Neural Networks for Signal Processing-I

Prof. Shayan Srinivasa Garani

Department of Electronic System Engineering

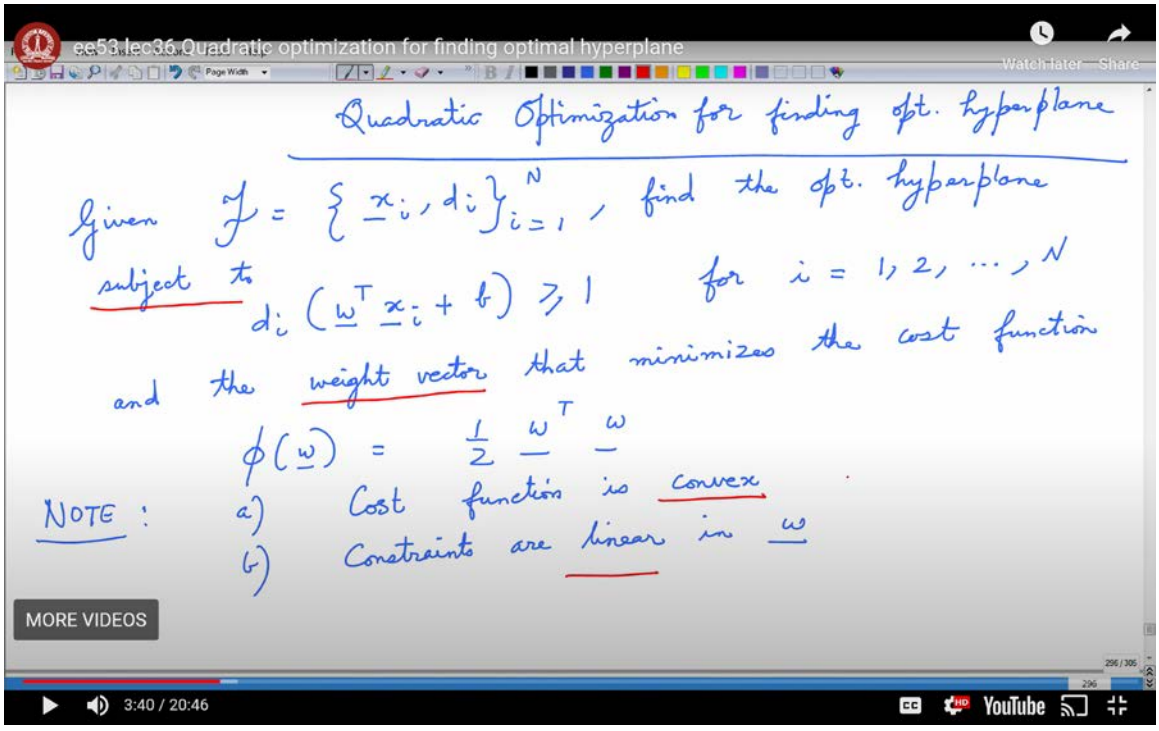
Indian Institute of Science, Bengaluru

Lecture – 36

Quadratic Optimization for Finding Optimal Hyperplane

In our previous discussions on hyperplanes within Support Vector Machines (SVMs), we highlighted the goal of maximizing the separation margin between classes. This margin is inversely related to the norm of the weight vector W. To reiterate, maximizing the margin of separation is equivalent to minimizing the norm of W.

(Refer Slide Time: 03:40)



Now, let's formulate the SVM problem for linearly separable patterns using an optimization framework. We start with a set F comprising pairs (x_i, d_i) for i = 1 to n, where

we have n examples. Our objective is to find an optimal hyperplane subject to the following constraints:

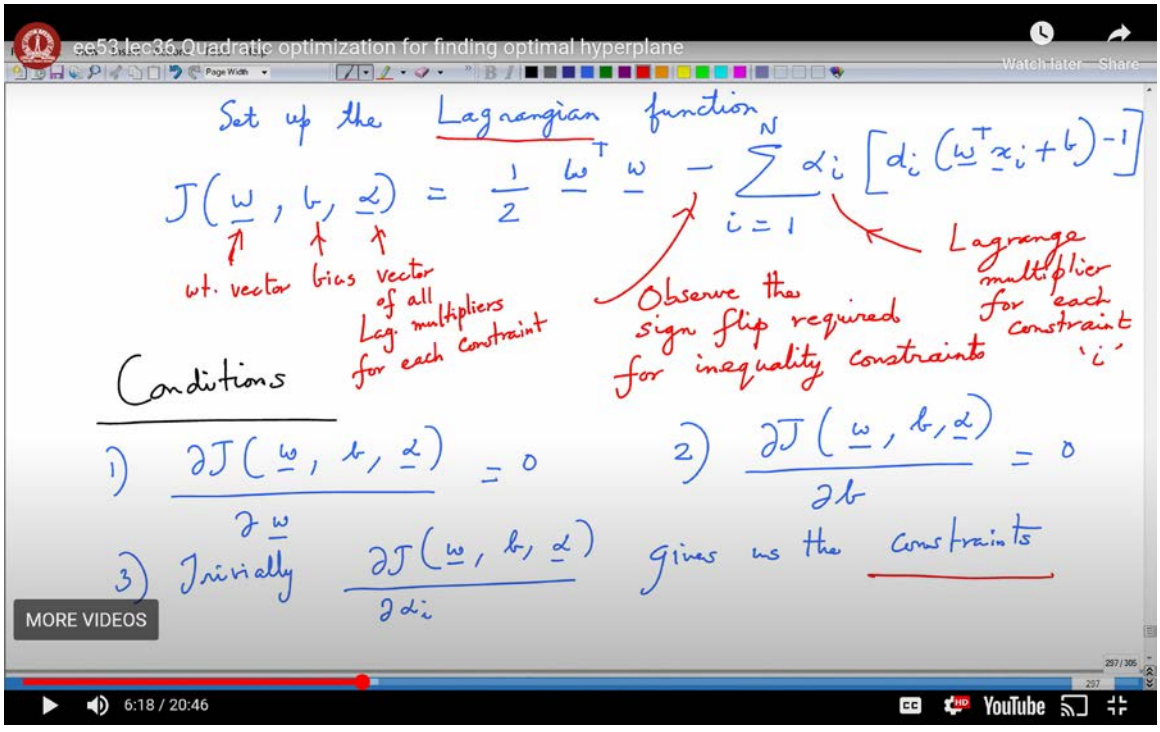
$$d_i(W^T x_i + b) \geq 1 \text{ for } i = 1 \text{ through } n.$$

Here's what these constraints imply:

- If $W^T xi + b \geq 1$, then d_i is 1.
- If $W^T xi + b < 0$, then d_i is -1.
- Conversely, if $W^T xi + b > 0$, then d_i is +1.

So, the term $d_i(W^T x_i + b)$ should be greater than or equal to 1. This means you have a positive term multiplied by +1 or -1 depending on the side of the hyperplane x_i lies.

(Refer Slide Time: 06:18)



Our task is to select the weight vector W that minimizes the cost function φ(W), defined as:

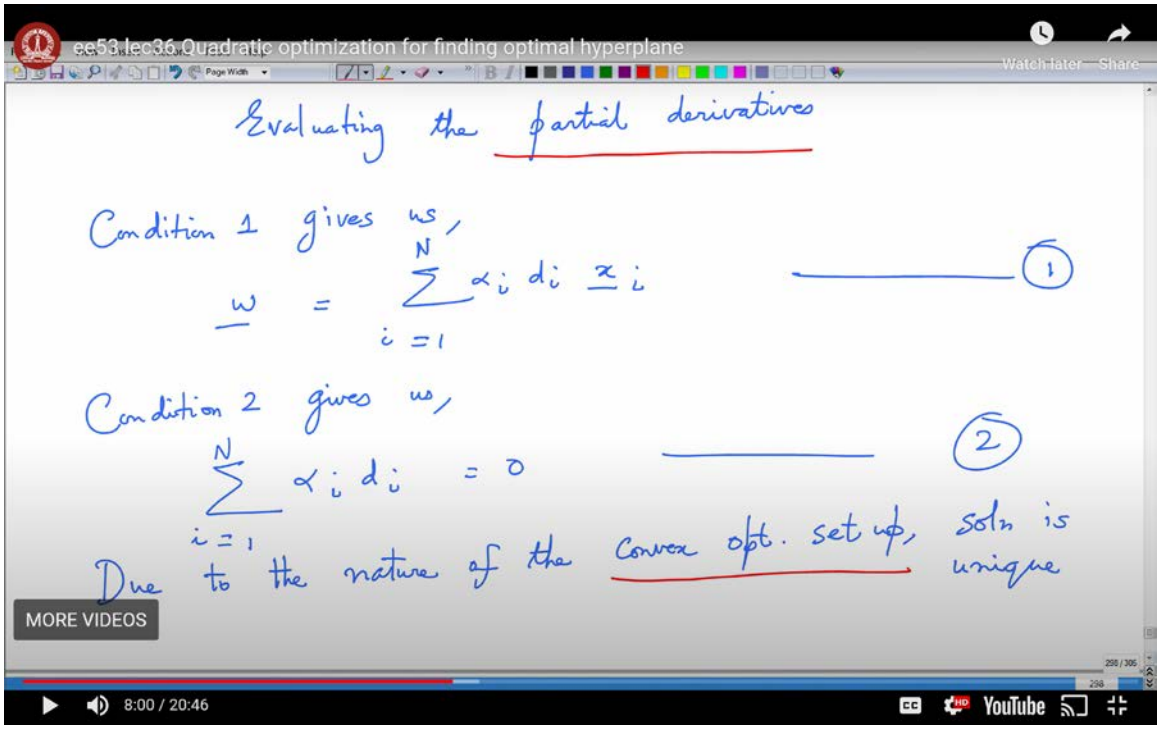
$$\phi(W) = \frac{1}{2} W^T W.$$

The factor of $\frac{1}{2}$ is included to simplify the differentiation, as the derivative of a quadratic function typically involves a factor of 2.

This cost function is quadratic and convex, and the constraints are linear with respect to W. The constraints are:

$$d_i(W^T x_i + b) \geq 1 \text{ for all } i = 1 \text{ to } n.$$

(Refer Slide Time: 08:00)



To set up the Lagrangian for this problem, we incorporate the Lagrange multipliers for each constraint. The overall cost function J or the Lagrangian is given by:

$$J(W, b, \alpha) = \frac{1}{2} W^T W - \sum_{i=1}^{n} \alpha_i [d_i(W^T x_i + b) - 1],$$

where α_i are the Lagrange multipliers associated with each constraint.

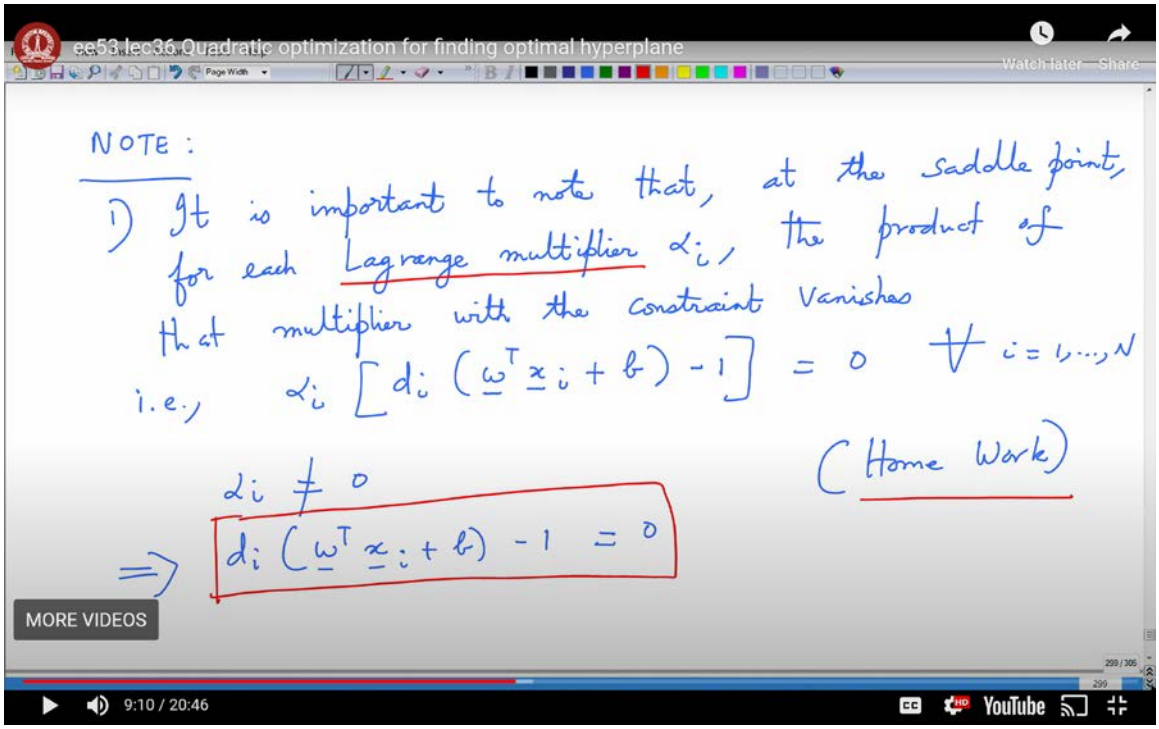
For optimality, we need to:

1. Take the derivative of the Lagrangian J with respect to W and set it equal to zero.

2. Take the partial derivative of J with respect to b and set it equal to zero.

3. Ensure that the gradient of J with respect to the Lagrange multipliers α_i yields the original constraints.

(Refer Slide Time: 09:10)



By solving these equations, we determine the optimal values for W, b, and α_i, which will give us the optimal hyperplane and the maximum margin of separation.

Let's walk through the key conditions and steps in solving the optimization problem for SVMs.

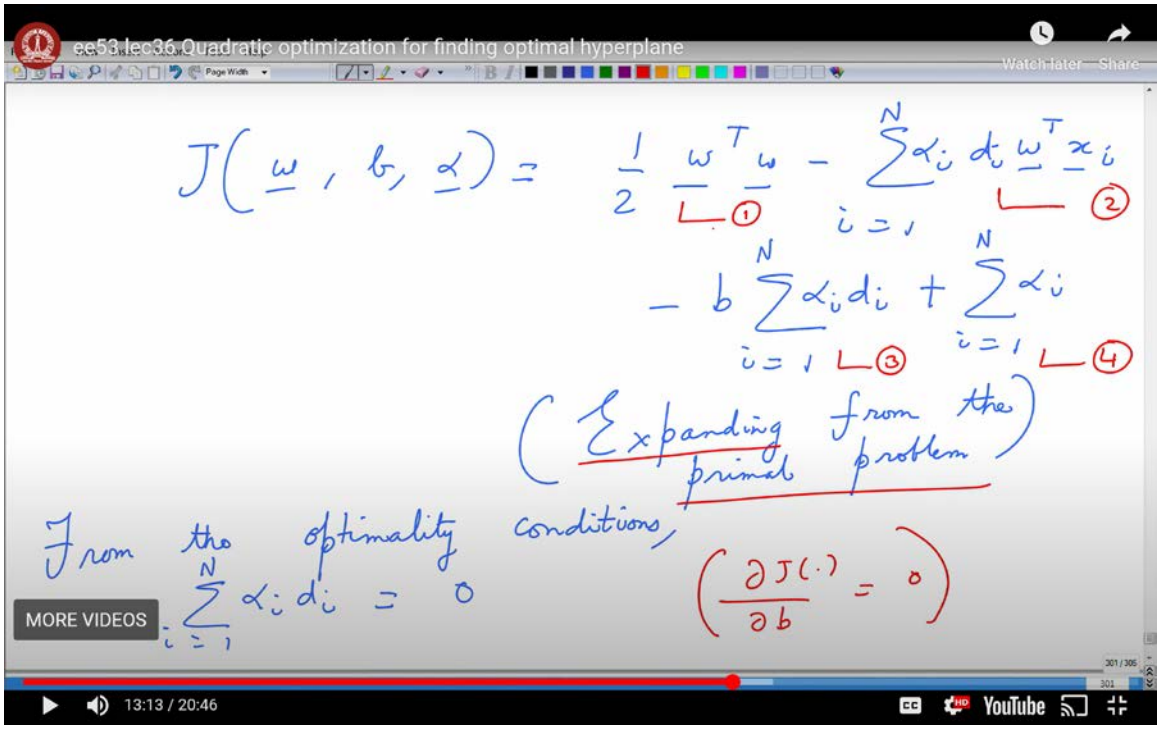
Firstly, conditions 1 and 2 are crucial for solving the problem, while condition 3 is straightforward since it ensures that the constraints are satisfied. Let's dive into evaluating the partial derivatives.

Condition 1 involves setting the partial derivative of the Lagrangian $\frac{\partial J}{\partial W}$ equal to zero. By performing matrix differentiation and following the rules of matrix calculus, we find that:

$$\frac{\partial J}{\partial W} = \sum_{i=1}^{N} \alpha_i d_i x_i = 0.$$

This gives us our first condition.

(Refer Slide Time: 13:13)



Condition 2 requires taking the partial derivative of the Lagrangian with respect to the bias b. Here, we focus on the coefficients associated with b, which are α_i and d_i. Setting the derivative to zero yields:

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

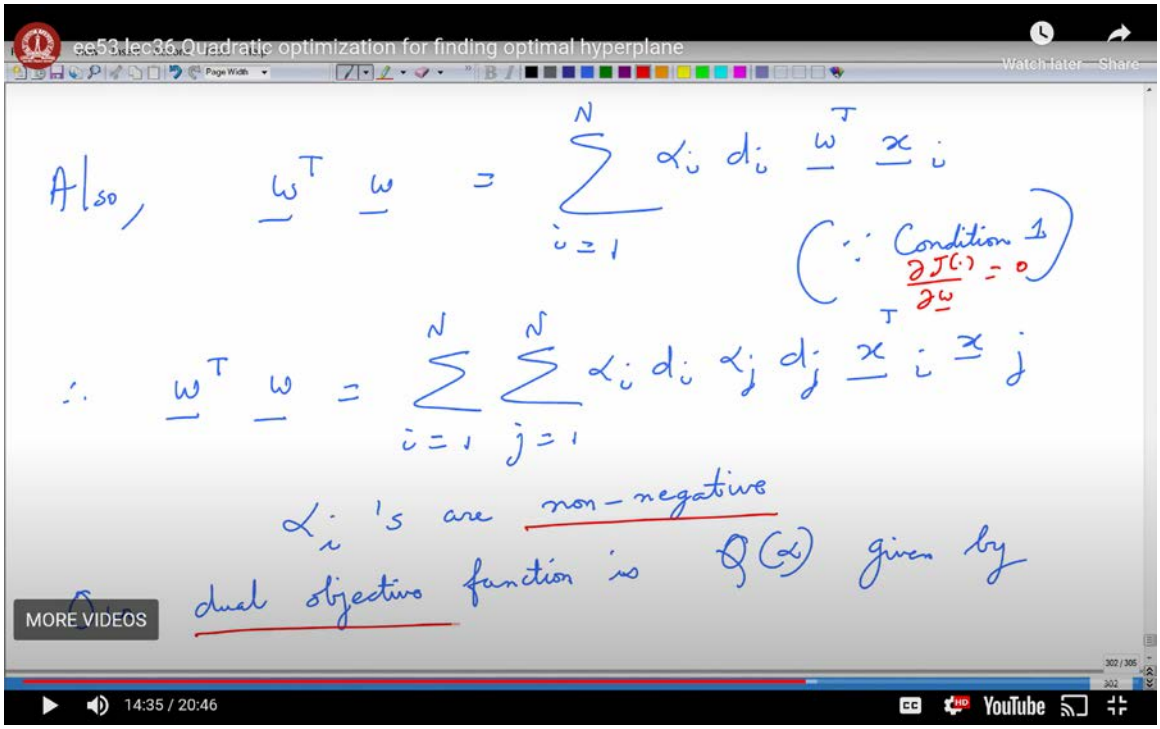
This is our second condition.

Since we are solving a convex optimization problem, the solution is unique due to the convex nature of the setup. Once we solve for these conditions subject to the constraints, we achieve the optimal solution.

Important Note: At the saddle point for each Lagrange multiplier α_i, the quantity $\alpha_i(d_i(W^T x_i + b) - 1)$ must be zero for all i = 1 to n, where $\alpha_i \neq 0$. This implies that for $\alpha_i \neq 0$, we have:

$$d_i(W^T x_i + b) = 1.$$

(Refer Slide Time: 14:35)



I encourage you to reflect on this condition, as it's a key aspect of understanding the SVM solution.

In general optimization problems, we often explore the relationship between primal and dual problems. For instance, if we are maximizing a quantity subject to constraints, we can formulate a dual problem by negating the sign in the objective function and minimizing that new objective subject to constraints. The optimality conditions derived from the primal

problem can be used to formulate the dual problem, which will also have its own optimal values.

In non-convex problems, there can be a gap between the primal and dual problems, which optimization techniques aim to minimize. However, for convex problems like SVMs, this gap is zero.

For SVMs, the primal problem focuses on finding the optimal weight vector W. The dual problem involves setting up and solving the derivative of the dual Lagrangian with respect to an alternative variable, and finding the optimal values for this variable.

I will not delve further into primal and dual problems here, as they are covered in detail in optimization courses. However, once we have formulated the problem in its primal form, we will invoke the dual form to continue our discussion on SVMs.

Thus, we expand the Lagrangian J(W, b, α) as follows:

$$J(W, b, \alpha) = \frac{1}{2} W^T W - \sum_{i=1}^{N} \alpha_i [d_i (W^T x_i + b) - 1],$$

where W and α are vectors. This formulation incorporates the summation of Lagrange multipliers scaled by their corresponding constraints.

Let's break down the optimization setup for our SVM problem.

We have reduced the problem to four key terms:

1. Objective Term: $\frac{1}{2} W^T W$, which is our primary objective function.

2. Constraint Terms: Terms 2, 3, and 4, which originate from the constraints of the problem.

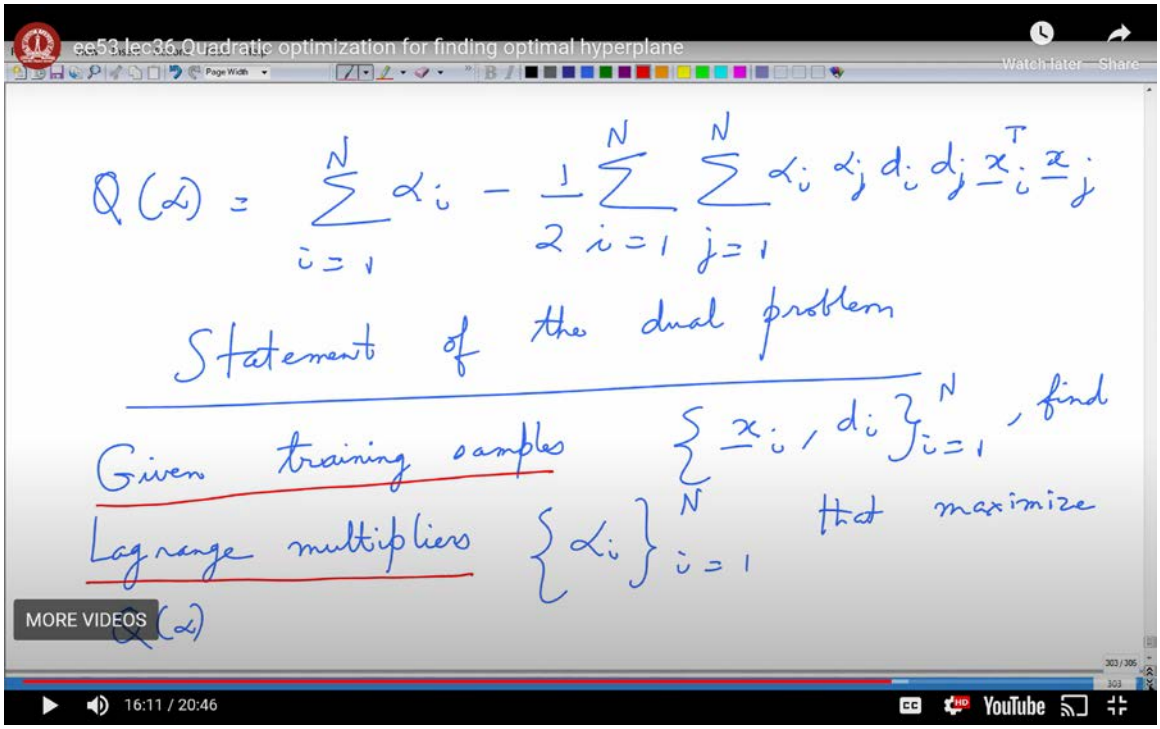
By expanding the primal problem and considering the optimality conditions, we obtained that the gradient of the Lagrangian with respect to the bias b set to zero gives us:

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

As a result, Term 3, which involves this sum, effectively vanishes. Therefore, our focus remains on:

- Term 1: The objective function $\frac{1}{2} W^T W$.
- Terms 2 and 4: These arise from the constraints and can be simplified further.

(Refer Slide Time: 16:11)



To compute $W^T W$, we substitute W with:

$$W = \sum_{i=1}^{N} \alpha_i d_i x_i,$$

where α_i and d_i are scalars, and x_i is the vector. Thus:

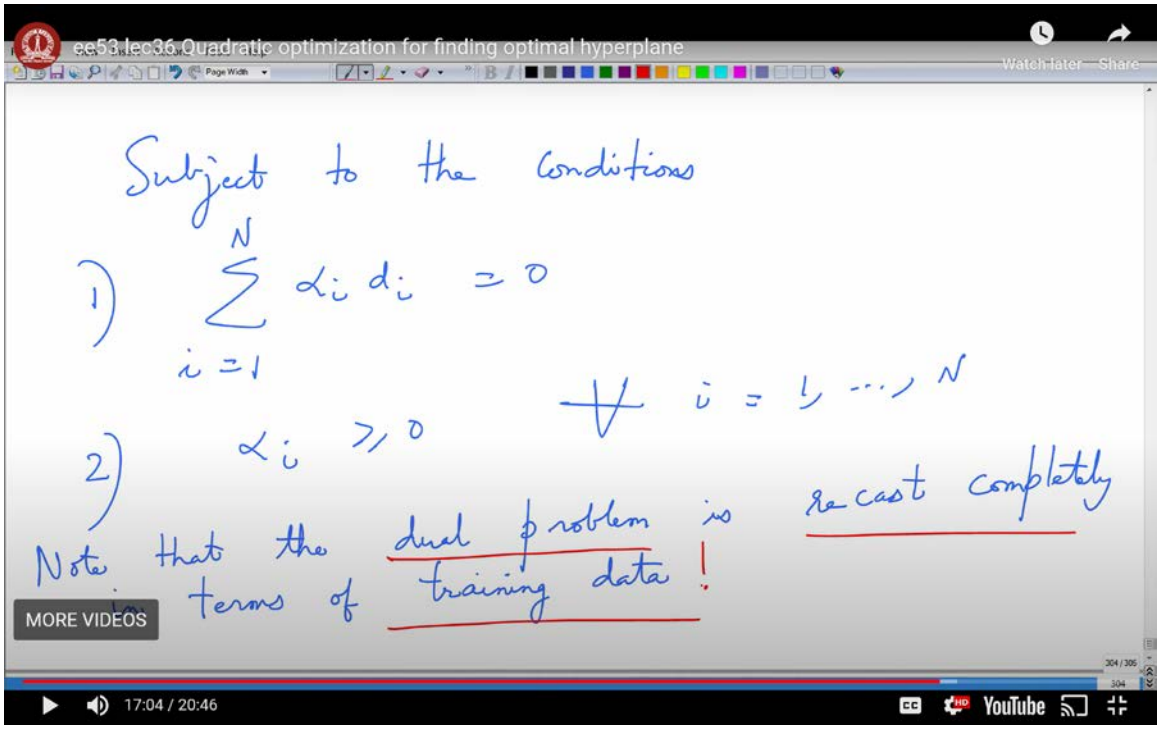
$$W^T W = \left(\sum_{i=1}^{N} \alpha_i d_i x_i\right)^T \left(\sum_{i=1}^{N} \alpha_i d_i x_i\right).$$

Expanding this:

$$W^T W = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j x_i^T x_j.$$

Here, α_i are non-negative, i.e., $\alpha_i \geq 0$.

(Refer Slide Time: 17:04)



We can now formulate our dual objective function. Let Q(α) be the dual objective function. It is expressed as:

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j x_i^T x_j.$$

This formulation is interesting because, while the primal problem depends on the weight vector W, we have eliminated W in favor of the Lagrange multipliers α_i and the training samples x_i and d_i. Our goal is to find the set of Lagrange multipliers α_i (for i = 1 to N) that maximizes Q(α).

Let's go over the conditions and steps involved in our support vector machine (SVM) optimization.

Conditions to Satisfy:

1. The first condition is:

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

This equation links the Lagrange multipliers α_i with the labels d_i.

2. The second condition is:

$$\alpha_i \geq 0 \text{ for all } i = 1,2, \dots, N$$

This ensures that the Lagrange multipliers are non-negative.

Given the training data and these constraints, our goal is to maximize Q(α). In this dual problem formulation, we optimize Q(α) subject to these conditions. This reformulation is crucial because it translates the problem from working directly with the weight vector W to working with the Lagrange multipliers α_i.

Obtaining the Optimum Weight Vector:

Once we determine the optimal Lagrange multipliers α_i^{opt} for each constraint, we compute the optimal weight vector W. For clarity, I will denote this as W^{opt}, but sometimes I might use just W or W^{OPT} interchangeably. The optimal weight vector W^{opt} is given by:

$$W^{\text{opt}} = \sum_{i=1}^{N} \alpha_i^{\text{opt}} d_i x_i$$

By substituting these optimal α_i values into the equation, we compute the weight vector W^{opt}.

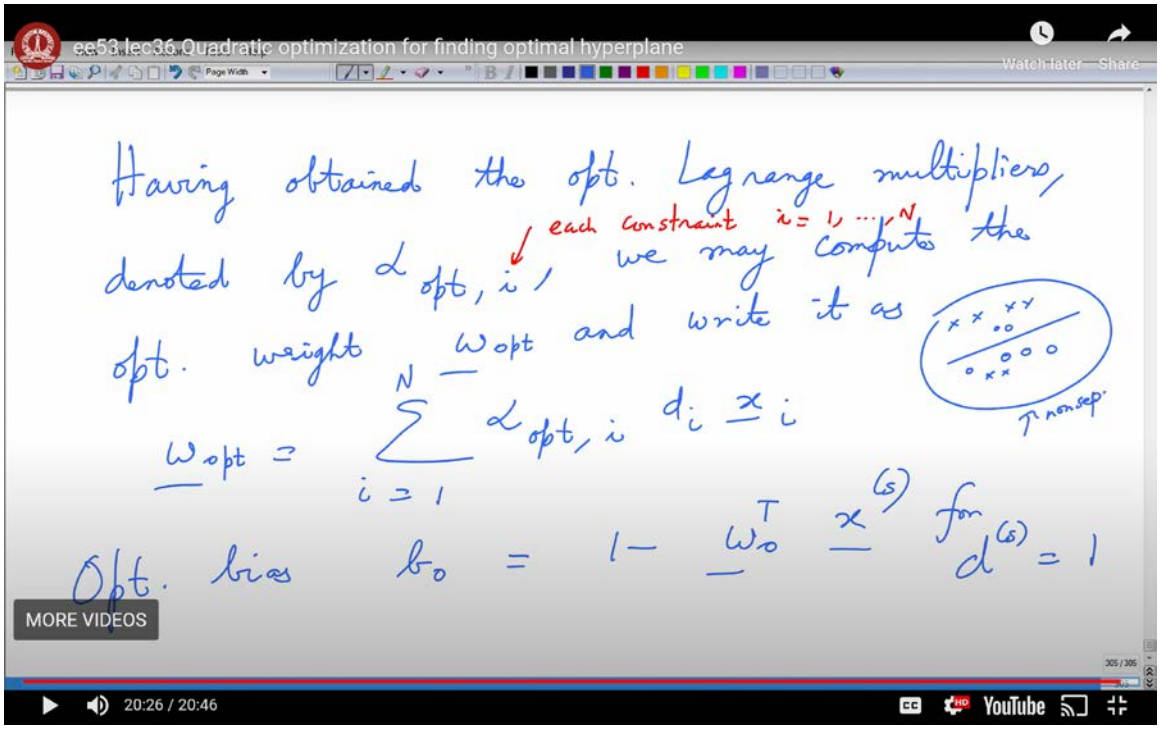
Computing the Bias b_0:

Once we have W^{opt}, the bias term b_0 can be calculated as:

$$b_0 = 1 - W^{\text{opt}} \cdot x_s$$

where d_s is 1 (for a support vector x_s with $d_s = 1$).

(Refer Slide Time: 20:26)



For Linearly Separable Patterns:

We've successfully formulated the SVM to solve for the optimal hyperplane. This involves finding the appropriate bias b_0 and weight vector W^{opt} using our training data (x_i, d_i). This

setup works well when the patterns are linearly separable, meaning we can find a hyperplane that perfectly separates the classes with zero error.

Handling Non-Linearly Separable Patterns. However, in cases where patterns are not linearly separable, it's impossible to find a hyperplane that separates all data points perfectly with zero error. For instance, some points from one class might end up on the wrong side of the decision boundary. Visualizing this, you might see instances where points from different classes are intermingled despite the best-fit hyperplane.

We will discuss how to handle such scenarios with non-zero errors and develop a linear decision boundary in the next segment.