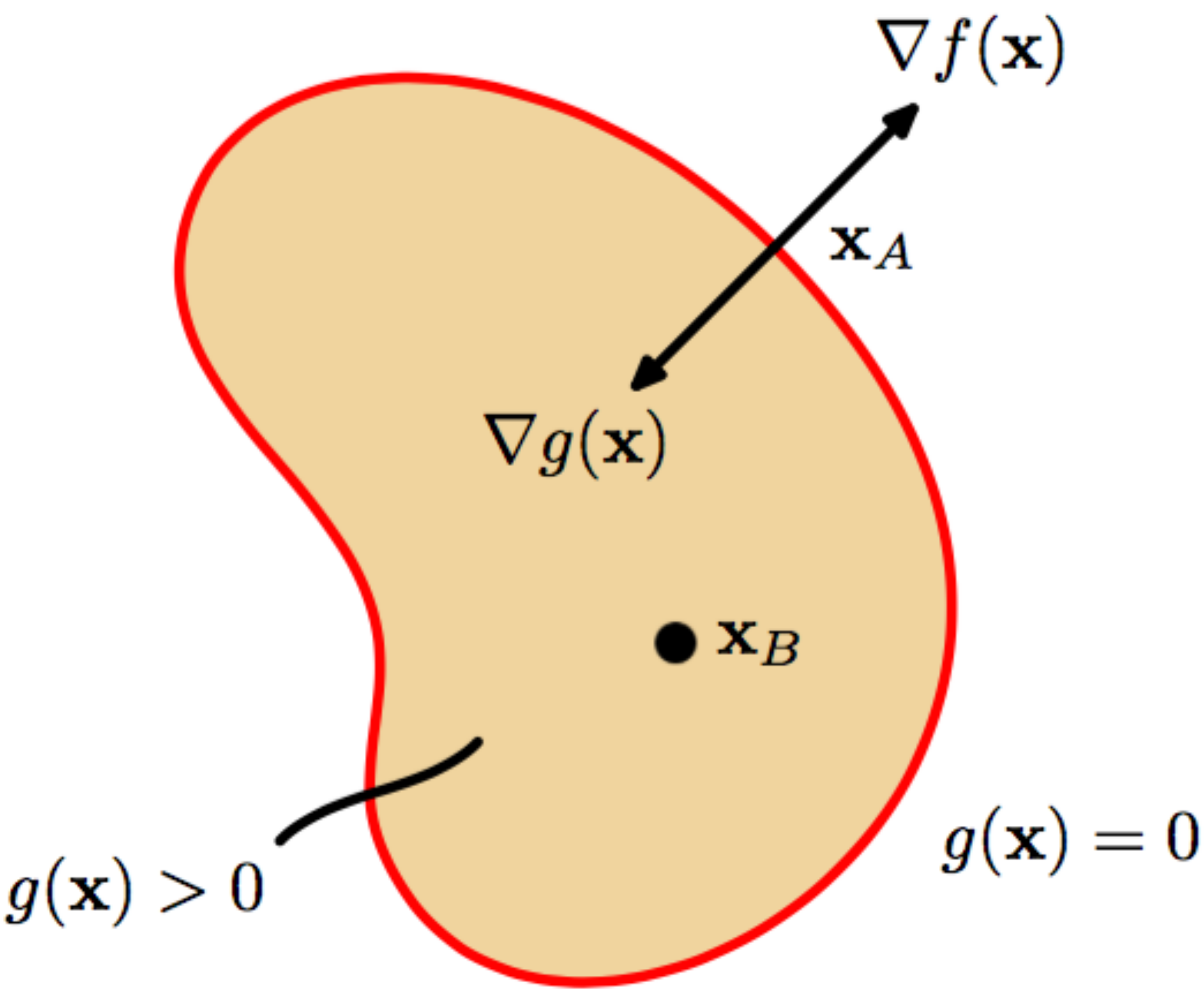
Support Vectors Continued

Kernels,Slack variables and SMO

with help from A.Zizzerman, A. Ng & S. Haykin

Inequality constraints



$\max f(x)$ stc $g(x) \geq 0$

KKT: $g(x) \geq 0$

$\lambda \geq 0$

$\lambda g(x) = 0$

KERNEL FUNCTIONS

Now the big bonus occurs because all the machinery we have developed will work if we map the points $\mathbf{x}_i$ to a higher dimensional space, provided we observe certain conventions.

Let $\phi(\mathbf{x}_i)$ be a function that does the mapping. So the new hyperplane is

$$\sum_{i=1}^{N} w_i \phi_i(\mathbf{x}) + b = 0$$

For simplicity in notation define

$$\phi(\mathbf{x}) = (\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_{m_1}(\mathbf{x}))$$

where m_1 is the new dimension size and by convention $\phi_0(\mathbf{x}) = 1$.

Then all the work we did with $\mathbf{x}$ works with $\phi(\mathbf{x})$. The only issue is that instead of $\mathbf{x}_i^T \mathbf{x}_j$ we have a *Kernel function*, $K(\mathbf{x}_i, \mathbf{x}_j)$ where

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi_i(\mathbf{x})^T \phi_j(\mathbf{x})$$

and Kernel functions need to have certain nice properties. :)

Examples

Polynomials

$$(\mathbf{x}_i^T \mathbf{x}_j + 1)^p$$

Radial Basis Functions

$$\exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_i||^2}{2\sigma^2}\right)$$

Lets see an example. Suppose $x, z \in \mathbb{R}^n$, and consider

$$K(x,z) = (x^T z)^2.$$

We can also write this as

$$\begin{aligned} K(x,z) &= \left(\sum_{i=1}^n x_i z_i\right)\left(\sum_{j=1}^n x_i z_i\right) \\ &= \sum_{i=1}^n \sum_{j=1}^n x_i x_j z_i z_j \\ &= \sum_{i,j=1}^n (x_i x_j)(z_i z_j) \end{aligned}$$

Thus, we see that $K(x,z) = \phi(x)^T\phi(z)$, where the feature mapping ϕ is given (shown here for the case of $n = 3$) by

$$\phi(x) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \end{bmatrix}.$$

Note that whereas calculating the high-dimensional $\phi(x)$ requires $O(n^2)$ time, finding $K(x,z)$ takes only $O(n)$ time—linear in the dimension of the input attributes.

This is known as the 'Kernel Trick'
When is a Kernel a Kernel?
K being **symmetric** and **positive definite** is necessary and sufficient

A reprise of last time …

The problem statement

Given a set of training data $\{(\mathbf{x}_i, d_i), i = 1, \ldots, N\}$, minimize

$$\Phi(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

subject to the constraint that

$$d_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1, i = 1, \ldots, N$$

The problem statement

Given a set of training data $\{(\mathbf{x}_i, d_i), i = 1, \dots, N\}$, minimize

$$\Phi(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

subject to the constraint that

$$d_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1, i = 1, \dots, N$$

Looks like a job for LAGRANGE MULTIPLIERS!

$$J(\mathbf{w}, b, \lambda) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \lambda_i(d_i(\mathbf{w}^T\mathbf{x_i} + b) - 1)$$

So that

$$J_{\mathbf{w}} = \mathbf{0} = \mathbf{w} - \sum_{i=1}^{N} \lambda_i d_i \mathbf{x_i} \tag{3}$$

and

$$J_b = 0 = \sum_{i=1}^{N} \lambda_i d_i \tag{4}$$

Now for the DUAL PROBLEM

$$J(\mathbf{w}, b, \lambda) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \lambda_i d_i \mathbf{w}^T\mathbf{x_i} + b\sum_{i=1}^{N} \lambda_i d_i + \sum_{i=1}^{N} \lambda_i$$

Note that from (4) third term is zero. Using Eq. (3):

$$Q(\lambda) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i\lambda_j d_i d_j \mathbf{x}_i^T\mathbf{x}_j$$

DUAL PROBLEM

$$\max Q(\lambda) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j$$

Subject to constraints

$$\sum_{i=1}^{N} \lambda_i d_i = 0$$

$$\lambda_i \geq 0, \;\; i = 1, \ldots, N$$

This is *easier to solve* than the original. Furthermore it only depends on the training samples $\{(\mathbf{x}_i, d_i), i = 1, \ldots, N\}$.

Once you have the λ_is, get the $\mathbf{w}$ from

$$\mathbf{w} = \sum_{i=1}^{N} \lambda_i d_i \mathbf{x}_\mathrm{i}$$

and the b from a support vector that has $d_i = 1$,

$$b = 1 - \mathbf{w}^T \mathbf{x}_s$$

Given the training sample $\{(\mathbf{x}_i, d_i)\}_{i=1}^{N}$, *find the optimum values of the weight vector* $\mathbf{w}$ *and bias* b *such that they satisfy the constraint*

$$d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \qquad \text{for } i = 1, 2, ..., N \tag{6.24}$$

$$\xi_i \geq 0 \qquad \text{for all } i \tag{6.25}$$

and such that the weight vector $\mathbf{w}$ *and the slack variables* ξ_i *minimize the cost functional*

$$\Phi(\mathbf{w}, \xi) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i \tag{6.26}$$

where C *is a user-specified positive parameter.*

Given the training sample $\{(\mathbf{x}_i, d_i)\}_{i=1}^{N}$, *find the Lagrange multipliers* $\{\alpha_i\}_{i=1}^{N}$ *that maximize the objective function*

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j \tag{6.27}$$

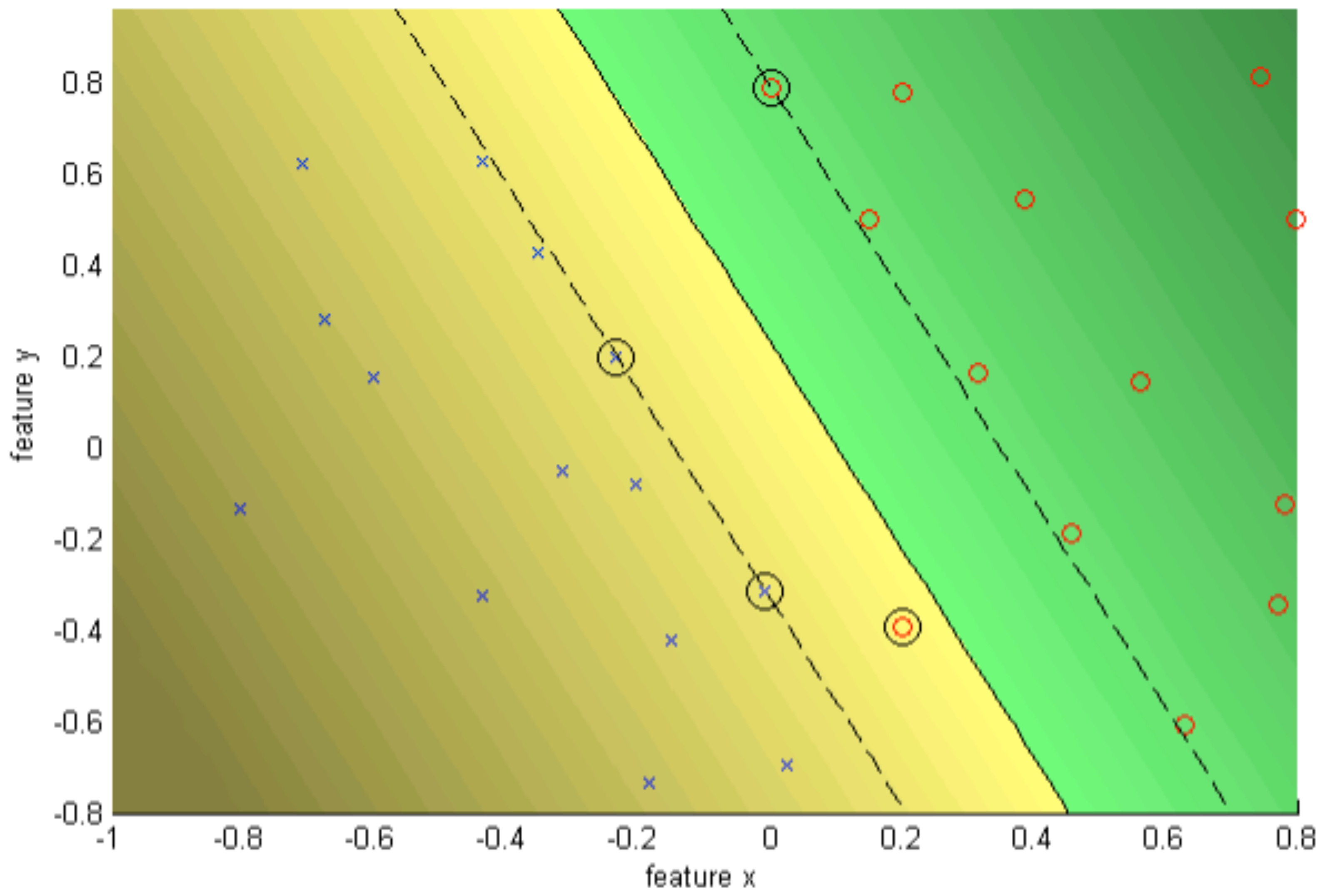
subject to the constraints

$$(1) \quad \sum_{i=1}^{N} \alpha_i d_i = 0$$

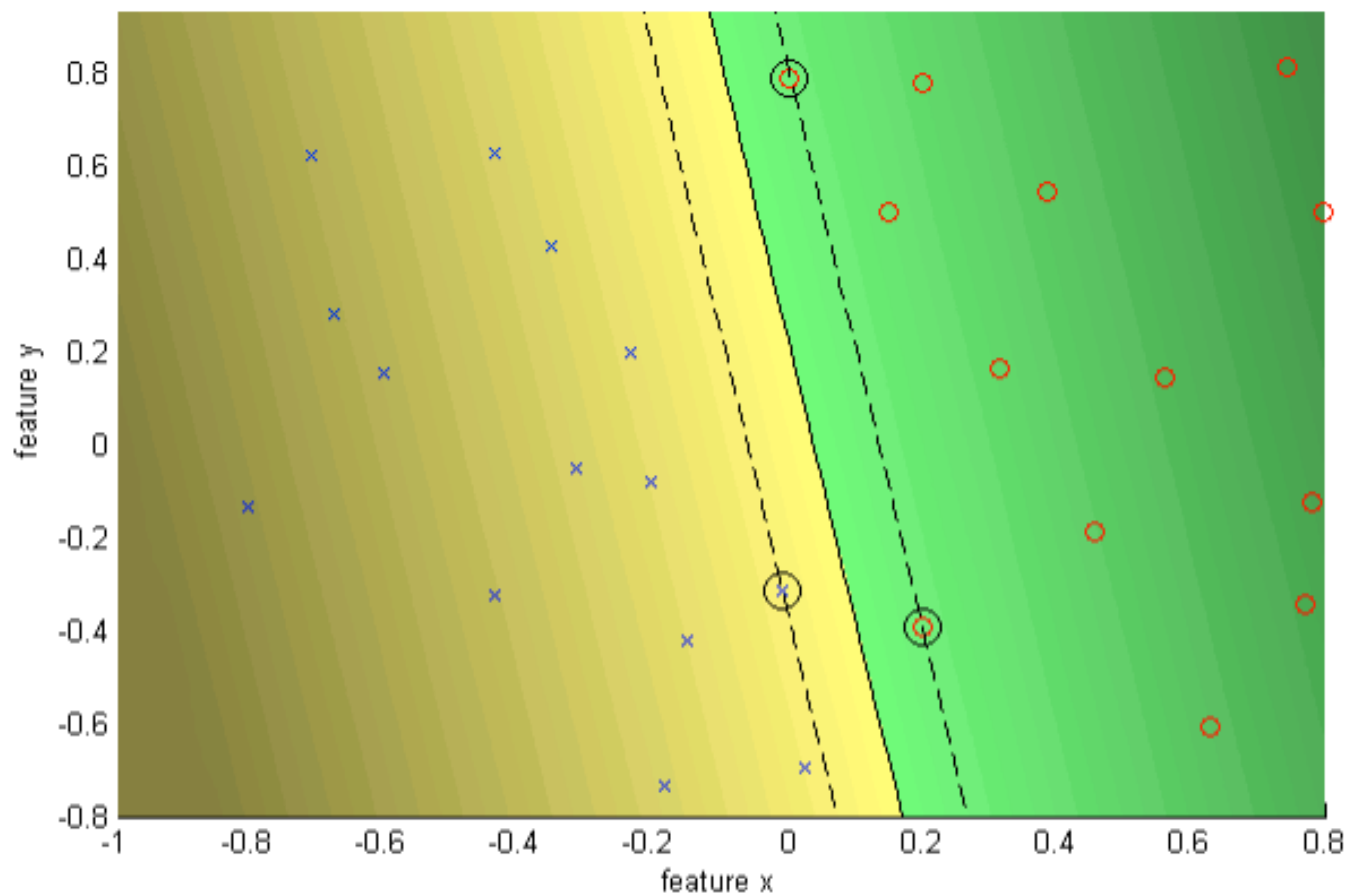
$$(2) \quad 0 \le \alpha_i \le C \qquad \text{for } i = 1, 2, \ldots, N$$

where C is a user-specified positive parameter.

C = 10 soft margin



C = Infinity   hard margin



SMO: Sequential Minimization Optimization

Let's go back to the dual formulation.
Note the soft constraint formulation

$$\max_\alpha \quad W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle. \tag{17}$$

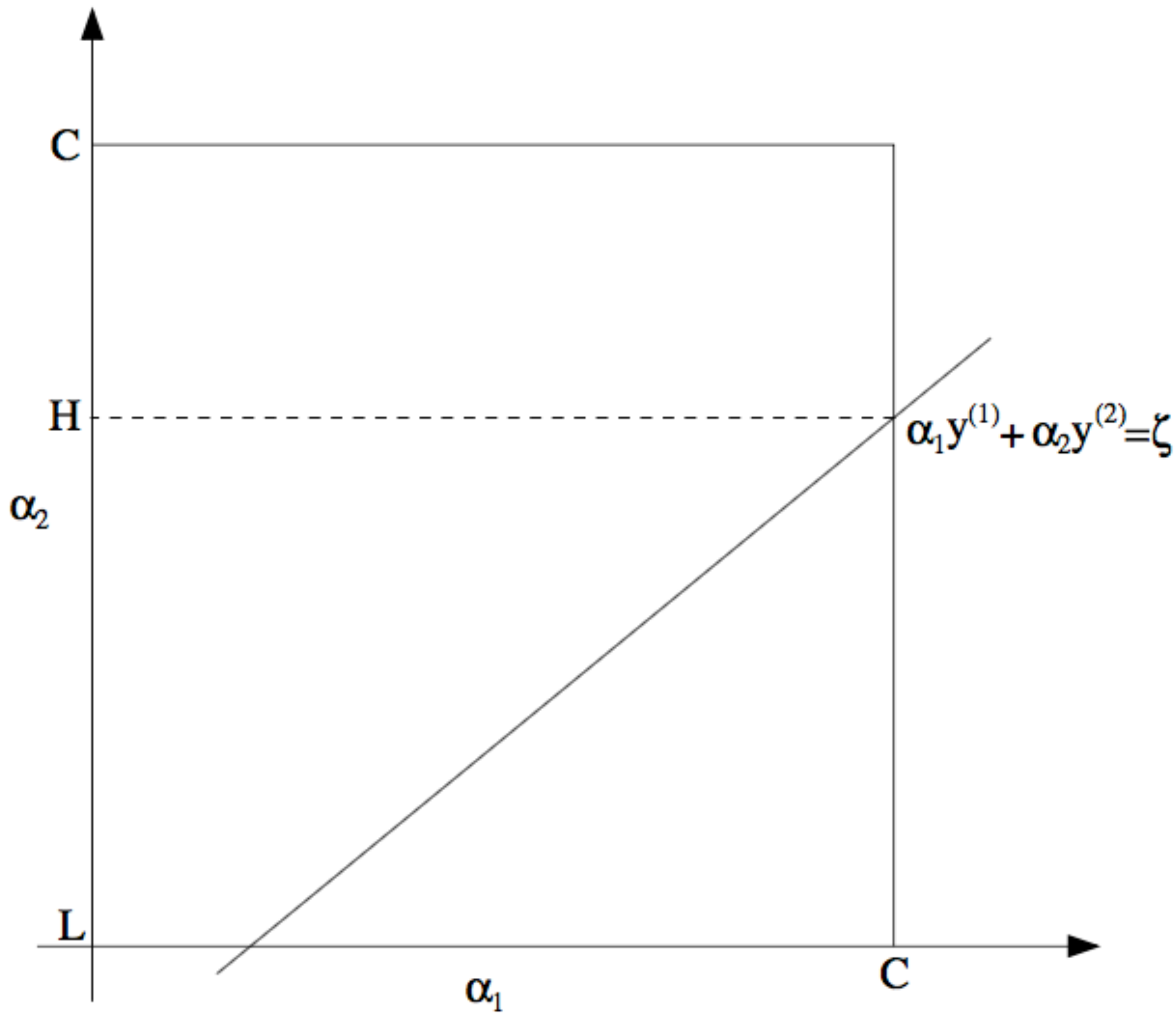
$$\text{s.t.} \quad 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, m \tag{18}$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0. \tag{19}$$

$$\alpha_1 y^{(1)} = -\sum_{i=2}^{m} \alpha_i y^{(i)}$$

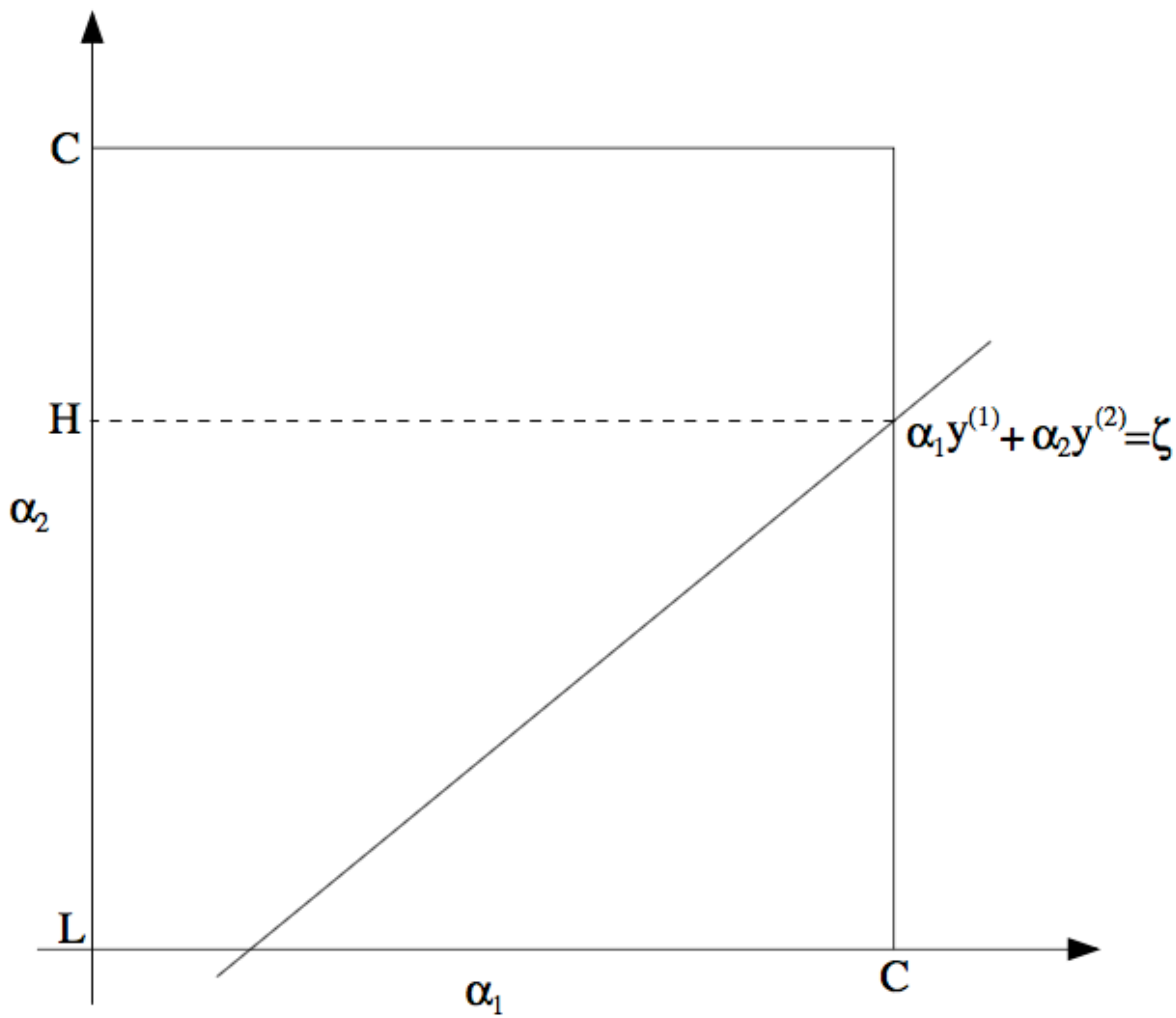
$$\alpha_1 = -y^{(1)} \sum_{i=2}^{m} \alpha_i y^{(i)}$$

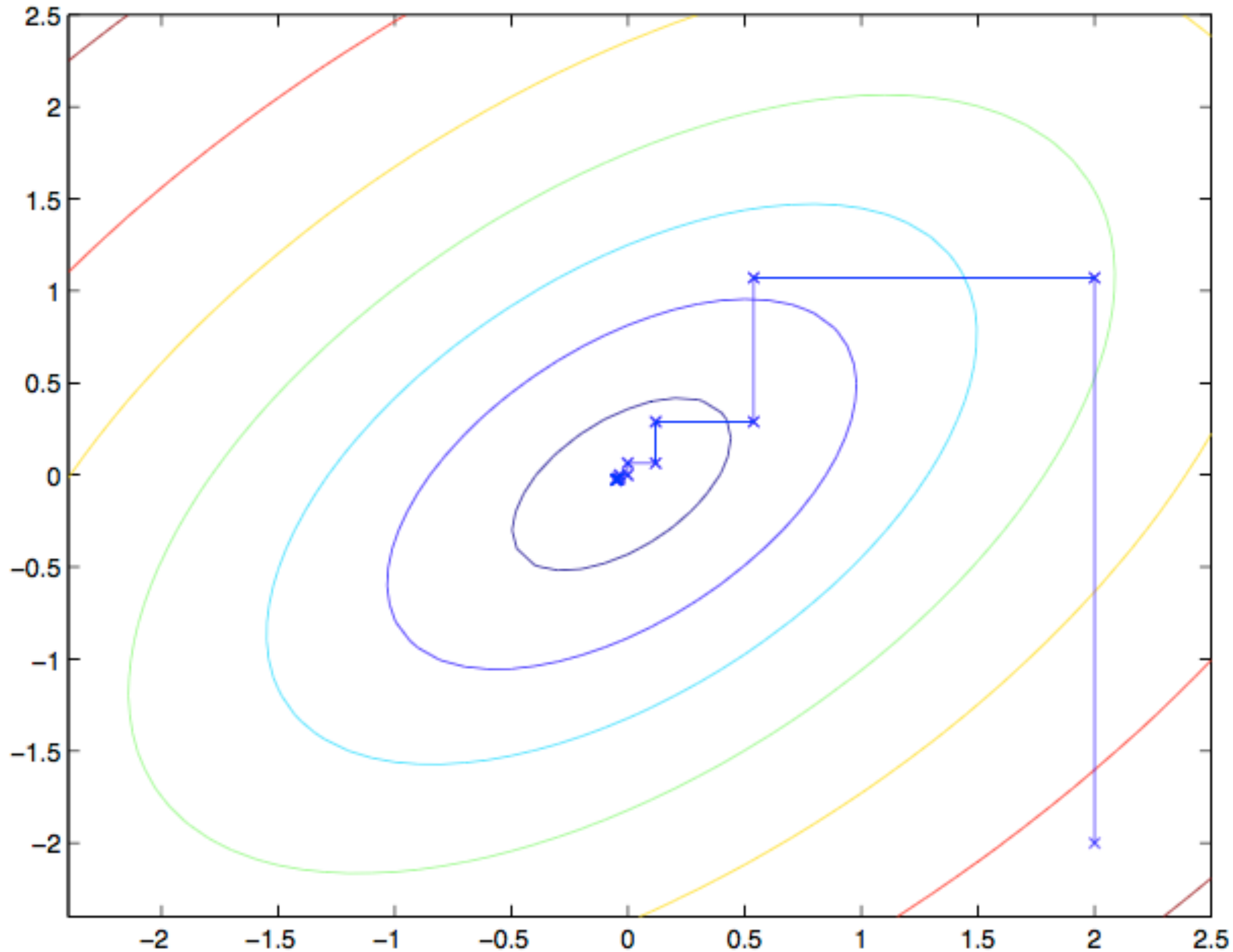
$$\alpha_1 y^{(1)} + \alpha_2 y^{(2)} = -\sum_{i=3}^{m} \alpha_i y^{(i)}$$


C
H
α_2
L
α_1
C
$\alpha_1 y^{(1)} + \alpha_2 y^{(2)} = \zeta$

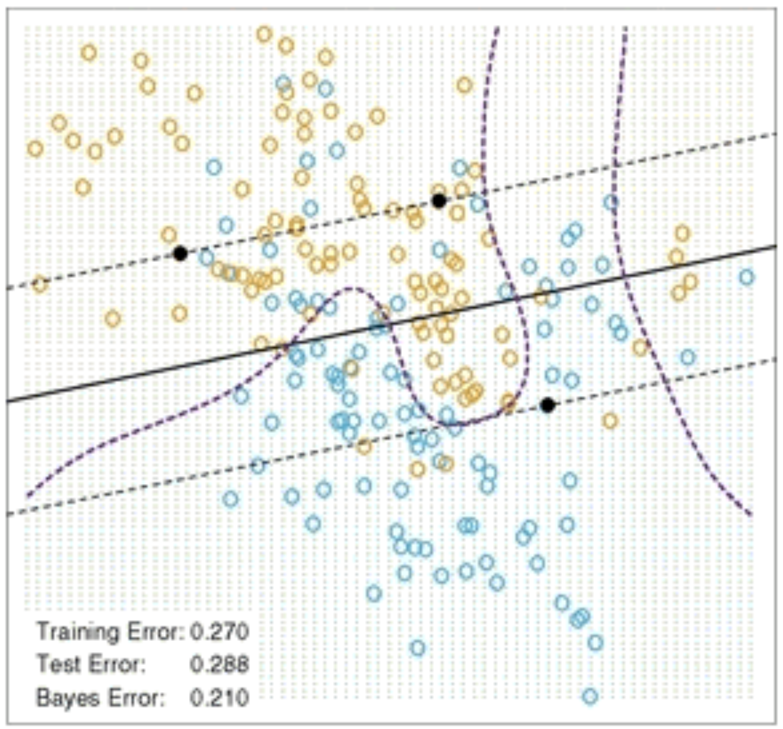
$$\alpha_1 = (\zeta - \alpha_2 y^{(2)}) y^{(1)}$$

$$\alpha_2^{new} = \begin{cases} H & \text{if } \alpha_2^{new,unclipped} > H \\ \alpha_2^{new,unclipped} & \text{if } L \leq \alpha_2^{new,unclipped} \leq H \\ L & \text{if } \alpha_2^{new,unclipped} < L \end{cases}$$

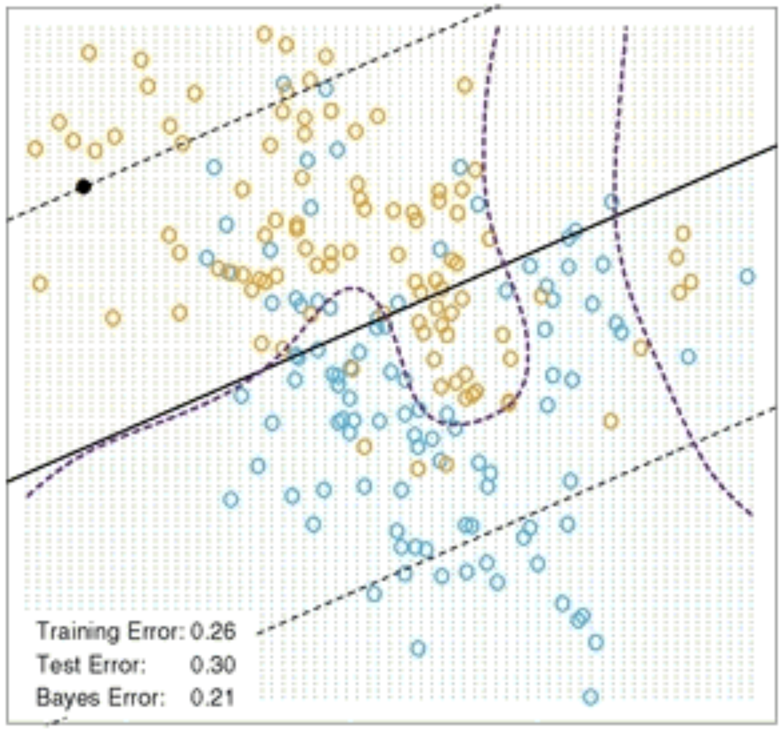




2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
2.5

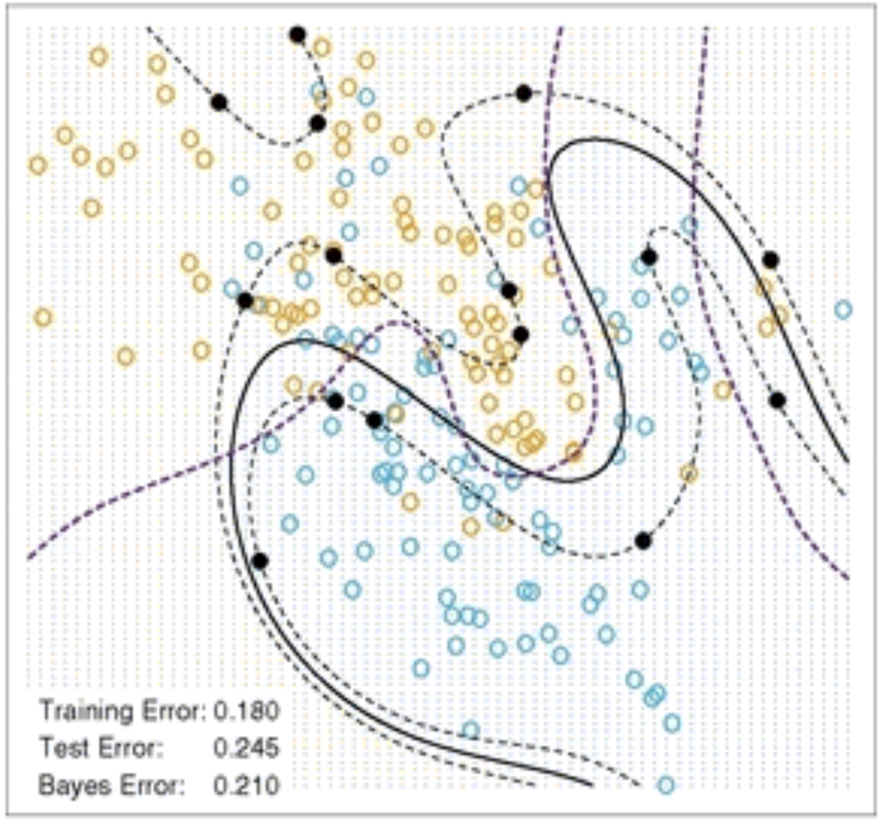

Training Error: 0.270
Test Error: 0.288
Bayes Error: 0.210

$C = 10000$

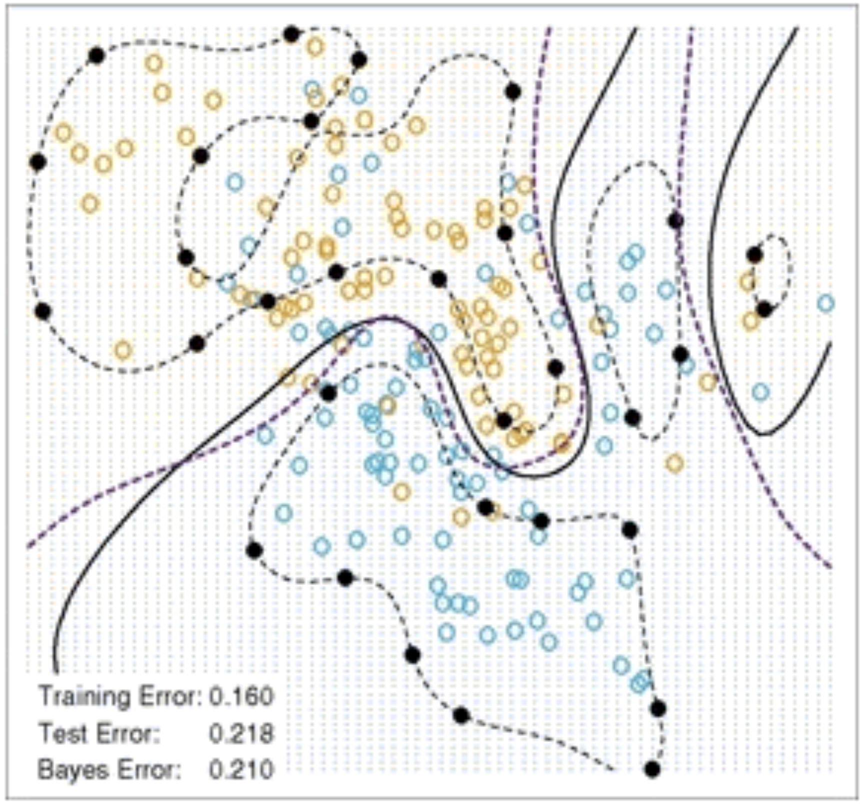

Training Error: 0.26
Test Error: 0.30
Bayes Error: 0.21

$C = 0.01$

SVM - Degree-4 Polynomial in Feature Space



SVM - Radial Kernel in Feature Space



Soft constraints

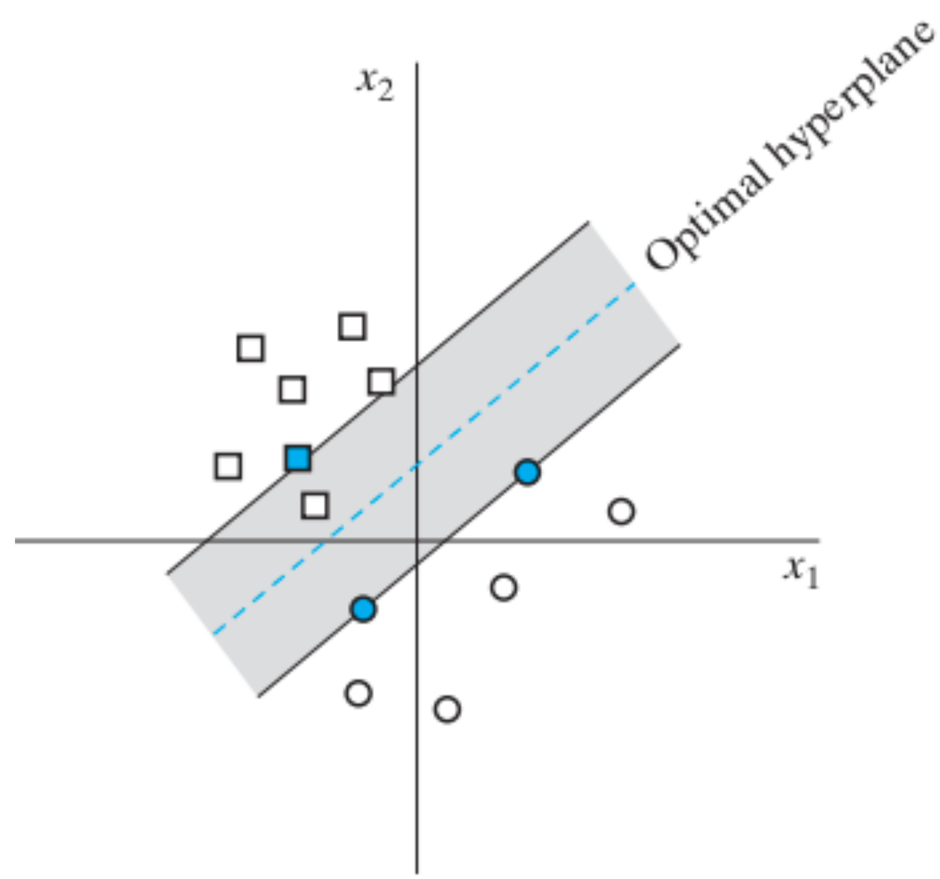
The margin of separation between classes is said to be *soft* if a data point $(\mathbf{x}_i, d_i)$ violates the following condition (see Eq. (6.10)):

$$d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq +1, \qquad i = 1, 2, ..., N$$

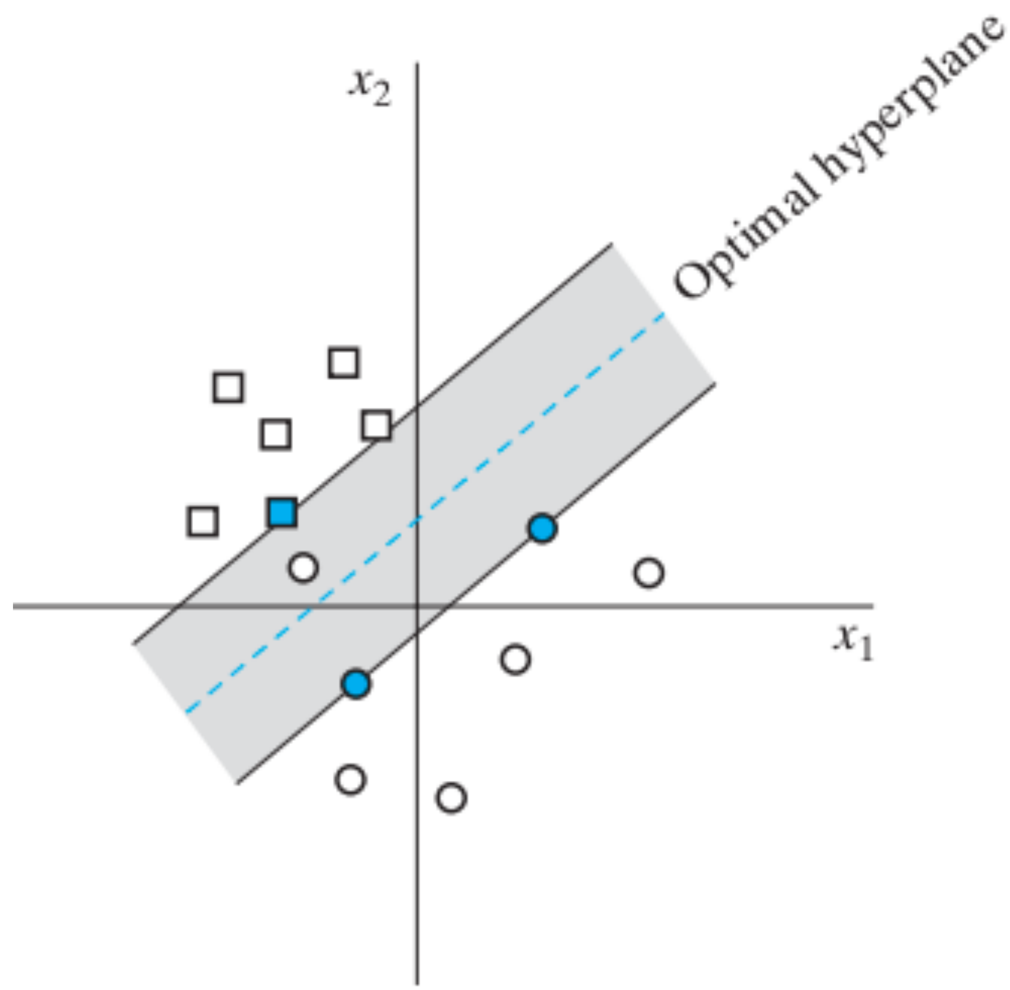
- The data point $(\mathbf{x}_i, d_i)$ falls on the wrong side of the decision surface, as illustrated in Fig. 6.3b.

This violation can arise in one of two ways:

- The data point $(\mathbf{x}_i, d_i)$ falls inside the region of separation, but on the correct side of the decision surface, as illustrated in Fig. 6.3a.



- The data point $(\mathbf{x}_i, d_i)$ falls on the wrong side of the decision surface, as illustrated in Fig. 6.3b.



To make the optimization problem mathematically tractable, we approximate the functional $\Phi(\xi)$ by writing

$$\Phi(\xi) = \sum_{i=1}^{N} \xi_i$$

Moreover, we simplify the computation by formulating the functional to be minimized with respect to the weight vector **w** as follows:

$$\Phi(\mathbf{w}, \xi) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} \xi_i \tag{6.23}$$