Decision Feedback Equalizer Design Using Support Vector Machines

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

Department of Electronics and Computer Science
University of Southampton, Highfield
Southampton SO17 1BJ, U.K.

Abstract

We consider the conventional decision feedback equalizer (DFE) that employs a linear combination of the channel observations and the past decisions. 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. We propose a strategy for designing the DFE based on support vector machines (SVM). 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 much more efficiently.

Keywords. Decision feedback equalizer, support vector machines, mean square error, bit error rate.

1 Introduction

Equalization 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 in practice 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

*Contact author. Tel./Fax: +44 (0)23 8059 6660/4508; E-mail: sqc@ecs.soton.ac.uk

investigated other DFE structures that employs nonlinear combinations of the channel observations and the past decisions [5]–[10]. These non-conventional 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 will 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, as 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 equalization. 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.

A better design in terms of performance is to choose the equalizer coefficients to minimize 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 equalizers without decision feedback [16],[17]. This approximate MBER solution was derived for the special case of equalizable channels. Equalizability 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 equalizer 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, due to the nature of minimizing BER.

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

In this study, we apply SVM 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 SVM becomes a simple matter, and the SVM solution is obtained by maximizing 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 The decision feedback equalizer 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

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

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

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

where σ_s^2 is the symbol variance. In this study, for the purpose of easy geometrical visualisation, we will 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 equalization process defined in Fig.1 uses the information present in the channel observation vector

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

and the past detected symbol vector

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

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 quantizing the filter output

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

where

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

are the coefficients of the feedforward and feedback filters, respectively. Without the loss of generality, we will 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 (5). The observation vector (3) can be arranged as

$$\mathbf{r}(k) = \mathbf{F} \; \mathbf{s}(k) + \mathbf{e}(k) \,, \tag{7}$$

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

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

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

$$\mathbf{F} = [\mathbf{F}_1 \ \mathbf{F}_2] \tag{9}$$

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

$$\mathbf{F}_1 = \begin{bmatrix} a_0 & a_1 & \cdots & a_{n_a-1} \\ 0 & a_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_1 \\ 0 & \cdots & 0 & a_0 \end{bmatrix} \tag{10}$$

and

$$\mathbf{F}_2 = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ a_{n_a-1} & 0 & \ddots & \vdots \\ a_{n_a-2} & a_{n_a-1} & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 \\ a_1 & \cdots & a_{n_a-2} & a_{n_a-1} \end{bmatrix}, \tag{11}$$

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

$$\mathbf{r}(k) = \mathbf{F}_1 \ \mathbf{s}_f(k) + \mathbf{F}_2 \ \hat{\mathbf{s}}_b(k) + \mathbf{e}(k)\,. \tag{12}$$

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

$$\mathbf{r}^{'}(k) \triangleq \mathbf{r}(k) - \mathbf{F}_2 \ \hat{\mathbf{s}}_b(k)\,, \tag{13}$$

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

$$f^{'}(\mathbf{r}^{'}(k)) = \mathbf{w}^T \mathbf{r}^{'}(k) \tag{14}$$

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

$$\left.\begin{array}{l} r^{'}(k-i) = z^{-1} r^{'}(k-i+1) - a_{n_a-i}\hat{s}(k-d-1), \quad i = m-1, \cdots, 2, 1 \\ r^{'}(k) = r(k) \end{array}\right\}, \tag{15}$$

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

By considering the alternative but equivalent DFE structure of Fig. 2, the geometric insight of the equalization process becomes evident: the DFE design is basically to choose the weight vector $\mathbf{w}$ of the hyperplane $\{\mathbf{r} : \mathbf{w}^T \mathbf{r} = 0\}$. 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 $\mathbf{s}_f(k)$ be $\mathbf{s}_{f,j}$, $1 \le j \le N_f$. The set of the noiseless channel states in the translated space is defined by:

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

This set can be partitioned into two subsets conditioned on $s(k-d) = \pm 1$:

$$\mathcal{R}^{\pm} \triangleq \{\mathbf{r}_j \in \mathcal{R} \mid s(k-d) = \pm 1\} . \tag{17}$$

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

Proof: Choose the weight vector of a hyperplane $\mathbf{w}^T\mathbf{r} = 0$ to be: $\mathbf{w} = \left[0 \cdots 0 \, \frac{1}{a_0}\right]^T$. For any $\mathbf{r}^+ \in \mathcal{R}^+$ and $\mathbf{r}^- \in \mathcal{R}^-$, we have $\mathbf{w}^T\mathbf{r}^+ = 1 > 0$ and $\mathbf{w}^T\mathbf{r}^- = -1 < 0$.

2.1 The MMSE decision feedback equalizer

From the above discussion, 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 minimizing $\mathrm{E}[(\mathbf{w}^T\mathbf{r}'(k) - s(k-d))^2]$, where $\mathrm{E}[\cdot]$ denotes the expectation operation, is given by

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

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

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

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

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

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

Lemma 2

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

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

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

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 (18) will rotate and is no longer orthogonal to the last axis of the translated observation space. Consider the channel 1 of (22) again. Given SNR=15 dB, the MMSE hyperplane has a slope of -0.27, which is clearly non-optimal. In fact, the numerical solution of the MBER DFE in this case gives a slope of -1.02.

2.2 The MBER decision feedback equalizer

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

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

where

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

The MBER DFE is obtained by minimizing $P_E(\mathbf{w})$:

$$\mathbf{w}_{\text{MBER}} = \arg \min_{\mathbf{w}} P_E(\mathbf{w}) \,. \tag{25}$$

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

Obviously, $\mathbf{w}_{\text{MBER}}$ is optimal for the linear-combiner DFE structure discussed here. Unlike the MMSE solution, there is no closed-form expression for $\mathbf{w}_{\text{MBER}}$ and the solution must be sought via some gradient-based iterative procedure. In general, the BER surface

$P_E(\mathbf{w})$ 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 equalizer [16],[17] has a complexity per sample similar to the LMS algorithm but it requires a very long training sequence.

2.3 The SVM decision feedback equalizer

SVM 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 the completeness, SVM for the linearly separable classification case are described in Appendix. The minimum distance from the nearest point in $\mathcal{R}$ to a separating hyperplane $\mathbf{w}^T\mathbf{r} = 0$

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

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

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

Define the integer set

$$I_{\mathcal{R}} \triangleq \{i \mid \mathbf{r}_i \in \mathcal{R}\}\,, \tag{28}$$

and the class indicator

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

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

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

where

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

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

The optimization problem (31) with (32) is a standard quadratic programming, whose solution $\bar{\mathbf{g}}$ 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]

$$\bar{g}_i \left(y_i \mathbf{w}_{\mathrm{SVM}}^T \mathbf{r}_i - 1 \right) = 0\,, \tag{33}$$

and only those points $\mathbf{r}_i$, which satisfy $y_i \mathbf{w}_{\mathrm{SVM}}^T \mathbf{r}_i = 1$, will have non-zero 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 $\mathcal{R}_{\mathrm{SV}}$ be the set of SVs. The hyperplane $\mathbf{w}_{\mathrm{SVM}}$ is determined by $\mathcal{R}_{\mathrm{SV}}$ only. Thus, the identical solution is obtained by substituting $\mathcal{R}$ with $\mathcal{R}_{\mathrm{SV}}$ in (30) to (32) with massive saving in computational efforts. However, SVs are unknown *a priori*. Let us consider how to select a smaller subset $\mathcal{R}_{\mathrm{Sub}}$ of $\mathcal{R}$, which contains all the SVs. We first point out that, as we restrict to the linear structure (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 equalizer 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.

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

Subset Selection Algorithm

As $\mathcal{R}_{\text{Sub}}$ can be a much smaller subset of $\mathcal{R}$, using it to substitute $\mathcal{R}$ in the quadratic programming (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 $\mathbf{w}_{\text{SVM}}$ does not depend on the noise variance σ_e^2 and it is the asymptotic MBER solution, that is, $\mathbf{w}_{\text{SVM}} \rightarrow \mathbf{w}_{\text{MBER}}$ as SNR$\rightarrow \infty$. In general, $\mathbf{w}_{\text{SVM}}$ will not be identical to $\mathbf{w}_{\text{MBER}}$ but the difference is practically negligible for useful SNR conditions. Consider the channel 1 of (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, $\mathbf{w}_{\text{SVM}}$ remains unchanged with a slope of -1 but $\mathbf{w}_{\text{MBER}}$ 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 (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 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:

$$\text{Channel 2}: \ \mathbf{a} = [0.35\ 0.80\ 1.00\ 0.80]^T . \tag{34}$$

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

$$\text{Channel 3}: \ \mathbf{a} = [0.227\ 0.466\ 0.688\ 0.466\ 0.227]^T . \tag{35}$$

The structural parameters of the DFE were set to $d = 4$, $m = 5$ and $n = 4$. The full set $\mathcal{R}$ contains thirty-two states. The subset $\mathcal{R}_{\text{Sub}}$ used in computing the SVM solution has eighteen 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.

2.5 Adaptive implementation

An indirect adaptation scheme is suitable for adaptive implementation of the SVM DFE. The scheme first estimates a channel model $\hat{\mathbf{a}}$ using for example the LMS algorithm and then computes the weight vector $\mathbf{w}_{\text{SVM}}$ of the SVM DFE based on the channel estimate $\hat{\mathbf{a}}$. This indirect approach, as opposed to a direct adaptation of the equalizer 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 equalizer 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 disk drives, the equalizers 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 organized 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 optimization 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 realize the SVM DFE in data storage systems and slow time-varying communication links.

References

[1] S.U.H. Qureshi, "Adaptive equalization," *Proc. IEEE*, Vol.73, No.9, pp.1349–1387, 1985.

[2] J.G. Proakis, *Digital Communications.* 3rd edition, New York: McGraw-Hill, 1995.

[3] J. Moon, "The role of SP in data-storage systems," *IEEE Signal Processing Magazine*, Vol.15, No.4, pp.54–72, 1998.

[4] J.G. Proakis, "Equalization techniques for high-density magnetic recording," *IEEE Signal Processing Magazine*, Vol.15, No.4, pp.73–82, 1998.

[5] S. Siu, G.J. Gibson and C.F.N. Cowan, "Decision feedback equalisation using neural network structures and performance comparison with the standard architecture," *IEE Proc., Pt.I*, Vol.137, No.4, pp.221–225, 1990.

[6] D. Williamson, R.A. Kennedy and G.W. Pulford, “Block decision feedback equalization,” *IEEE Trans. Communications*, Vol.40, No.2, pp.255-264, 1992.

[7] S. Chen, B. Mulgrew and S. McLaughlin, “Adaptive Bayesian equaliser with decision feedback,” *IEEE Trans. Signal Processing*, Vol.41, No.9, pp.2918–2927, 1993.

[8] S. Chen, S. McLaughlin and B. Mulgrew, “Complex-valued radial basis function network, Part II: application to digital communications channel equalisation,” *EURASIP Signal Processing J.*, Vol.36, pp.175–188, 1994.

[9] S. Chen, S. McLaughlin, B. Mulgrew and P.M. Grant, “Adaptive Bayesian decision feedback equaliser for dispersive mobile radio channels,” *IEEE Trans. Communications*, Vol.43 No.5, pp.1937–1946, 1995.

[10] I. Cha and S.A. Kassam, “Channel equalization using adaptive complex radial basis function networks,” *IEEE J. Selected Areas in Communications*, Vol.13, No.1, pp.122–131, 1995.

[11] V. Vapnik, *The Nature of Statistical Learning Theory.* New York: Springer-Verlag, 1995.

[12] S. Chen, B. Mulgrew, E.S. Chng and G. Gibson, “Space translation properties and the minimum-BER linear-combiner DFE,” *IEE Proc. Communications*, Vol.145, No.5, pp.316–322, 1998.

[13] E. Shamash and K. Yao, “On the structure and performance of a linear decision feedback equalizer based on the minimum error probability criterion,” in *Proc. ICC’74*, 1974, pp.25F1–25F5.

[14] S. Chen, E.S. Chng, B. Mulgrew and G. Gibson, “Minimum-BER linear-combiner DFE,” in *Proc. ICC’96* (Dallas, Texas), 1996, Vol.2, pp.1173–1177.

[15] S. Chen and B. Mulgrew, “The minimum-SER linear-combiner decision feedback equalizer,” *IEE Proc. Communications*, accepted, 1999.

[16] C.C. Yeh and J.R. Barry, “Approximate minimum bit-error rate equalization for binary signaling,” in *Proc. ICC’97* (Montreal, Canada), 1997, Vol.2, pp.1095-1099.

[17] C.C. Yeh and J.R. Barry, "Approximate minimum bit-error rate equalization for pulse-amplitude and quadrature-amplitude modulation," in *Proc. ICC'98* (Atlanta, USA), 1998, Vol.1, pp.16-20.

[18] G.J. Gibson, S. Siu and C.F.N. Cowan, "The application of nonlinear structures to the reconstruction of binary signals," *IEEE Trans. Signal Processing*, Vol.39, No.8, pp.1877–1884, 1991.

[19] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, Vol.20, pp.273–297, 1995.

[20] V. Vapnik, S. Golowich and A. Smola, "Support vector method for function approximation, regression estimation, and signal processing," in: M. Mozer, M. Jordan and T. Petsche, eds., *Neural Information Processing Systems*, Vol.9, Cambridge, MA: MIT Press, 1997.

[21] S. Gunn, "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] M. Minoux, *Mathematical Programming: Theory and Algorithms.* Chichester: John Wiley and Sons, 1986.

[23] M.S. Bazaraa, H.D. Sherali and C.M. Shetty, *Nonlinear Programming: Theory and Algorithms.* New York: John Wiley, 1993.

[24] S. Chen, S. McLaughlin, B. Mulgrew and P.M. Grant, "Bayesian decision feedback equaliser for overcoming co-channel interference," *IEE Proc. Communications*, Vol.143, No.4, pp.219–225, 1996.

[25] R.A. Iltis, "A randomized bias technique for the importance sampling simulation of Bayesian equalizers," *IEEE Trans. Communications*, Vol.43, No.2/3/4, pp.1107–1115, 1995.

[26] S. Haykin, *Adaptive Filter Theory.* 2nd edition, Englewood Cliffs, NJ: Prentice Hall, 1991.

Appendix

Consider the problem of separating the set $\mathcal{X}$ of N training data belonging to two separate classes,

$$(y_1, \mathbf{x}_1), \cdots, (y_N, \mathbf{x}_N)\,, \tag{36}$$

with a hyperplane

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

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

$$\min_{\mathbf{x}_i \in \mathcal{X}} |\mathbf{w}^T\mathbf{x}_i + c| = 1\,. \tag{38}$$

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

$$y_i\left(\mathbf{w}^T\mathbf{x}_i + c\right) \geq 1,\ \forall \mathbf{x}_i \in \mathcal{X}\,. \tag{39}$$

Referring to Fig. 7, there are innumerable hyperplanes which can correctly separate $\mathcal{X}$ 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 maximizing the margin, which is defined as

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

subject to the constraints (39). Thus, the optimal hyperplane is the one that minimizes

$$\Phi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2\,, \tag{41}$$

subject to the constraints (39). It is interesting to notice that minimizing the function (41) is in fact equivalent to implementing the powerful SRM principle [11],[21].

The solution of the optimization problem with the cost function (41) and under the constraints (39) is given by the saddle point of the Lagrangian [22],[23]

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

where g_i are the Lagrange multipliers. The Lagrangian has to be minimized with respect to (w.r.t.) $\mathbf{w}, c$ and maximized w.r.t. $g_i \geq 0$. Classical Lagrangian duality enables the primal problem (42) to be transformed to its dual problem:

$$\max_{\mathbf{g}} \Psi(\mathbf{g}) = \max_{\mathbf{g}} \left\{ \min_{\mathbf{w},c} L(\mathbf{w}, c, \mathbf{g}) \right\} . \tag{43}$$

The minimum w.r.t. $\mathbf{w}$ and c of the Lagrangian, L, is given by

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

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{N} g_i \, y_i \, \mathbf{x}_i . \tag{45}$$

Substituting (45) into (43) yields

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

and the solution of the dual problem is given by

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

with the constraints

$$g_i \geq 0, \;\; i = 1, \cdots, N , \tag{48}$$

$$\sum_{i=1}^{N} g_i \, y_i = 0 . \tag{49}$$

Solving the quadratic optimization problem (47) subject to the constraints (48) and (49) determines the Lagrange multipliers, and the optimal separating hyperplane is given by

$$\bar{\mathbf{w}} = \sum_{i=1}^{N} \bar{g}_i \, y_i \, \mathbf{x}_i , \tag{50}$$

$$\bar{c} = -\frac{1}{2} \bar{\mathbf{w}}^T \left(\mathbf{x}_+ + \mathbf{x}_- \right) , \tag{51}$$

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

$$\bar{g}_+ > 0, \;\; y_+ = +1; \;\; \bar{g}_- > 0, \;\; y_- = -1 . \tag{52}$$

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

From the Kuhn-Tucker conditions

$$\bar{g}_i \left(y_i \left[\bar{\mathbf{w}}^T \mathbf{x}_i + \bar{c} \right] - 1 \right) = 0 \,, \tag{53}$$

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

$$y_i \left[\bar{\mathbf{w}}^T \mathbf{x}_i + \bar{c} \right] = 1 \tag{54}$$

will have non-zero 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 $\mathcal{X}$ 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 (47) under the constraints (48) only can be computed very efficiently.

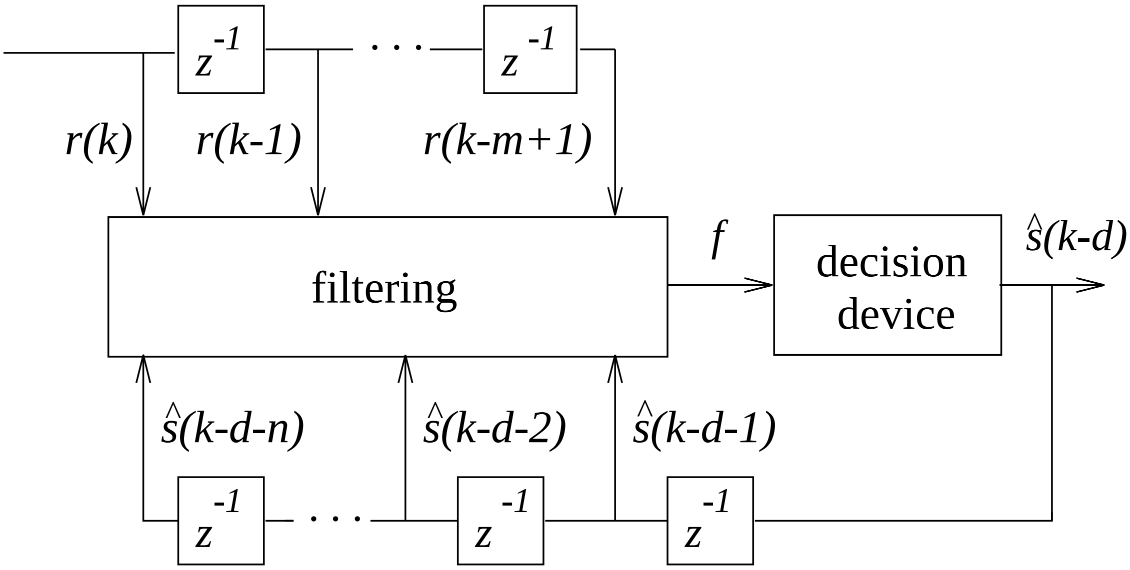

Figure 1: Schematic of a generic decision feedback equalizer.

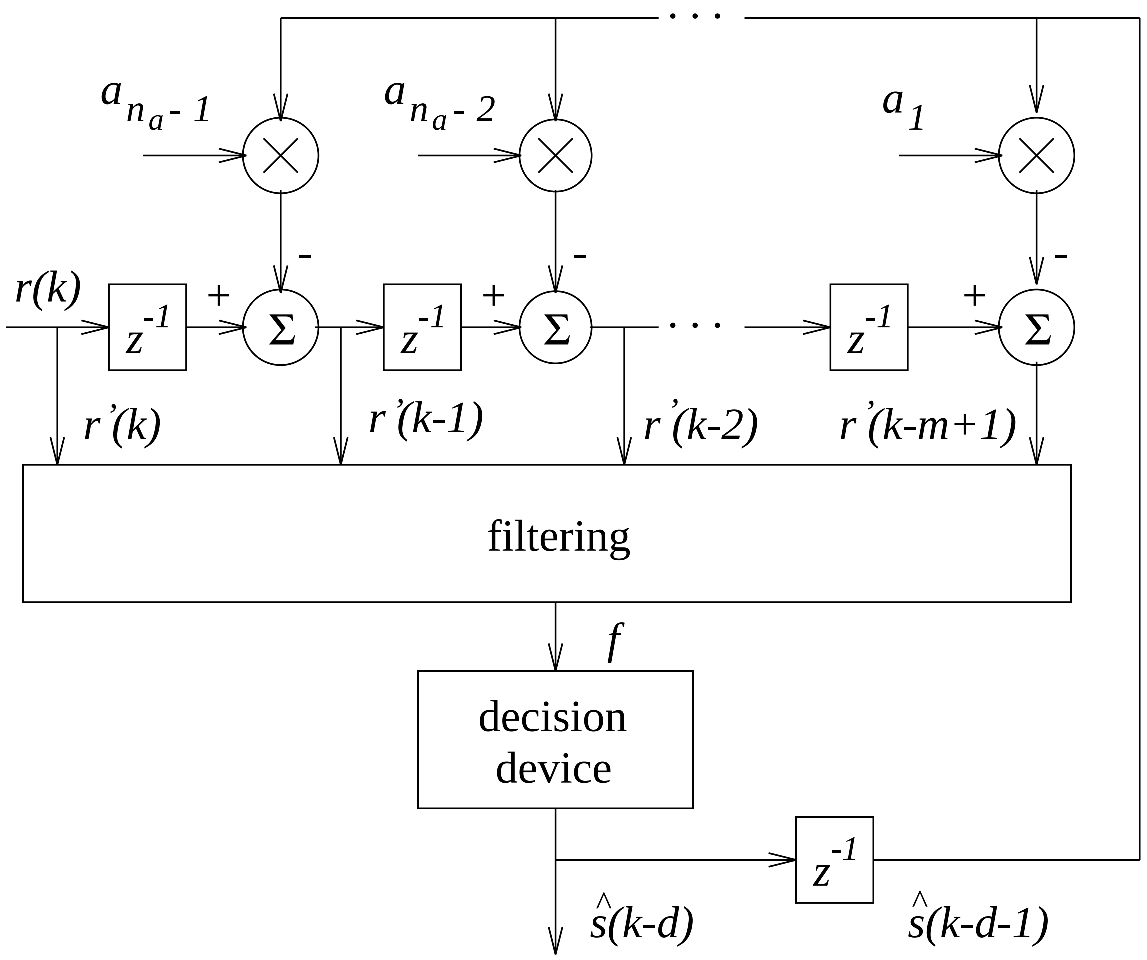

Figure 2: Schematic of the translated decision feedback equalizer.

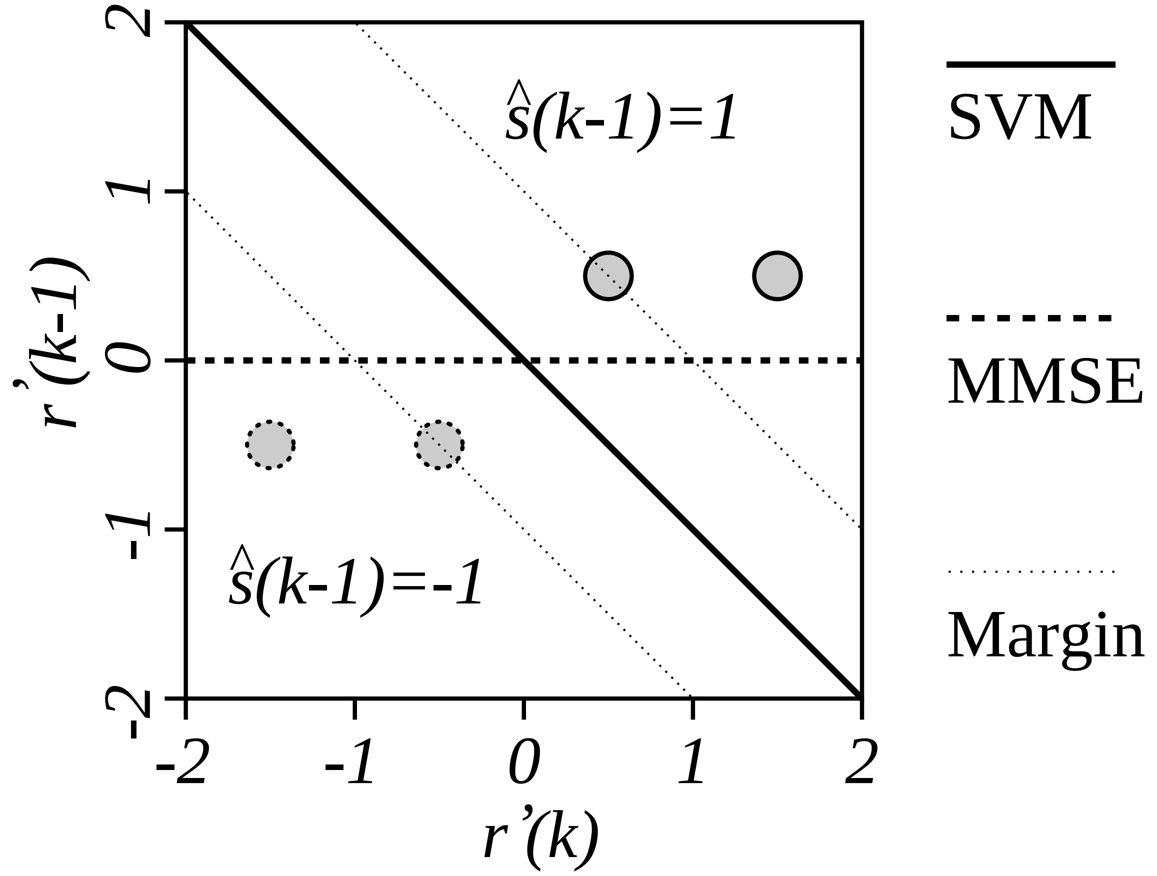

Figure 3: Two asymptotic decision hyperplanes corresponding to large SNR for channel $\mathbf{a} = [0.5\ 1.0]^T$. Circles denote the two subsets of states $\mathcal{R}^-$ and $\mathcal{R}^+$.

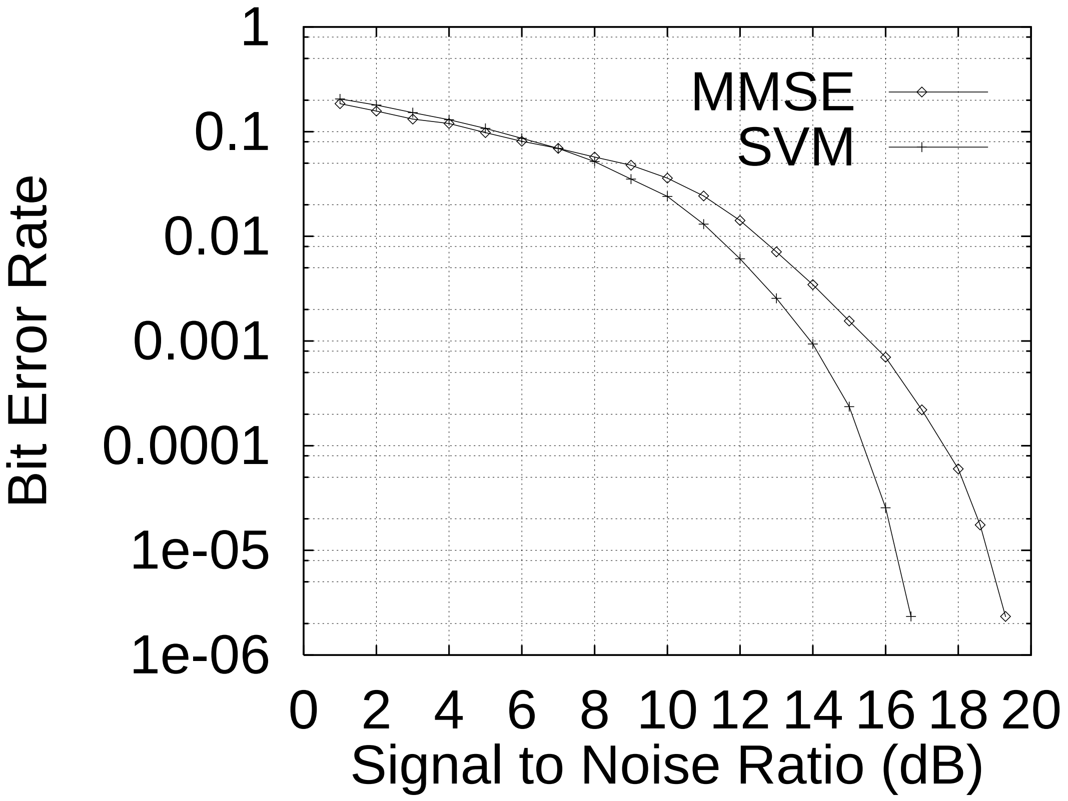

Figure 4: Performance comparison for channel $\mathbf{a} = [0.5\ 1.0]^T$ with detected symbols being fed back. MMSE/SVM: MMSE/SVM DFEs.

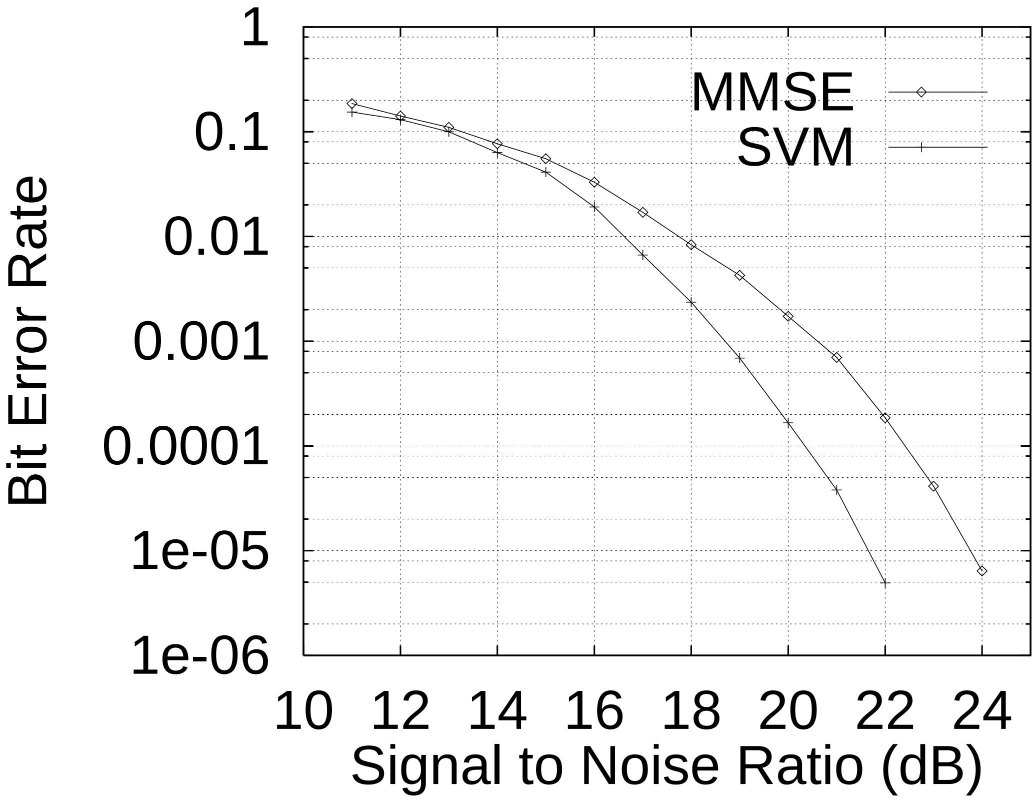

Figure 5: Performance comparison for channel $\mathbf{a} = [0.35\ 0.80\ 1.00\ 0.80]^T$ with detected symbols being fed back. MMSE/SVM: MMSE/SVM DFEs.

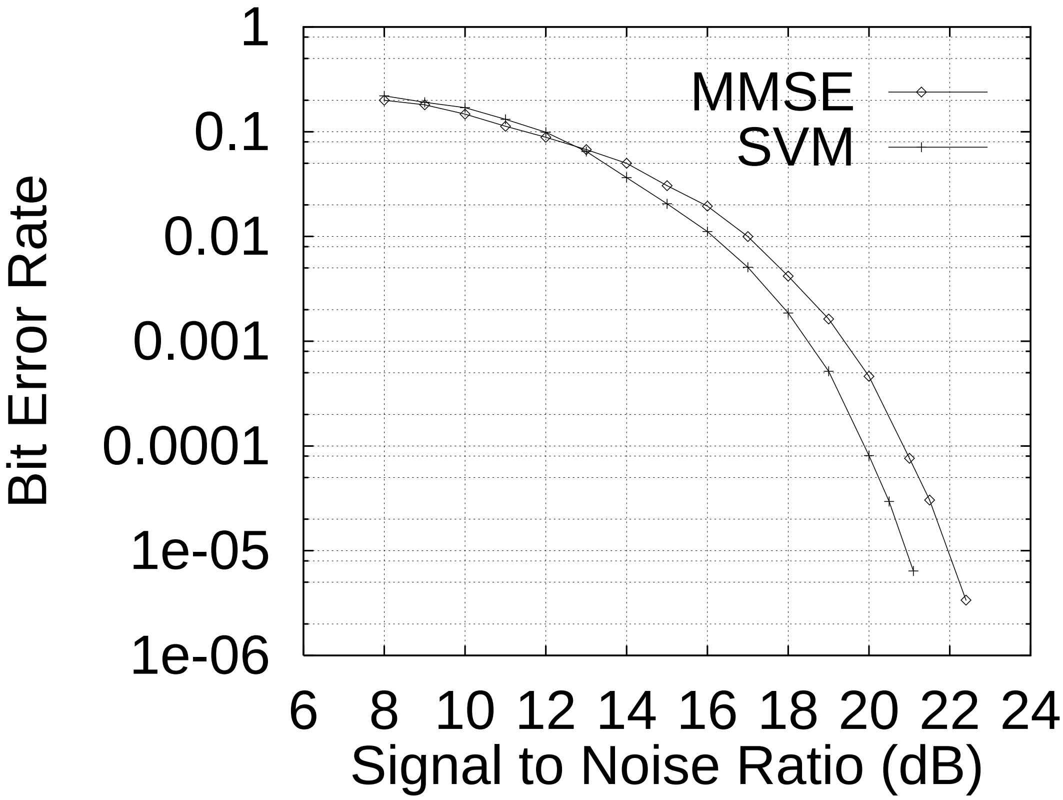

Figure 6: Performance comparison for channel $\mathbf{a} = [0.227\ 0.466\ 0.688\ 0.466\ 0.227]^T$ with detected symbols being fed back. MMSE/SVM: MMSE/SVM DFEs.

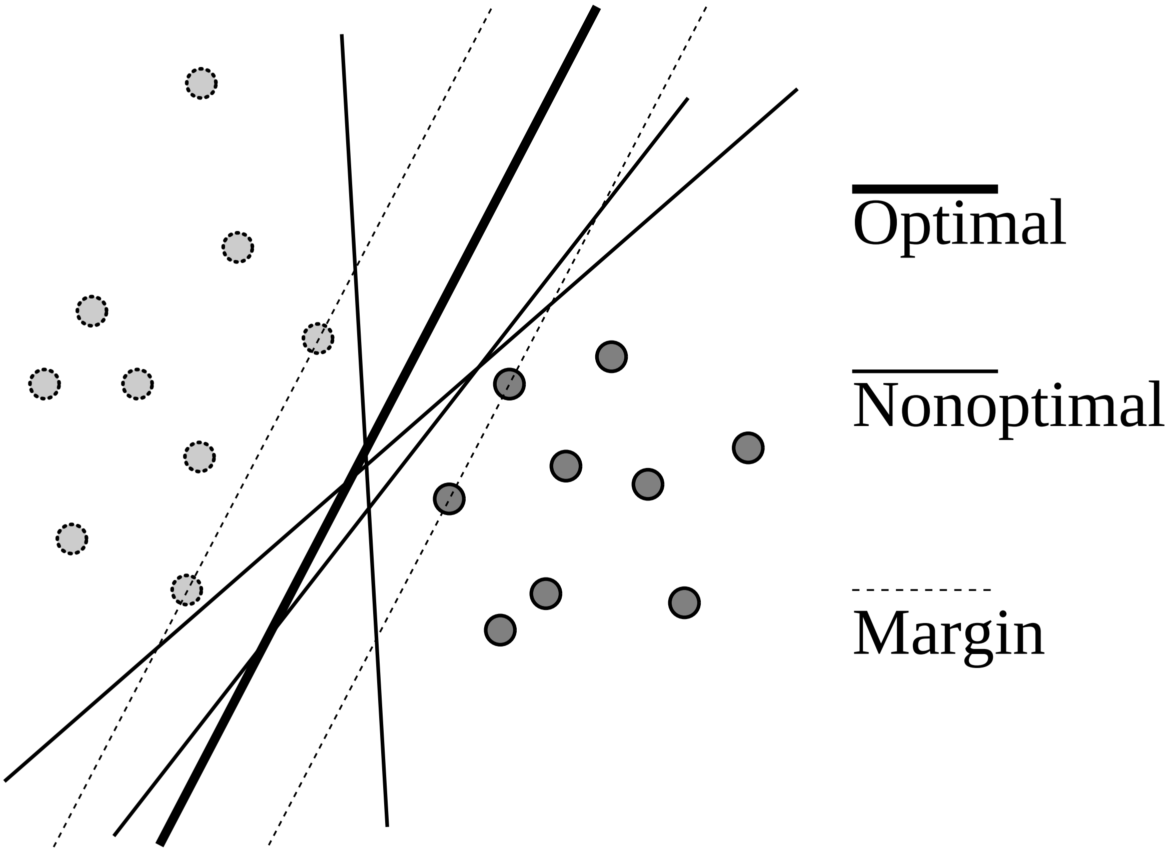

Figure 7: Optimal and non-optimal separating hyperplanes.

List of figures