient-based iterative procedure. In general, the BER surface $P_E(\boldsymbol{w})$ can be highly irregular and may have local minima. An adaptive implementation of the MBER DFE is described in [12] but its computational complexity per sample is much higher than the LMS algorithm. An adaptive implementation of the approximate MBER linear equaliser [16, 17] has a complexity per sample similar to the LMS algorithm but it requires a very long training sequence.

2.3 SVM decision feedback equaliser

SVMs are a powerful learning approach to data modelling [11, 19–21]. Notice that SVM are very general and classification problems with nonlinear separable or overlapping classes can successfully been solved using SVM. For our DFE problem. as $\mathcal{R}^+$ and $\mathcal{R}^-$ are always linearly separable, only the basic part of SVM is required. For completeness, SVM for the linearly separable classification case are described in the Appendix. The minimum distance from the nearest point in $\mathcal{R}$ to a separating hyperplane $\boldsymbol{w}^T\boldsymbol{r} = 0$

$$\rho(\boldsymbol{w}) = \min_{\boldsymbol{r}_i \in \mathcal{R}^+} \frac{|\boldsymbol{w}^T\boldsymbol{r}_i|}{\|\boldsymbol{w}\|} + \min_{\boldsymbol{r}_j \in \mathcal{R}^-} \frac{|\boldsymbol{w}^T\boldsymbol{r}_j|}{\|\boldsymbol{w}\|} \tag{26}$$

is called the margin. The SVM design finds the hyperplane that maximises this margin. As there is some redundancy in $\boldsymbol{w}^T\boldsymbol{r} = 0$, it is appropriate to consider a canonical hyperplane [11] where $\boldsymbol{w}$ is constrained by

$$\min_{\boldsymbol{r}_i \in \mathcal{R}} |\boldsymbol{w}^T\boldsymbol{r}_i| = 1 \tag{27}$$

Define the integer set

$$I_{\mathcal{R}} \triangleq \{i | \boldsymbol{r}_i \in \mathcal{R}\} \tag{28}$$

and the class indicator

$$y_i = \pm 1, \ \forall \boldsymbol{r}_i \in \mathcal{R}^\pm \tag{29}$$

The maximisation of the margin (eqn. 26) with the constraint (eqn. 27) using the classical lagrangian theory [22, 23] gives rise to the optimal separating hyperplane

$$\boldsymbol{w}_{\text{SVM}} = \sum_{i \in I_{\mathcal{R}}} \bar{g}_i y_i \boldsymbol{r}_i \tag{30}$$

where

$$\bar{\boldsymbol{g}} = \arg\min_{\boldsymbol{g}} \frac{1}{2} \sum_{i \in I_{\mathcal{R}}} \sum_{j \in I_{\mathcal{R}}} g_i g_j y_i y_j \boldsymbol{r}_i^T \boldsymbol{r}_j - \sum_{i \in I_{\mathcal{R}}} g_i \tag{31}$$

$$g_i \geq 0, \ \forall i \in I_{\mathcal{R}} \tag{32}$$

The optimisation problem of eqn. 31 with eqn. 32 is a standard quadratic programming, whose solution $\bar{\boldsymbol{g}}$ can readily be computed efficiently [22, 23]. Notice that g_i are the Lagrange multipliers of the primal problem. From the Kuhn–Tucker conditions [22, 23]

$$\bar{g}_i (y_i \boldsymbol{w}_{\text{SVM}}^T \boldsymbol{r}_i - 1) = 0 \tag{33}$$

and only those points $\boldsymbol{r}_i$, which satisfy $y_i \boldsymbol{w}_{\text{SVM}}^T \boldsymbol{r}_i = 1$, will have nonzero Lagrange multipliers. These points are termed support vectors. All the SVs lie on the margin and the number of SVs can be very small. Let $\mathcal{R}_{\text{SV}}$ be the set of SVs. The hyperplane $\boldsymbol{w}_{\text{SVM}}$ is determined by $\mathcal{R}_{\text{SV}}$ only. Thus, the identical solution is obtained by substituting $\mathcal{R}$ with $\mathcal{R}_{\text{SV}}$ in eqns. 30 to 32 with massive saving in computational efforts. However, SVs are unknown *a priori.*

Consider how to select a smaller subset $\mathcal{R}_{\text{Sub}}$ of $\mathcal{R}$, which contains all the SVs. We first point out that, as we restrict ourselves to the linear structure of eqn. 14; the decision boundary is a hyperplane. If we allow a nonlinear structure, however, the decision boundary will become a hypersurface. The true optimal solution for the equaliser structure of Fig. 1 without restricting to linear filtering is the nonlinear bayesian solution [7, 24]. Asymptotically, the bayesian hypersurface consists of a set of hyperplanes [25]. Each of these hyperplanes is defined by a pair of the *dominant* states in $\mathcal{R}^+$ and $\mathcal{R}^-$, respectively, and the line connecting this pair is perpendicular to the hyperplane with the midpoint of the line at the hyperplane. The following algorithm can be used to select these pairs of the states which define the set of the asymptotic hyperplanes.

Subset selection algorithm

FOR $\boldsymbol{r}_i^+ \in \mathcal{R}^+$
  FOR $\boldsymbol{r}_j^- \in \mathcal{R}^-$
    $\boldsymbol{x} = (\boldsymbol{r}_i^+ + \boldsymbol{r}_j^-)/2$
    $d_0 = \|\boldsymbol{r}_i^+ - \boldsymbol{x}\|$
    IF $(\|\boldsymbol{r}_l^+ - \boldsymbol{x}\| > d_0, \ \forall \boldsymbol{r}_l^+ \in \mathcal{R}^+, l \neq i)$ AND
      $(\|\boldsymbol{r}_l^- - \boldsymbol{x}\| > d_0, \ \forall \boldsymbol{r}_l^- \in \mathcal{R}^-, l \neq j)$
      $\mathcal{R}_{\text{Sub}} \leftarrow (\boldsymbol{r}_i^+, \boldsymbol{r}_j^-)$
    END IF
  NEXT $\boldsymbol{r}_j^-$
NEXT $\boldsymbol{r}_i^+$

As $\mathcal{R}_{\text{Sub}}$ can be a much smaller subset of $\mathcal{R}$, using it to substitute $\mathcal{R}$ in the quadratic programming of eqns. 30 to 32 will result in considerable saving in computation. Obviously, this is allowed, as all the SVs are included in $\mathcal{R}_{\text{Sub}}$, and we have

Proposition 1: $\mathcal{R}_{\text{SV}} \subset \mathcal{R}_{\text{Sub}}$.

We comment that $\boldsymbol{w}_{\text{SVM}}$ does not depend on the noise variance σ_e^2 and it is the asymptotic MBER solution, that is, $\boldsymbol{w}_{\text{SVM}} \to \boldsymbol{w}_{\text{MBER}}$ as SNR $\to \infty$. In general, $\boldsymbol{w}_{\text{SVM}}$ will not be identical to $\boldsymbol{w}_{\text{MBER}}$ but the difference is practically negligible for useful SNR conditions. Consider channel 1 of eqn. 22. The hyperplane of the SVM solution is depicted in Fig. 3. This is identical to the asymptotic MBER solution for large SNR. When the SNR is reduced to 15 dB, $\boldsymbol{w}_{\text{SVM}}$ remains unchanged with a slope of -1 but $\boldsymbol{w}_{\text{MBER}}$ is changed, from a slope of -1 to -1.02. Such a small difference will hardly cause any difference in BER performance between the SVM DFE and the MBER DFE.

2.4 Numerical examples

Three examples were used to compare the SVM and MMSE solutions of the DFE. All the BERs were evaluated with detected symbols being fed back. For all the three examples, the BERs of the MBER DFE were practically indistinguishable from those of the SVM DFE. Therefore the BER curves of the MBER DFE are not included. The first example was the two-tap channel given in eqn. 22. The full set of states $\mathcal{R}$ contains four points, and the subset selection algorithm selected a $\mathcal{R}_{\text{Sub}}$ of two states, which are the two SVs. Fig. 4 compares the BERs of the SVM DFE

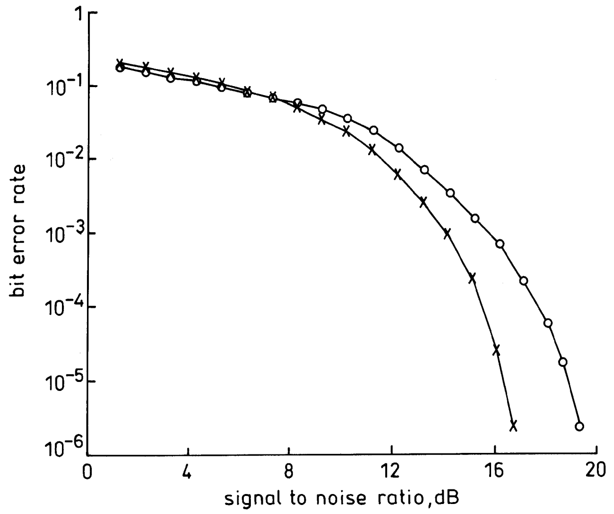

Fig. 4 *Performance comparison for channel* $\boldsymbol{a} = [0.5\ 1.0]^T$ *with detected symbols being fed back*
–X– SVM
–O– MMSE

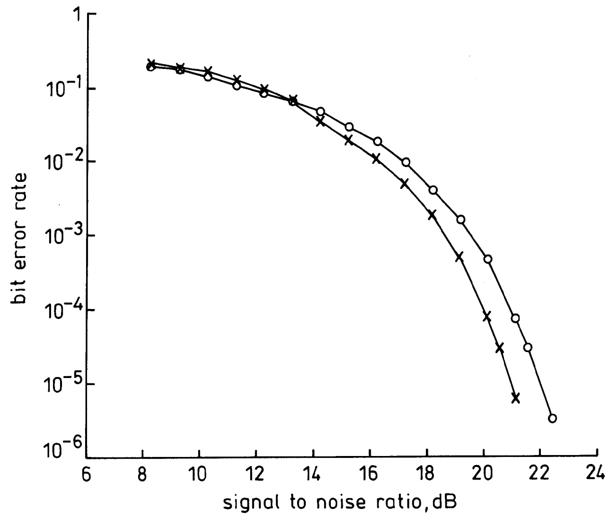

Fig. 6 *Performance comparison for channel* $\boldsymbol{a} = [0.227\ 0.466\ 0.688\ 0.466\ 0.227]^T$ *with detected symbols being fed back*
–X– SVM
–O– MMSE

with those of the MMSE DFE for a range of SNR conditions. For this example, the SVM DFE has a SNR gain of about 2 dB over the MMSE solution at the BER of 10^{-4}.

The second example was a four-tap channel given by

$$\text{channel} \quad 2 : \boldsymbol{a} = [0.35\ 0.80\ 1.00\ 0.80]^T \tag{34}$$

The structure of the DFE was accordingly chosen to be $d = 3$, $m = 4$ and $n = 3$. The full set of states $\mathcal{R}$ has sixteen points. The subset selection produced a subset $\mathcal{R}_{\text{Sub}}$ of eight states, four of them being the SVs. The BERs of the MMSE and SVM DFEs with detected symbols being fed back are plotted in Fig. 5, where it can be seen that the performance of the SVM DFE is significantly better than that of the MMSE DFE. At the BER of 10^{-4}, the SVM DFE has a SNR gain of about 2 dB over the MMSE solution.

The third example was a five-tap channel defined by

$$\text{channel} \quad 3 : \boldsymbol{a} = [0.227\ 0.466\ 0.688\ 0.466\ 0.227]^T \tag{35}$$

The structural parameters of the DFE were set to $d = 4$, $m = 5$ and $n = 4$. The full set $\mathcal{R}$ contains 32 states. The subset $\mathcal{R}_{\text{Sub}}$ used in computing the SVM solution has 18 points, eight of them being the SVs. The BERs of the SVM and MMSE DFEs with detected symbols being fed back are depicted in Fig. 6. Again, the SVM solution is superior in performance over the MMSE solution.

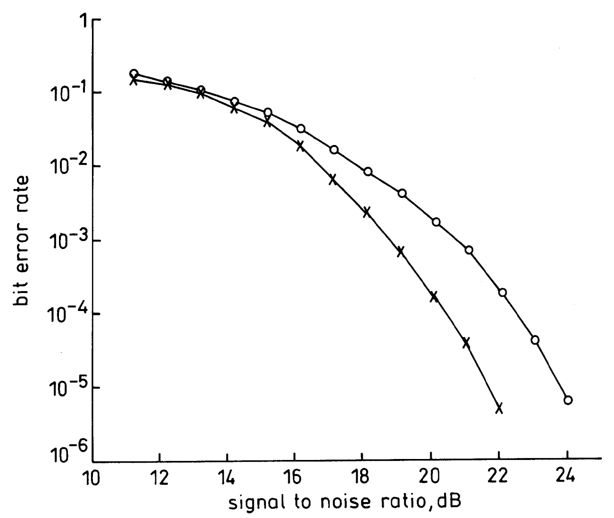

Fig. 5 *Performance comparison for channel* $\boldsymbol{a} = [0.35\ 0.80\ 1.00\ 0.80]^T$ *with detected symbols being fed back*
–X– SVM
–O– MMSE

2.5 Adaptive implementation

An indirect adaptation scheme is suitable for adaptive implementation of the SVM DFE. The scheme first estimates a channel model $\hat{\boldsymbol{a}}$ using for example the LMS algorithm and then computes the weight vector $\boldsymbol{w}_{\text{SVM}}$ of the SVM DFE based on the channel estimate $\hat{\boldsymbol{a}}$. This indirect approach, as opposed to a direct adaptation of the equaliser weight vector using the LMS algorithm, has an advantage of shorter training period. This is because the correlation matrix of the LMS channel estimator has an eigenvalue spread of one, while the correlation matrix of the LMS algorithm for updating the equaliser weight vector can have a large eigenvalue spread [26]. The LMS channel estimator typically requires training samples ten times of the channel length.

Implementing the SVM DFE in data storage systems is particularly simple, as in many commercial disc drives, the equalisers are trained at the factory floor and then 'frozen' before shipping. Thus training can be done off-line in one go. For time-varying communication links, it is possible to implement the SVM-DFE in a block-by-block adaptation. For example, in many communication systems, transmission is organised in frames. Each frame contains a training portion, which can be used in channel estimation. The estimated channel is then used to design the SVM DFE to detect data in the frame. The adaptive SVM DFE is computationally more complex than the adaptive MMSE DFE, due to the need of solving a quadratic programming. The increase in computation, however, can partly be justified by an improved performance.

3 Conclusions

The linear-combiner DFE partitions the signal space with a hyperplane. This geometric visualisation provides insights into various DFE designs. The best-known DFE design is the MMSE DFE with its computational simplicity. At the other end, the sophisticated MBER DFE offers the optimal performance for the linear-combiner DFE structure, at the cost of solving a complex nonlinear optimisation problem. We have proposed a new DFE design based on SVM. The SVM DFE, being the asymptotic MBER solution, is superior in performance to the MMSE DFE. The SVM approach results in a simple quadratic programming, and the solution can be computed very efficiently. Adaptive implementation of the SVM DFE is also discussed, and it is possible to realise the SVM DFE in data storage systems and slow time-varying communication links.

4 References

1 QURESHI, S.U.H.: 'Adaptive equalization', *Proc. IEEE*, 1985, **73**, (9), pp. 1349–1387
2 PROAKIS, J.G.: 'Digital communications' (McGraw-Hill, New York, 1995, 3rd edn.)
3 MOON, J.: 'The role of SP in data-storage systems', *IEEE Signal Process. Mag.*, 1998, **15**, (4), pp. 54–72
4 PROAKIS, J.G.: 'Equalization techniques for high-density magnetic recording', *IEEE Signal Process. Mag.*, 1998, **15**, (4), pp. 73–82
5 SIU, S., GIBSON, G.J., and COWAN, C.F.N.: 'Decision feedback equalisation using neural network structures and performance comparison with the standard architecture', *IEE Proc. I, Commun. Speech Vis.*, 1990, **137**, (4), pp. 221–225
6 WILLIAMSON, D., KENNEDY, R.A., and PULFORD, G.W.: 'Block decision feedback equalization', *IEEE Trans. Commun.*, 1992, **40**, (2), pp. 255–264
7 CHEN, S., MULGREW, B., and McLAUGHLIN, S.: 'Adaptive Bayesian equaliser with decision feedback', *IEEE Trans. Signal Process.*, 1993, **41**, (9), pp. 2918–2927
8 CHEN, S., McLAUGHLIN, S., and MULGREW, B.: 'Complex-valued radial basis function network, Part II: application to digital communications channel equalisation', *EURASIP Signal Process. J.*, 1994, **36**, pp. 175–188
9 CHEN, S., McLAUGHLIN, S., MULGREW, B., and GRANT, P.M.: 'Adaptive bayesian decision feedback equaliser for dispersive mobile radio channels', *IEEE Trans. Commun.*, 1995, **43**, (5), pp. 1937–1946
10 CHA, I., and KASSAM, S.A.: 'Channel equalization using adaptive complex radial basis function networks', *IEEE J. Sel. Areas Commun.*, 1995, **13**, (1), pp. 122–131
11 VAPNIK, V.: 'The nature of statistical learning theory' (Springer-Verlag, New York, 1995)
12 CHEN, S., MULGREW, B., CHNG, E.S., and GIBSON, G.: 'Space translation properties and the minimum-BER linear-combiner DFE', *IEE Proc. Commun.*, 1998, **145**, (5), pp. 316–322
13 SHAMASH, E., and YAO, K.: 'On the structure and performance of a linear decision feed-back equalizer based on the minimum error probability criterion'. Proceedings of ICC'74, 1974, pp. 25F1–25F5
14 CHEN, S., CHNG, E.S., MULGREW, B., and GIBSON, G.: 'Minimum-BER linear-combiner DFE'. Proceedings of ICC'96, 1996, Dallas, Texas, **2**, pp. 1173–1177
15 CHEN, S., and MULGREW, B.: 'Minimum-SER linear-combiner decision feedback equaliser', *IEE Proc. Commun.*, 1999, **146**, pp. 347–353
16 YEH, C.C., and BARRY, J.R.: 'Approximate minimum bit-error rate equalization for binary signaling'. Proceedings of ICC'97, Montreal, Canada, 1997, **2**, pp. 1095–1099
17 YEH, C.C., and BARRY, J.R.: 'Approximate minimum bit-error rate equalization for pulse-amplitude and quadrature-amplitude modulation'. Proceedings of ICC'98, 1998, Atlanta, USA, **1**, pp. 16–20
18 GIBSON, G.J., SIU, S., and COWAN, C.F.N.: 'The application of nonlinear structures to the reconstruction of binary signals', *IEEE Trans. Signal Process.*, 1991, **39**, (8), pp. 1877–1884
19 CORTES, C., and VAPNIK, V.: 'Support vector networks', *Mach. Learn.*, 1995, **20**, pp. 273–297
20 VAPNIK, V., GOLOWICH, S., and SMOLA, A.: 'Support vector method for function approximation, regression estimation, and signal processing', *in* MOZER, M., JORDAN, M., and PETSCHE, T. (Eds.): 'Neural information processing systems' (MIT Press, Cambridge, MA, **9**, 1997)
21 GUNN, S.: 'Support vector machines for classification and regression' Technical report, ISIS research group, Department of Electronics and Computer Science, University of Southampton, UK, May 1998
22 MINOUX, M.: 'Mathematical programming: Theory and algorithms' (Wiley, Chichester, 1986)
23 BAZARAA, M.S., SHERALI, H.D., and SHETTY, C.M.: 'Nonlinear programming: Theory and algorithms' (Wiley, New York, 1993)
24 CHEN, S., MCLAUGHLIN, S., MULGREW, B., and GRANT, P.M.: 'Bayesian decision feedback equaliser for overcoming co-channel interference', *IEE Proc. Commun.*, 1996, **143**, (4), pp. 219–225
25 ILTIS, R.A.: 'A randomized bias technique for the importance sampling simulation of Bayesian equalizers', *IEEE Trans. Commun.*, 1995, **43**, (2/3/4), pp. 1107–1115
26 HAYKIN, S.: 'Adaptive filter Theory' (Prentice Hall, Englewood Cliffs, NJ, 1991, 2nd edn.)

5 Appendix

Consider the problem of separating the set χ of N training data belonging to two separate classes

$$(y_1, \boldsymbol{x}_l), \ldots, (y_N, \boldsymbol{x}_N) \tag{36}$$

with a hyperplane

$$\boldsymbol{w}^T\boldsymbol{x} + c = 0 \tag{37}$$

where $\boldsymbol{x}_l \in R^m$ is an m-dimensional training data vector and $y_l \in \{-1, +1\}$ its class indicator. It is assumed that χ is linearly separable. As there is some redundancy in the parameters of the hyperplane (eqn. 37), it is appropriate to consider a canonical hyperplane [11], where the parameters $\boldsymbol{w}$, c are constrained by

$$\min_{x_l \in \chi} |\boldsymbol{w}^T\boldsymbol{x}_i + c| = 1 \tag{38}$$

This constraint is preferred to other alternatives, as it simplifies the formulation of the problem. A canonical separating hyperplane must satisfy the constraints

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + c) \geq 1, \ \ \forall \boldsymbol{x}_i \in \chi \tag{39}$$

Referring to Fig. 7, there are innumerable hyperplanes which can correctly separate χ into two classes. Intuitively, the best hyperplane is the one with the following property: the distance between the closest training vector to the hyperplane is maximal. Based on this belief, the optimal hyperplane is given by maximising the margin, which is defined as

$$\begin{aligned}\rho(\boldsymbol{w}, c) &= \min_{\{\boldsymbol{x}_i|y_i=+1\}} \frac{|\boldsymbol{w}^T\boldsymbol{x}_i + c|}{\|\boldsymbol{w}\|} + \min_{\{\boldsymbol{x}_j|y_j=-1\}} \frac{|\boldsymbol{w}^T\boldsymbol{x}_j + c|}{\|\boldsymbol{w}\|} \\ &= \frac{1}{\|\boldsymbol{w}\|}\left(\min_{\{\boldsymbol{x}_i|y_i=+1\}} |\boldsymbol{w}^T\boldsymbol{x}_i + c| + \min_{\{\boldsymbol{x}_j|y_j=-1\}} |\boldsymbol{w}^T\boldsymbol{x}_j + c|\right) \\ &= \frac{2}{\|\boldsymbol{w}\|}\end{aligned} \tag{40}$$

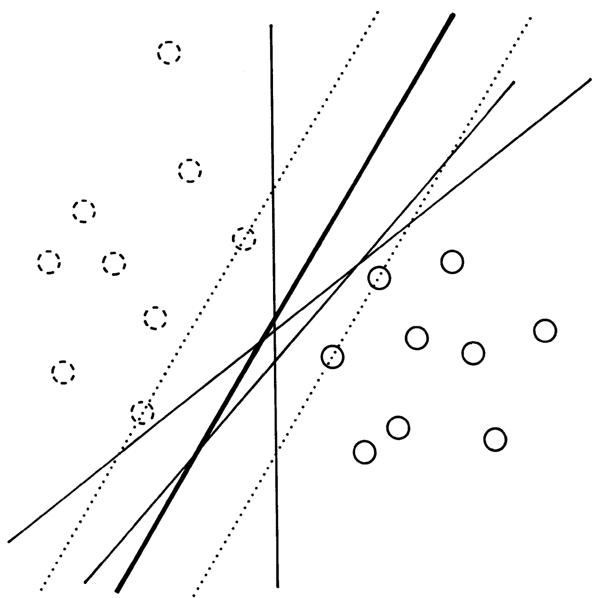

Fig. 7 *Optimal and nonoptimal separating hyperplanes*
—— optimal
—— nonoptimal
····· margin

subject to the constraints of expr. 39. Thus, the optimal hyperplane is the one that minimises

$$\Phi(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2 \quad (41)$$

subject to the constraints of expr. 39. It is interesting to notice that minimising the function of eqn. 41 is in fact equivalent to implementing the powerful SRM principle [11, 21].

The solution of the optimisation problem with the cost function of eqn. 41 under the constraints of expr. 39 is given by the saddle point of the Lagrangian [22, 23]

$$L(\boldsymbol{w}, c, \boldsymbol{g}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{N} g_i[(\boldsymbol{w}^T\boldsymbol{x}_i + c)y_i - 1] \quad (42)$$

where g_i are the Lagrange multipliers. The Lagrangian has to be minimised with respect to $\boldsymbol{w}$, c and maximised with respect to $g_i \geq 0$. Classical Lagrangian duality enables the primal problem (eqn. 42) to be transformed to its dual problem

$$\max_{\boldsymbol{g}} \psi(\boldsymbol{g}) = \max_{\boldsymbol{g}}\{\min_{\boldsymbol{w},c} L(\boldsymbol{w}, c, \boldsymbol{g})\} \quad (43)$$

The minimum with respect to $\boldsymbol{w}$ and c of the Lagrangian L is given by

$$\frac{\partial L}{\partial c} = 0 \Rightarrow \sum_{i=1}^{N} g_i y_i = 0 \quad (44)$$

$$\frac{\partial L}{\partial \boldsymbol{w}} = 0 \Rightarrow \boldsymbol{w} = \sum_{i=1}^{N} g_i y_i \boldsymbol{x}_i \quad (45)$$

Substituting eqn. 45 into eqn. 43 yields

$$\max_{\boldsymbol{g}} \psi(\boldsymbol{g}) = \max_{\boldsymbol{g}}\left\{-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} g_i g_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j + \sum_{i=1}^{N} g_i\right\} \quad (46)$$

and the solution of the dual problem is given by

$$\bar{\boldsymbol{g}} = \arg\min_{\boldsymbol{g}}\left\{\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} g_i g_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j - \sum_{i=1}^{N} g_i\right\} \quad (47)$$

with the constraints

$$g_i \geq 0, \; i = 1, \cdots, N \quad (48)$$

$$\sum_{i=1}^{N} g_i y_i = 0 \quad (49)$$

Solving the quadratic optimisation problem of eqn. 47 subject to the constraints of expr. 48 and eqn. 49 determines the Lagrange multipliers, and the optimal separating hyperplane is given by

$$\bar{\boldsymbol{w}} = \sum_{i=1}^{N} \bar{g}_i y_i \boldsymbol{x}_i \quad (50)$$

$$\bar{c} = -\frac{1}{2}\bar{\boldsymbol{w}}^T(\boldsymbol{x}_+ + \boldsymbol{x}_-) \quad (51)$$

where $\boldsymbol{x}_+$ and $\boldsymbol{x}_-$ are any 'support vector' from each class with the corresponding Lagrange multipliers and class indicators satisfying

$$\bar{g}_+ > 0, \; y_+ = +1; \; \bar{g}_- > 0, \; y_- = -1 \quad (52)$$

Notice that $\bar{c}$ is simply obtained from $y_+(\bar{\boldsymbol{w}}^T\boldsymbol{x}_+ + c) = 1$ and $y_-(\bar{\boldsymbol{w}}^T\boldsymbol{x}_- + c) = 1$.

From the Kuhn–Tucker conditions

$$\bar{g}_i(y_i[\bar{\boldsymbol{w}}^T\boldsymbol{x}_i + \bar{c}] - 1) = 0 \quad (53)$$

and only those points $\boldsymbol{x}_i$ which satisfy

$$y_i[\bar{\boldsymbol{w}}^T\boldsymbol{x}_i + \bar{c}] = 1 \quad (54)$$

will have nonzero Lagrange multipliers. These points are termed SVs. All the SVs lie on the margin and the number of SVs can be very small. The optimal hyperplane is determined by the SVs, and all the other points in the training set χ can be removed without affecting the solution. Finally, the following observation can be made. In the special case where the hyperplane passes through the origin of the space, $c = 0$ and the solution of the quadratic programming (eqn. 47) under the constraints of expr. 48 only can be computed very efficiently.