Adaptive Multiuser Receiver Using a Support Vector Machine Technique

S. Chen, A.K. Samingan and L. Hanzo

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

Abstract— **The paper investigates the application of an emerging learning technique, called support vector machines (SVMs), to construct an adaptive nonlinear multiuser detector (MUD) for direct-sequence code-division multiple-access (DS-CDMA) signals transmitted through multipath channels. Computer simulation is used to study this adaptive SVM MUD, and the results show that it can closely match the performance of the optimal Bayesian one-shot detector, using a relatively small training data block.**

I. Introduction

DS-CDMA constitutes an attractive multiuser scheme that allows users to transmit at the same carrier frequency. However, this creates multiuser interference which, if not controlled, can seriously degrade the quality of reception. For the downlink scenario, the linear minimum mean square error (MMSE) multiuser detector (MUD) [1]–[5] is widely used, as its adaptive implementation is very simple. The linear MUD, however, can only work when the underlying noise-free signal classes are linearly separable. As nonlinear separable cases are common in DS-CDMA channels, neural networks have been considered as nonlinear MUDs [6]–[9]. Training times for these nonlinear MUDs, however, are often long and unpredictable. Furthermore, the structures of these neural network MUDs are usually determined by trial and error.

A learning technique known as the support vector machines SVM has gained popularity due to its many attractive features and promising empirical performance [10]–[12]. For a brief introduction to SVMs please refer to the Appendix. For binary classification tasks, the SVM approach nonlinearly maps the input space into a high dimensional feature space via simple kernel representations. In the high dimensional feature space, a linear classifier with maximum margin is constructed. Apart from good generalisation properties, the learning process of SVMs is intriguing. A SVM classifier is determined only by a sparse set of support vectors (SVs), and these SVs are automatically selected from the training data during the learning process.

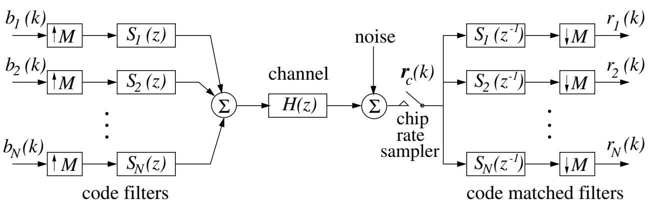

Fig. 1. Discrete-time model of synchronous CDMA downlink.

Since the idea of SVMs originates from finding an optimum hyperplane for separating two classes with maximum margin, it is also very relevant to multiuser detection in DS-CDMA. In this paper, the SVM technique is investigated as an adaptive nonlinear MUD. Our study shows that a SVM-based MUD trained using a relatively small block of noisy received signal samples can closely approximate the performance of the optimal MUD [13], which requires a complete knowledge of the system in terms of the so-called system matrix $\mathbf{P}$ to be introduced during our further discourse in (3) and the noise variance. Another advantage of the SVM approach over the existing nonlinear MUDs is an automatic determination of the detector structure. The main drawback of the SVM method is that it is a block-data based method.

II. System model

The discrete-time model of the synchronous DS-CDMA system supporting N users and transmitting M $(> N)$ chips per bit is depicted in Fig. 1, where $b_i(k) \in \{\pm 1\}$ denotes the k-th bit of user i, the unit-length signature code sequence for user i is $\bar{\mathbf{s}}_i = [\bar{s}_{i,1} \cdots \bar{s}_{i,M}]^T$, and the transfer function associated with the channel's impulse response (CIR) is

$$H(z) = \sum_{i=0}^{n_h-1} h_i z^{-i}. \quad (1)$$

The bit vector of N users at instant k is $\mathbf{b}(k) = [b_1(k) \cdots b_N(k)]^T$, and the received signal vector after the chip-matched filters is $\mathbf{r}(k) = [r_1(k) \cdots r_N(k)]^T$. It can be shown that the baseband model for $\mathbf{r}(k)$ is:

$$\mathbf{r}(k) = \mathbf{P} \begin{bmatrix} \mathbf{b}(k) \\ \mathbf{b}(k-1) \\ \vdots \\ \mathbf{b}(k-L+1) \end{bmatrix} + \tilde{\mathbf{n}}(k), \quad (2)$$

where the $N \times LN$ system matrix is given by

$$\mathbf{P} = \bar{\mathbf{S}}^T \mathbf{H} \begin{bmatrix} \bar{\mathbf{S}}\mathbf{A} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{S}}\mathbf{A} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \bar{\mathbf{S}}\mathbf{A} \end{bmatrix}; \qquad (3)$$

the user signature sequence matrix $\bar{\mathbf{S}} = [\bar{\mathbf{s}}_1 \cdots \bar{\mathbf{s}}_N]$; the diagonal user signal amplitude matrix $\mathbf{A} = \mathrm{diag}\{A_1 \cdots A_N\}$; the $M \times LM$ CIR matrix $\mathbf{H}$ has the form

$$\mathbf{H} = \begin{bmatrix} h_0 & h_1 & \cdots & h_{n_h-1} & & & & \\ & h_0 & h_1 & \cdots & h_{n_h-1} & & & \\ & & \ddots & \ddots & \cdots & \ddots & & \\ & & & h_0 & h_1 & \cdots & h_{n_h-1} & \end{bmatrix}; \qquad (4)$$

and orthogonal code sequences are assumed, so that the noise vector $\tilde{\mathbf{n}}(k) = [\tilde{n}_1(k) \cdots \tilde{n}_N(k)]^T$ at the outputs of the chip-matched filters has a variance of $E[\tilde{\mathbf{n}}(k)\tilde{\mathbf{n}}^T(k)] = \sigma_n^2\mathbf{I}$. We note that the orthogonality of the codes is destroyed by the channel-induced intersymbol interference (ISI). The ISI span L depends on the length of the CIR, n_h, related to the length of the chip sequence, M. For $n_h = 1$, $L = 1$; for $1 < n_h \leq M$, $L = 2$; for $M < n_h \leq 2M$, $L = 3$; and so on.

III. Linear and optimal detectors

The linear MUD for user i has the form:

$$\hat{b}_i(k) = \mathrm{sgn}(y_L(k)) \ \text{ with } \ y_L(k) = \mathbf{w}^T\mathbf{r}(k), \qquad (5)$$

where $\mathbf{w} = [w_1 \cdots w_N]^T$ denotes the detector's weight vector. The most popular solution for the detector (5) is the MMSE solution given by

$$\mathbf{w}_{MMSE} = \left(\sigma_n^2\mathbf{I} + \mathbf{P}\mathbf{P}^T\right)^{-1}\mathbf{p}_i, \qquad (6)$$

where $\mathbf{p}_i$ denotes the i-th column of $\mathbf{P}$. The linear detector (5) is computationally very simple, and the standard LMS or RLS algorithms can be used to implement the MMSE solution adaptively.

However, a linear MUD only performs adequately in certain situations. Let the $N_b = 2^{LN}$ possible combinations of $[\mathbf{b}^T(k)\,\mathbf{b}^T(k-1)\cdots\mathbf{b}^T(k-L+1)]^T$ be

$$\mathbf{b}^{(j)} = \begin{bmatrix} \mathbf{b}^{(j)}(k) \\ \mathbf{b}^{(j)}(k-1) \\ \vdots \\ \mathbf{b}^{(j)}(k-L+1) \end{bmatrix}, \ 1 \leq j \leq N_b, \qquad (7)$$

and $b_i^{(j)}$ the ith element of $\mathbf{b}^{(j)}(k)$. Let us define the set of the N_b noise-free received signal states as

$$\mathcal{R} = \left\{\mathbf{r}_j = \mathbf{P}\mathbf{b}^{(j)},\ 1 \leq j \leq N_b\right\}, \qquad (8)$$

where $\mathcal{R}$ can be partitioned into two subsets:

$$\mathcal{R}_\pm = \left\{\mathbf{r}_j \in \mathcal{R} : b_i^{(j)} = \pm 1\right\}. \qquad (9)$$

If $\mathcal{R}_-$ and $\mathcal{R}_+$ are not linearly separable, a linear MUD will exhibit an irreducible error floor even in the noise-free case, as it can only form a hyperplane in the N-dimensional received signal space.

Applying the Bayesian classification theory in a manner similar to the channel equalization problem [14], it can be shown that the optimal detector has the form:

$$y_B(k) = f_B(\mathbf{r}(k)) = \sum_{j=1}^{N_b} \beta_j b_i^{(j)} \exp\left(-\frac{\|\mathbf{r}(k) - \mathbf{r}_j\|^2}{2\sigma_n^2}\right) \qquad (10)$$

with

$$\hat{b}_i(k) = \mathrm{sgn}(y_B(k)), \qquad (11)$$

where $b_i^{(j)} \in \{\pm 1\}$ serve as class labels, and all the channel states are assumed to be equiprobable with $\beta_j = \frac{1}{N_b(2\pi\sigma_n^2)^{\frac{m}{2}}}$.

IV. The support vector machine detector

The optimal detector obeying (10) requires the knowledge of all the noise-free signal sates $\mathbf{r}_j$, which are unknown to receiver i. In practice the receiver can have access to a block of K training samples $\{\mathbf{r}(k), b_i(k)\}_{k=1}^K$. Let us denote the training set of K noisy received signal vectors as

$$\mathcal{X} = \{\mathbf{x}_k = \mathbf{r}(k),\ 1 \leq k \leq K\} \qquad (12)$$

and the set of corresponding class labels as

$$\mathcal{C} = \{c_k = b_i(k),\ 1 \leq k \leq K\}. \qquad (13)$$

Applying the standard SVM method [10] (see Appendix for a tutorial), an SVM detector can be constructed for user i:

$$y_{SVM}(k) = \sum_{j=1}^{K} \bar{g}_j c_j F(\mathbf{r}(k), \mathbf{x}_j) + \bar{\eta}, \qquad (14)$$

where the set of Lagrangian multipliers $\{\bar{g}_j\}$, denoted in vectorial form as:

$$\bar{\mathbf{g}} = [\bar{g}_1 \cdots \bar{g}_K]^T, \qquad (15)$$

is the solution of the quadratic programming (QP)

$$\bar{\mathbf{g}} = \arg\min_{\mathbf{g}} \left\{ \frac{1}{2}\sum_{j=1}^{K}\sum_{l=1}^{K} g_j g_l c_j c_l F(\mathbf{x}_j, \mathbf{x}_l) - \sum_{j=1}^{K} g_j \right\} \qquad (16)$$

with the constraints

$$0 \leq g_j \leq C,\ \ 1 \leq j \leq K, \qquad (17)$$

and

$$\sum_{j=1}^{K} g_j c_j = 0. \tag{18}$$

In this application it was found advantageous to choose the Gaussian kernel function of:

$$F(\mathbf{x}_j, \mathbf{x}_l) = \exp\left(-\frac{\|\mathbf{x}_j - \mathbf{x}_l\|^2}{2\rho^2}\right), \tag{19}$$

where the width parameter ρ is related to the root mean square σ_n of the channel noise, an estimate of which can be obtained. The offset constant $\bar{\eta}$ is usually determined from the so-called "margin" SVs, i.e. from those particular $\mathbf{x}_j$s, for which the corresponding Lagrangian multipliers obey $0 < \bar{g}_j < C$. Because the optimal decision boundary, defined by $\{\mathbf{r} : f_B(\mathbf{r}) = 0\}$, passes through the origin of the received signal space and possesses certain symmetric properties due to the symmetric structure of $\mathcal{R}_-$ and $\mathcal{R}_+$, $\bar{\eta} = 0$ can be used. With this choice of the offset constant, the equality constraint (18) is no longer needed, and this leads to a simpler optimization task. The user-defined parameter C controls the trade-off between model complexity and training error. In our application, we will choose C empirically.

The set of SVs, denoted by $\mathcal{X}_{SVM}$, is given by those particular $\mathbf{x}_j$s, which have non-zero Lagrangian multipliers obeying $0 < \bar{g}_j \leq C$, where $\mathcal{X}_{SVM}$ is usually a small subset of the training data set $\mathcal{X}$. These SVs are determined during the optimization process. Thus the SVM-based MUD requires computing the decision variable

$$y_{SVM}(k) = \sum_{\mathbf{x}_j \in \mathcal{X}_{SVM}} \bar{g}_j c_j \exp\left(-\frac{\|\mathbf{r}(k) - \mathbf{x}_j\|^2}{2\rho^2}\right) \tag{20}$$

and making the decision according to:

$$\hat{b}_i(k) = \mathrm{sgn}(y_{SVM}(k)). \tag{21}$$

V. Simulation results

Two simulation examples were used for comparing the performance of the proposed SVM-based MUD to those of the linear MMSE and optimal MUDs. It is worth pointing out again that the linear MMSE MUD and the optimal MUD are designed based on the complete knowledge of the system - namely on that of the system matrix $\mathbf{P}$ and the noise variance - while the SVM MUD is trained using a block of the noisy received signal samples.

Example 1. A two-user system employing 4 chips per bit was constructed. The code sequences of the two users were $(+1, +1, -1, -1)$ and $(+1, -1, -1, +1)$, respectively, and the transfer function associated with the CIR was $H(z) = 0.3 + 0.7z^{-1} + 0.3z^{-2}$. The two users had equal signal power,

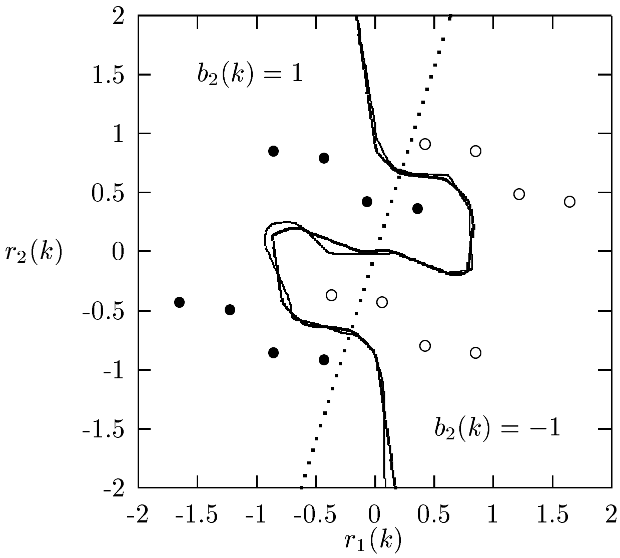

Fig. 2. The set of noise-free signal points and the three decision boundaries (dotted: linear MMSE, thick solid: optimal, thin solid: SVM) for user 2 of Example 1. $\mathrm{SNR}_1 = \mathrm{SNR}_2 = 20$ dB.

that is, the signal to noise ratio SNR_1 of user 1 was equal to SNR_2 of user 2. In order to construct an SVM-based MUD for user 2, 160 training data points were generated for each given noise variance. The number of SVs was found typically to be around 40.

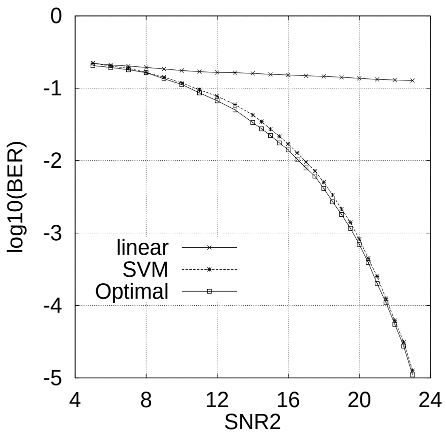

Fig. 3. Performance comparison of three MUDs, linear MMSE, adaptive SVM and optimal detectors, for user 2 of Example 1. WE had $\mathrm{SNR}_1 = \mathrm{SNR}_2$, and the training data set of the SVM had 160 samples.

Fig. 2 depicts the two subsets of noise-free signal states for user 2 together with the decision boundaries of the linear MMSE, as well as those of both the optimal and the SVM detectors, given $\mathrm{SNR}_1 = \mathrm{SNR}_2 = 20$ dB. It is clear that, for user 2, $\mathcal{R}_-$ and $\mathcal{R}_+$ are not linearly separable and the linear MMSE detector will have an irreducible error floor of 0.125,

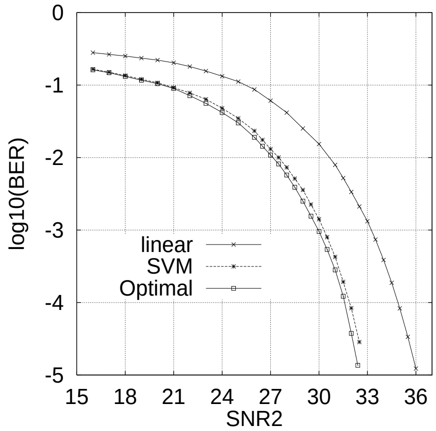

Fig. 4. Performance comparison of three MUDs, linear MMSE, adaptive SVM and optimal detectors, for user 2 of Example 2. SNR_i, $1 \le i \le 3$, were identical, and the training data set for SVM had 640 samples.

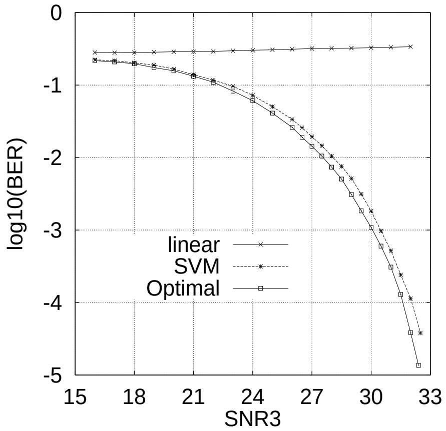

Fig. 5. Performance comparison of three MUDs, linear MMSE, adaptive SVM and optimal detectors, for user 3 of Example 2. SNR_i, $1 \le i \le 3$, were identical, and the training data set for SVM had 640 samples.

as can be seen in Figure 3, where the BERs of the optimal and the SVM detectors are also shown.

Example 2. A 3-user system employing 8 chips per bit was then also constructed. The code sequences for the three users were $(+1,+1,+1,+1,-1,-1,-1,-1)$, $(+1,-1,+1,-1,-1,+1,-1,+1)$ and $(+1,-1,-1,+1,-1,+1,+1,-1)$, respectively, and the transfer function of the CIR was $H(z) = 0.4 + 0.9z^{-1} + 0.4z^{-2}$. The three users had equal signal power. The number of training data points used for constructing SVM models was 640 for each given noise condition. For user 2 and 3, typically 180 SVs were selected from the training data set. The BERs of the resulting SVM-based MUDs for users 2 and 3 are given in Figs. 4 and 5, respectively, in comparison to the corresponding linear MMSE and optimal MUDs. The results again demonstrate that the SVM MUD can closely approximate the performance of the optimal detector.

VI. Conclusions

The SVM technique has been applied to adaptive nonlinear multiuser detection for DS-CDMA systems. It has been shown that the SVM-based MUD can closely match the performance of the optimal Bayesian one-shot detector, while having the important advantage of requiring a relatively small training data set, when compared to other neural network based multiuser detectors. A further advantage of the SVM approach is that the structure of the detector is automatically determined during training. A disadvantage of the SVM method is its block-based adaptation nature. Future research is required to investigate how to reduce the number of support vectors further without sacrificing the BER performance too much and how to incorporate the sample-by-sample adaptive methodology into the SVM approach.

References

[1] Z. Xie, R.T. Short and C.K. Rushforth, "A family of suboptimum detectors for coherent multiuser communications," *IEEE J. Selected Areas in Communications*, Vol.8, No.4, pp.683–690, 1990.

[2] U. Madhow and M.L. Honig, "MMSE interference suppression for direct-sequence spread-spectrum CDMA," *IEEE Trans. Communications*, Vol.42, No.12, pp.3178–3188, 1994.

[3] S.L. Miller, "An adaptive direct-sequence code-division multiple-access receiver for multiuser interference rejection," *IEEE Trans. Communications*, Vol.43, No.2/3/4, pp. 1746–1755, 1995.

[4] H.V. Poor and S. Verdú, "Probability of error in MMSE multiuser detection," *IEEE Trans. Information Theory*, Vol.43, No.3, pp.858–871, 1997.

[5] G. Woodward and B.S. Vucetic, "Adaptive detection for DS-CDMA," *Proc. IEEE*, Vol.86, No.7, pp.1413–1434, 1998.

[6] B. Aazhang, B.P. Paris and G.C. Orsak, "Neural networks for multiuser detection in code-division multiple-access communications," *IEEE Trans. Communications*, Vol.40, No.7, pp.1212–1222, 1992.

[7] U. Mitra and H.V. Poor, "Neural network techniques for adaptive multiuser demodulation," *IEEE J. Selected Areas in Communications*, Vol.12, No.9, pp.1460-1470, 1994.

[8] D.G.M. Cruickshank, "Radial basis function receivers for DS-CDMA," *Electronics Letters*, Vol.32, No.3, pp.188-190, 1996.

[9] R. Tanner and D.G.M. Cruickshank, "Volterra based receivers for DS-CDMA," in *Proc. 8th IEEE Int. Symp. Personal, Indoor and Mobile Radio Communications*, September 1997, Vol.3, pp.1166-1170.

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

[11] B. Schölkopf, K.K. Sung, C.J.C. Burges, F. Girosi, P. Niyogi, T. Poggio and V. Vapnik, "Comparing support vector machines with Gaussian kernels to radial basis function classifiers," *IEEE Trans. Signal Processing*, Vol.45, No.11, pp.2758–2765, 1997.

[12] C.J.C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, Vol.2, No.2, pp.121–167, 1998.

[13] S. Verdú, *Multiuser Detection*. Cambridge, UK: Cambridge University Press, 1998.

[14] S. Chen, B. Mulgrew and P.M. Grant, "A clustering technique for digital communications channel equalisation using radial basis function networks," *IEEE Trans. Neural Networks*, Vol.4, No.4, pp.570–579, 1993.

Appendix: Introduction to SVMs

The problem concerned is the separation of the set $\mathcal{X} = \{(c_l, \mathbf{x}_l)\}_{l=1}^{K}$ of K training data belonging to two classes, where $\mathbf{x}_l$ is an N-dimensional vector and $c_l \in \{\pm 1\}$ is its class label. Let us first consider the case, when $\mathcal{X}$ is linearly separable by a hyperplane

$$\mathbf{w}^T\mathbf{x} + \eta = 0, \tag{22}$$

In order to obtain a unique solution for the hyperplane parameters, it is appropriate to consider a canonical hyperplane [10], where $\mathbf{w} and \eta$ are constrained by:

$$\min_{\mathbf{x}_i \in \mathcal{X}} |\mathbf{w}^T\mathbf{x}_i + \eta| = 1. \tag{23}$$

A canonical separating hyperplane must satisfy:

$$c_i\left(\mathbf{w}^T\mathbf{x}_i + \eta\right) \geq 1,\ \forall \mathbf{x}_i \in \mathcal{X}. \tag{24}$$

Observe in Fig. 6 that there is an innumerable number of hyperplanes, which can correctly separate $\mathcal{X}$ into two classes. The best hyperplane is the one exibiting the property that the distance between the closest training vector to the hyperplane is maximal, that is, the optimal hyperplane can be found by maximizing the margin:

$$\rho(\mathbf{w}, \eta) = \min_{\{\mathbf{x}_i | c_i = +1\}} \frac{|\mathbf{w}^T\mathbf{x}_i + \eta|}{\|\mathbf{w}\|} + \min_{\{\mathbf{x}_j | c_j = -1\}} \frac{|\mathbf{w}^T\mathbf{x}_j + \eta|}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|}, \tag{25}$$

subject to the constraints (24). Thus, the optimal hyperplane is the one that minimizes

$$\Phi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2, \tag{26}$$

subject to (24). The solution to this constrained optimization problem is given by the saddle point of the Lagrangian:

$$L(\mathbf{w}, \eta, \mathbf{g}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{K} g_i\left(\left(\mathbf{w}^T\mathbf{x}_i + \eta\right)c_i - 1\right), \tag{27}$$

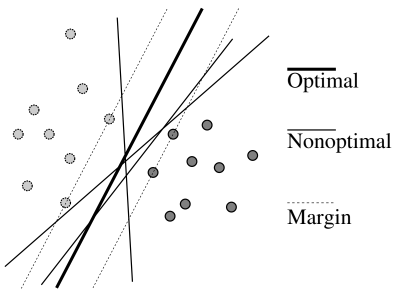

Fig. 6. Optimal and non-optimal separating hyperplanes.

where g_i are the Lagrange multipliers. The Lagrangian has to be minimized with respect to (w.r.t.) $\mathbf{w}, \eta$ and maximized w.r.t. $g_i \geq 0$. Classical Lagrangian duality enables the primary problem (27) to be transformed to its dual problem:

$$\max_{\mathbf{g}} \Psi(\mathbf{g}) = \max_{\mathbf{g}} \left\{ \min_{\mathbf{w}, \eta} L(\mathbf{w}, \eta, \mathbf{g}) \right\} = \max_{\mathbf{g}} \left\{ -\frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K} g_i g_j\, c_i c_j\, \mathbf{x}_i^T\mathbf{x}_j + \sum_{i=1}^{K} g_i \right\}, \tag{28}$$

and the solution of the dual problem is given by

$$\bar{\mathbf{g}} = \arg\min_{\mathbf{g}} \left\{ \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K} g_i g_j\, c_i c_j\, \mathbf{x}_i^T\mathbf{x}_j - \sum_{i=1}^{K} g_i \right\}, \tag{29}$$

with the constraints

$$g_i \geq 0,\ i = 1, \cdots, K, \ \text{ and } \sum_{i=1}^{K} g_i\, c_i = 0. \tag{30}$$

Solving the quadratic optimization problem (29) subject to the constraints (30) determines the Lagrange multipliers, and the optimal separating hyperplane is given by

$$\bar{\mathbf{w}} = \sum_{i=1}^{K} \bar{g}_i\, c_i\, \mathbf{x}_i \ \text{ and } \ \bar{\eta} = -\frac{1}{2}\bar{\mathbf{w}}^T\left(\mathbf{x}_+ + \mathbf{x}_-\right), \tag{31}$$

where $\mathbf{x}_+$ and $\mathbf{x}_-$ are any two support vectors, one from each class. Notice that $\bar{\mathbf{w}}, \bar{\eta}$ are defined by the set of support vectors, which are training points lying on the margin, since only these have non-zero Lagrange multipliers. The support vectors form a very small subset of $\mathcal{X}$.

Let us now consider the nonlinearly separable case. The basic idea is to nonlinearly map the data space onto a new feature space, on which the problem becomes linearly separable. It turns out that the solution is given in the form of (14), and the corresponding Lagrange multipliers are determined by substituting $\mathbf{x}_i^T\mathbf{x}_j$ in (29) with the kernel function $F(\mathbf{x}_i, \mathbf{x}_j)$.

In reality, a zero classification error may not be possible. In such situations, the Lagrange multipliers have an upper bound C, as given in (17). Notice that the support vectors - namely those, which have non-zero Lagrange multipliers - are not necessarily lying on the margin now. The parameter C can be chosen to provide an appropriate trade-off between the model's complexity - which is quantified in terms of the number of support vectors used - and training error.