Neural Networks for Signal Processing-I

Prof. Shayan Srinivasa Garani

Department of Electronic System Engineering

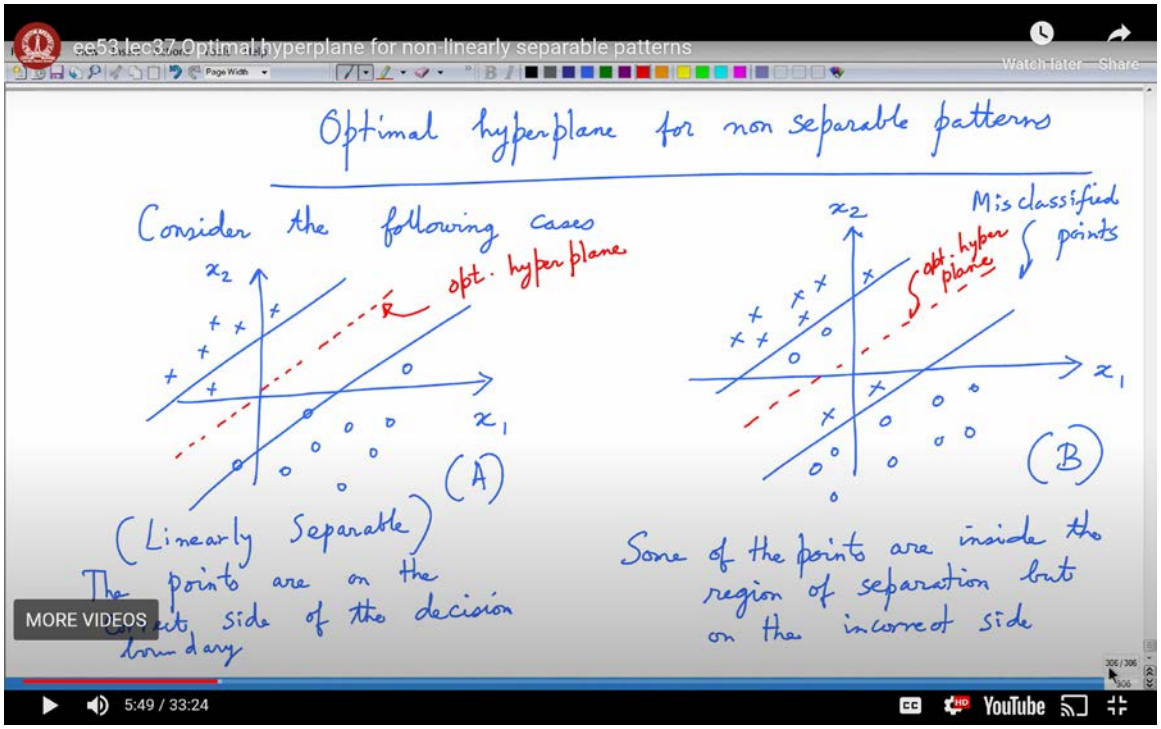
Indian Institute of Science, Bengaluru

Lecture – 37

Optimal Hyperplane for Non-Linearly Separable Patterns

In the last lecture, we derived the optimum hyperplane for linearly separable patterns, considering both the primal and dual formulations. We explored how to solve for the optimum hyperplane using the dual problem. Depending on your preference, you can choose either the primal or the dual route to optimize the problem and determine the optimum hyperplane.

(Refer Slide Time: 05:49)



Now, let's move forward and discuss how to derive an optimum hyperplane in the context of non-separable cases.

Consider a two-dimensional plane, say the x_1-x_2 plane. In this scenario, the points can be arranged in different configurations:

Case A: Linearly Separable Data

Here, we have an optimal hyperplane, indicated by the red line. Points belonging to class 1 and class 2, represented by stars and crosses respectively, are on the correct side of the decision boundary. In this case, the data is clearly linearly separable, with no misclassified points. All the points are on the correct side of the decision surface, resulting in zero errors.

Case B: Non-Separable Data

In this situation, while we still draw an optimal hyperplane, some points cross over to the other side of the decision boundary. For instance, crosses that belong to one class might spill over into the region meant for the other class, and vice versa. These points fall inside the margin but on the wrong side of the decision surface. This leads to misclassification, and thus, the error is not zero.

Given these scenarios, we now need to formally address the situation where data points are non-separable. To do so, we introduce a new set of non-negative scalar variables, denoted as ζ_i for i = 1 to n, into the definition of our hyperplane. These variables, known as slack variables, are incorporated into the hyperplane definition as follows:

$$d_i(W^{\mathrm{T}}x_i + \text{bias}) \geq 1 - \zeta_i \quad \text{for } i = 1,2,\dots,n$$

The slack variables ζ_i provide a measure of the deviation of a data point from the ideal conditions of pattern separability. In essence, they quantify how much a point violates the margin, allowing us to handle cases where data is not perfectly separable by the hyperplane.

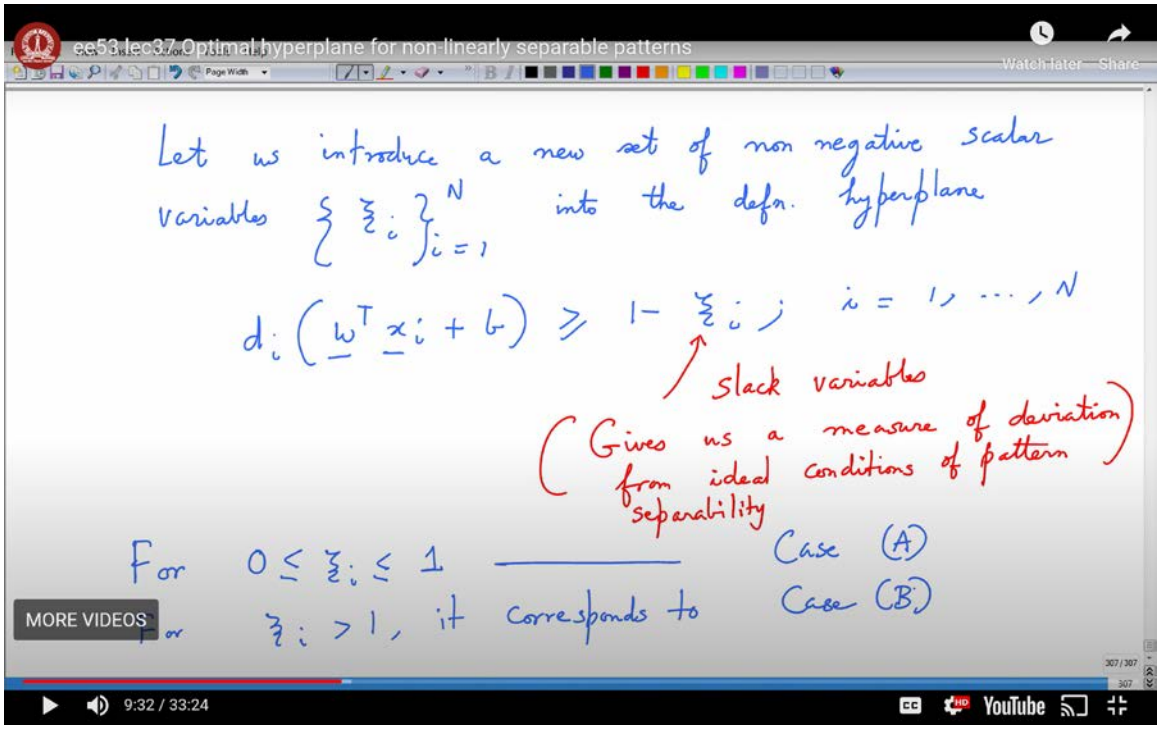
This concept gives us a way to measure how much each data point deviates from the ideal conditions of pattern separability. We can categorize this deviation into two cases:

Case A: When $0 \leq \zeta_i \leq 1$, the points are still on the correct side of the decision boundary.

Case B: When $\zeta_i > 1$, the points have crossed the optimum hyperplane, indicating misclassification, which corresponds to the second case.

Now, you might be curious about the location of the support vectors in this context. The support vectors are those data points that exactly satisfy the equality condition in the hyperplane equation, regardless of whether ζ_i is greater than zero. These are the critical points that lie on the margin, and they are key to defining the decision boundary.

(Refer Slide Time: 09:32)



The important task here is to find a separating hyperplane that minimizes the misclassification error when averaged over the entire training set. Since we are dealing with a situation where the data is not perfectly separable and some points are misclassified, our goal is to minimize this average error rate.

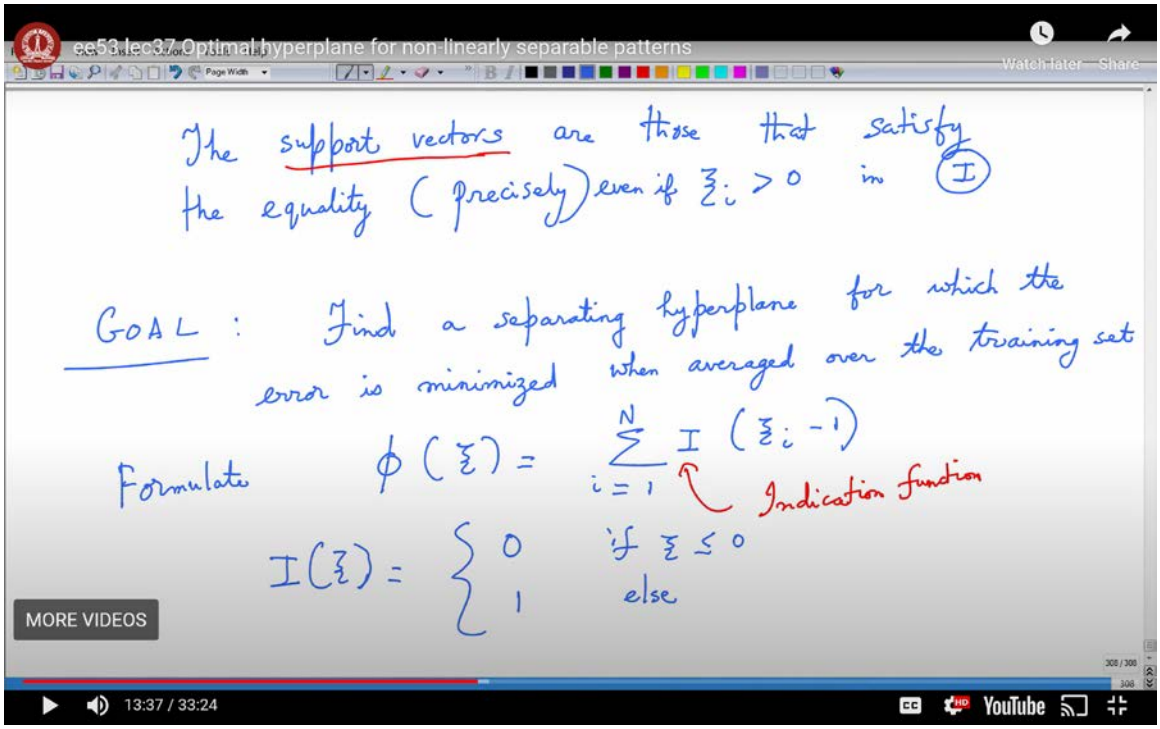
To achieve this, we formulate the problem by defining a function $\varphi(\zeta)$, which is expressed as:

$$\phi(\zeta) = \sum_{i=1}^{N} I(\zeta_i - 1)$$

Here, I(ζ) is an indicator function defined as:

$$I(\zeta) = \begin{cases} 0 & \text{if } \zeta \leq 0 \\ 1 & \text{otherwise} \end{cases}$$

(Refer Slide Time: 13:37)



This function φ(ζ) captures the error by counting the number of points that are misclassified (i.e., those for which $\zeta_i > 1$). However, dealing with a summation of indicator functions can be computationally challenging. To simplify the problem and make it more tractable, we approximate φ(ζ) as the summation of the ζ_i values:

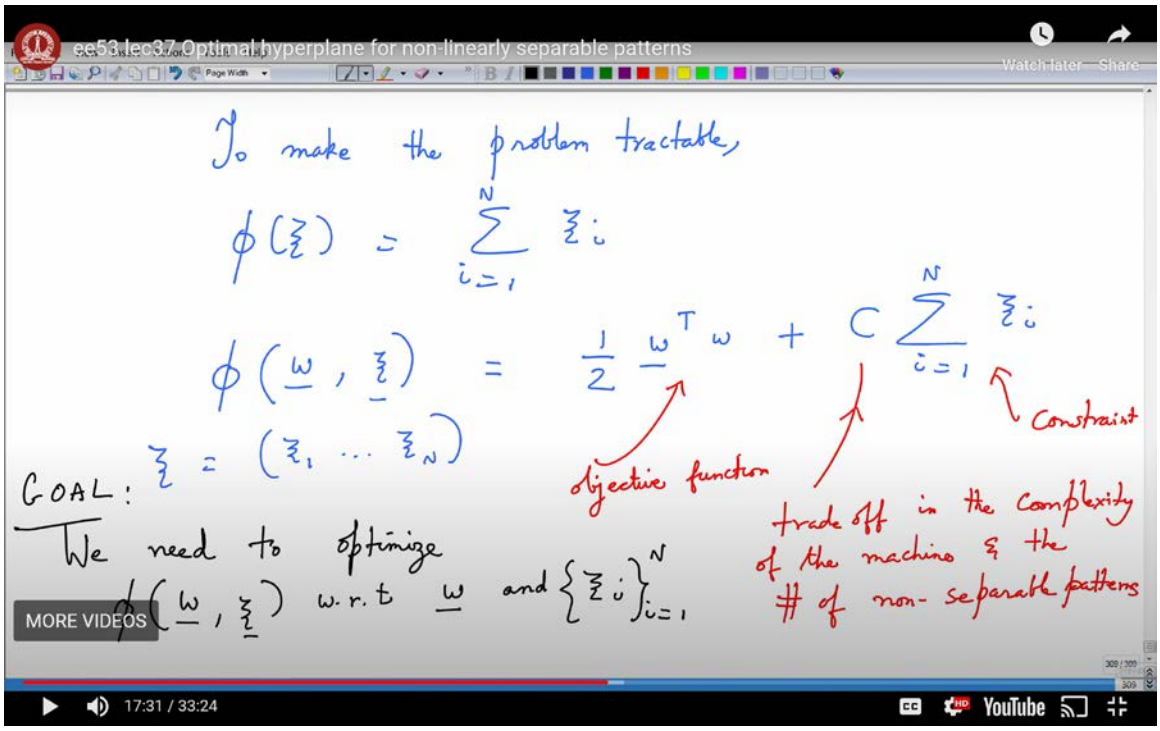
$$\varphi(\zeta) \approx \sum_{i=1}^{n} \zeta_i$$

This approximation allows us to reformulate the objective function in our optimization problem. The Lagrangian, which now includes both the weight vector W and the slack variables ζ, can be expressed as:

$$\frac{1}{2}W^{\mathrm{T}}W + C\sum_{i=1}^{n}\zeta_i$$

Here, C is a constant that controls the trade-off between the complexity of the model and the number of non-separable patterns. This C parameter plays a crucial role in the optimization process, and one might wonder how to choose an appropriate value for it. The choice of C will determine how the model balances the desire to minimize misclassification errors with the need to maintain a simple, generalizable decision boundary.

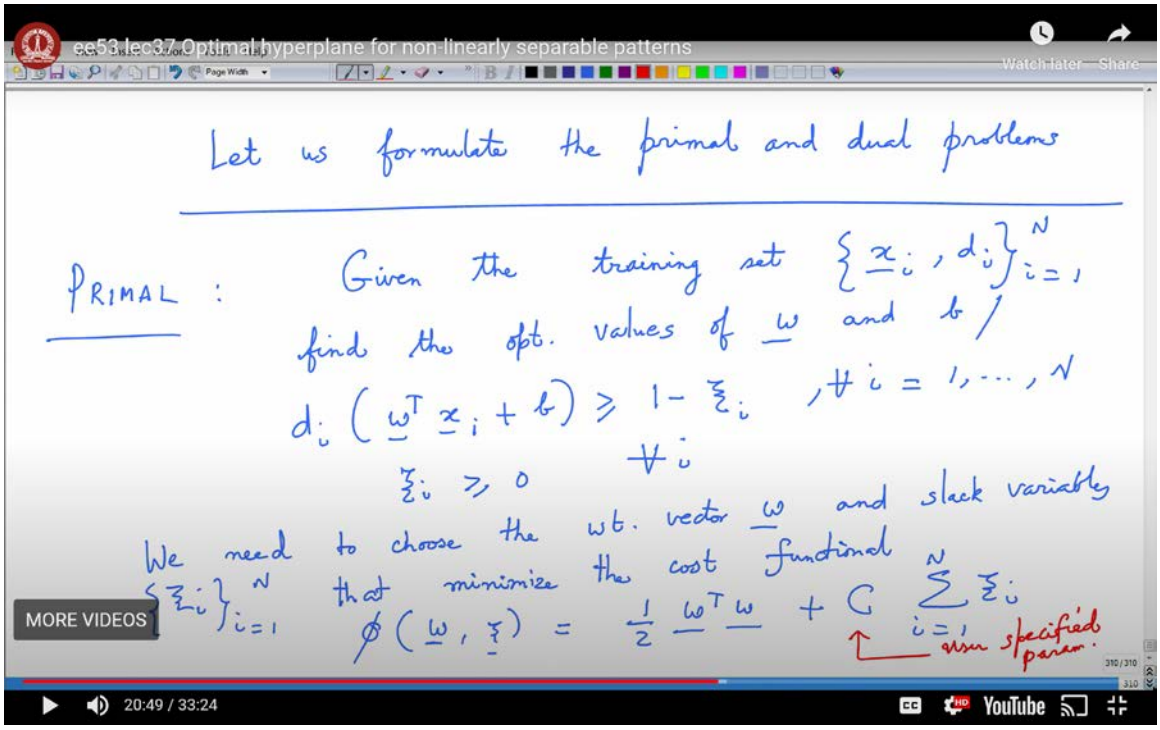
(Refer Slide Time: 17:31)



The parameter C can be determined experimentally through standard training and testing procedures, often involving some form of resampling. Alternatively, C can be chosen analytically by estimating a concept known as the VC dimension (Vapnik-Chervonenkis dimension), which I will discuss briefly. The VC dimension provides bounds on the

generalization performance of a model, which can guide the selection of C . The primary objective is to optimize the function φ(W, ζ) with respect to both the weight vector W and the slack variables ζ_i, where i ranges from 1 to n.

(Refer Slide Time: 20:49)



Now, let's move forward and formulate the primal and dual problems. We'll begin with the primal problem. Given a training set $\{x_i, d_i\}$ where i ranges from 1 to n, our goal is to determine the optimal values of the hyperplane parameters, the weight vector W and the bias B, such that the following condition holds:

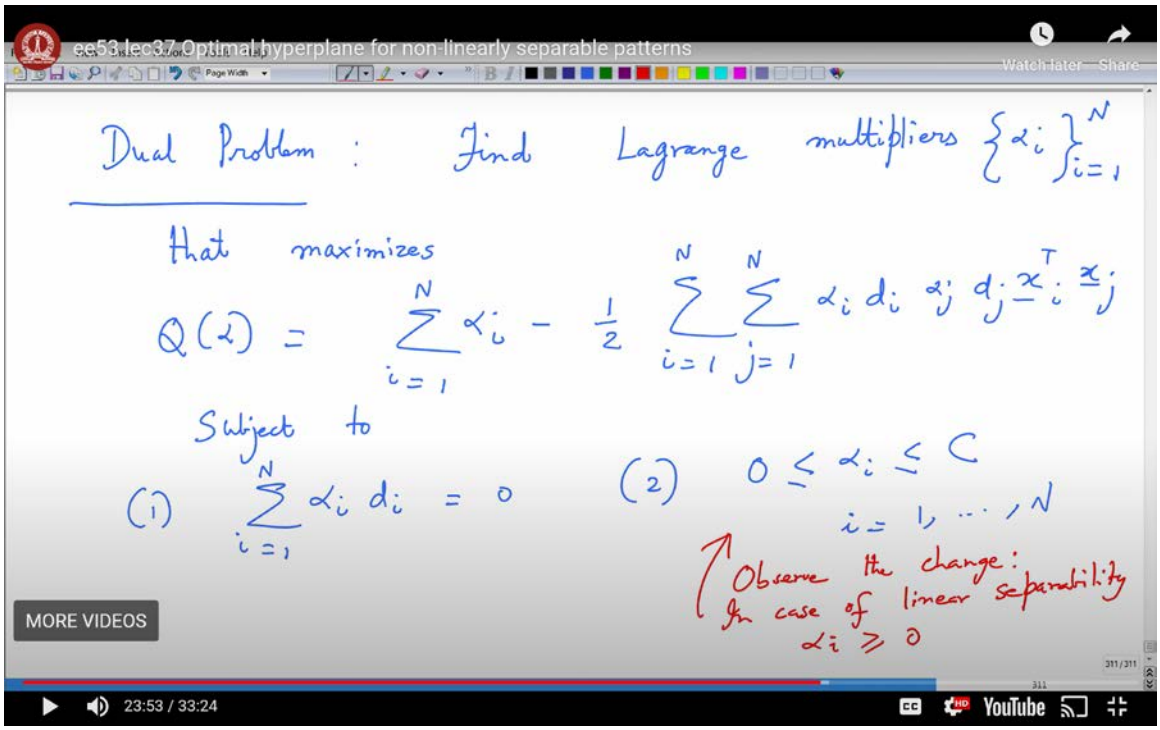
$$d_i \cdot (W^{\mathrm{T}} x_i + B) \geq 1 - \zeta_i \quad \text{for all } i = 1,2, \ldots, n,$$

with the additional constraint that $\zeta_i \geq 0$ for all i. This essentially means we need to select the weight vector W and the slack variables ζ_i in a way that minimizes the cost functional φ(W, ζ), defined as:

$$\varphi(W, \zeta) = \frac{1}{2} W^{\mathrm{T}} W + C \sum_{i=1}^{n} \zeta_i,$$

where C is a user-specified parameter. Following the same approach as before, we take derivatives with respect to W, B, and ζ_i, apply the constraints (which are inequalities), and then formulate the Lagrangian. By taking partial derivatives of the Lagrangian with respect to W, B, and ζ, and setting them equal to zero, we derive the conditions needed to solve the problem. Just as we did for the linearly separable case, we can now formulate the dual problem for the non-separable case.

(Refer Slide Time: 23:53)



The dual problem is as follows: We need to find the Lagrange multipliers α_i, where i ranges from 1 to n, that maximize the function $Q(\alpha)$:

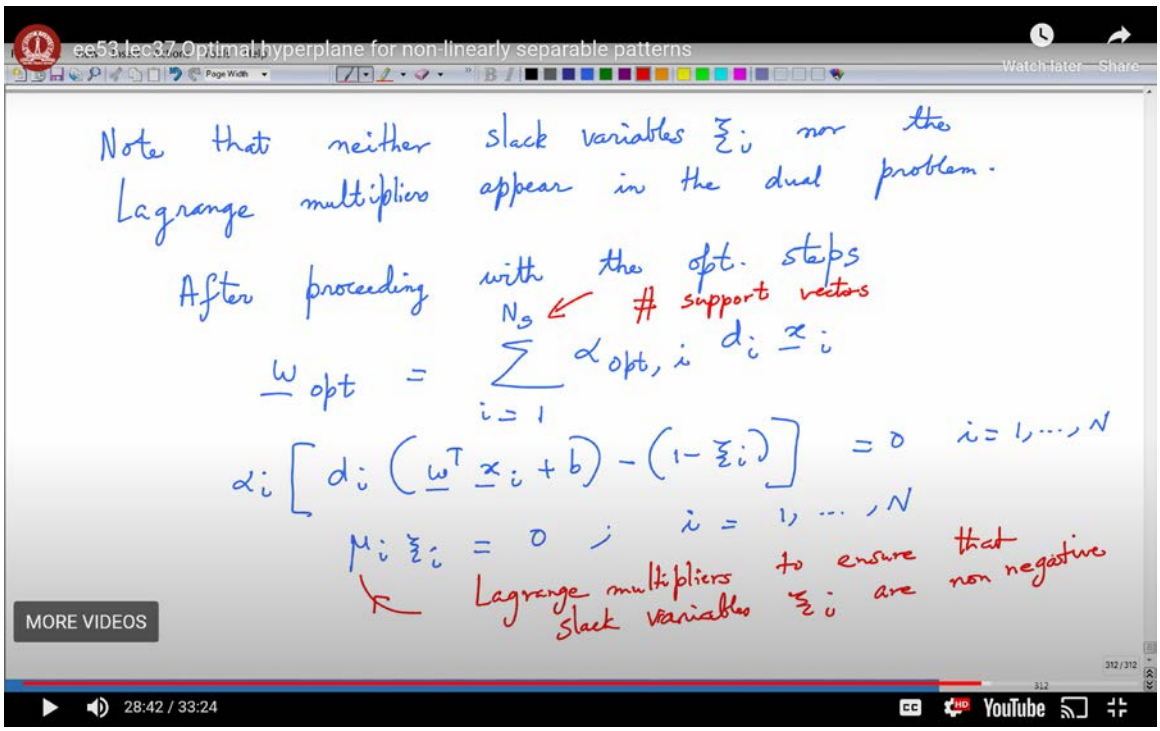
$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i d_i \alpha_j d_j (x_i^T x_j),$$

subject to the conditions:

$$\sum_{i=1}^{n} \alpha_i d_i = 0 \quad \text{and} \quad 0 \leq \alpha_i \leq C \quad \text{for all } i = 1,2,\ldots,n.$$

Notice the change here: In the linear separability case, the α_i values were simply non-negative ($\alpha_i \geq 0$), but now, due to the introduction of non-separable data, the α_i values are constrained by the user-specified constant C. This adjustment reflects the need to minimize misclassification over the training samples while accounting for the non-separable nature of the data.

(Refer Slide Time: 28:42)



I have deliberately skipped the detailed derivation of the dual problem because it follows the same steps as in the linearly separable case. Once you set up the primal problem and know the constraints, you can follow the exact same procedure to arrive at the dual problem. You will likely practice this in your homework exercises. An important thing to note is that neither the slack variables ζ_i nor the Lagrange multipliers appear in the dual problem directly.

The only subtle difference in this approach compared to the separable case is how we bound α_i between 0 and C, whereas in the separable case, α_i was simply required to be greater than or equal to 0. We follow the same optimization process: taking the partial derivatives

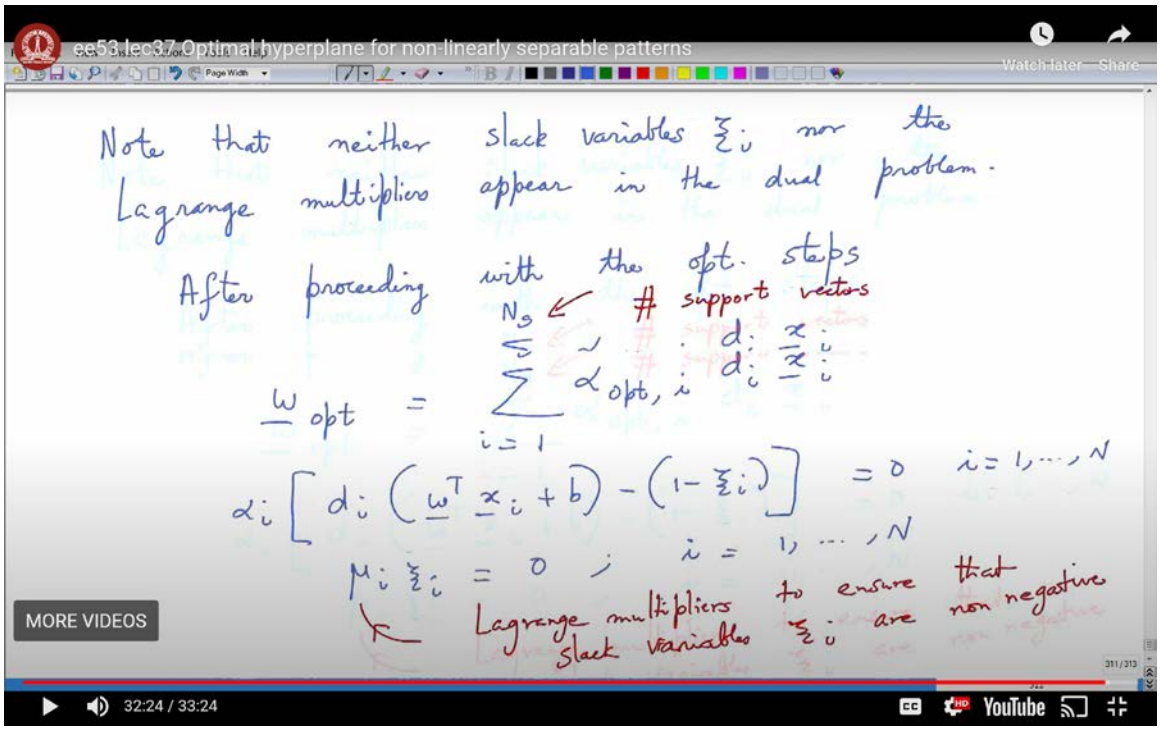
with respect to the variables in the Lagrangian formulation, setting them equal to 0, and then solving these equations to find the optimum value of the vector W.

This optimal vector W is given by the summation:

$$W_{\text{opt}} = \sum_{i=1}^{n} \alpha_i^{\text{opt}} d_i x_i,$$

where n_s represents the number of support vectors, which are crucial because they satisfy the constraints with equality. In our formulation, we encounter an inequality constraint involving ζ_i and an equality constraint, where the equality holds even if ζ_i is greater than 0. The vectors that satisfy this condition are the support vectors, and they are essential for computing the optimum hyperplane.

(Refer Slide Time: 32:24)



The key equation that governs this scenario is:

$$\alpha_i[d_i(W^{\mathrm{T}}x_i + B) - (1 - \zeta_i)] = 0 \quad \text{for all } i = 1,2, \ldots, n,$$

and

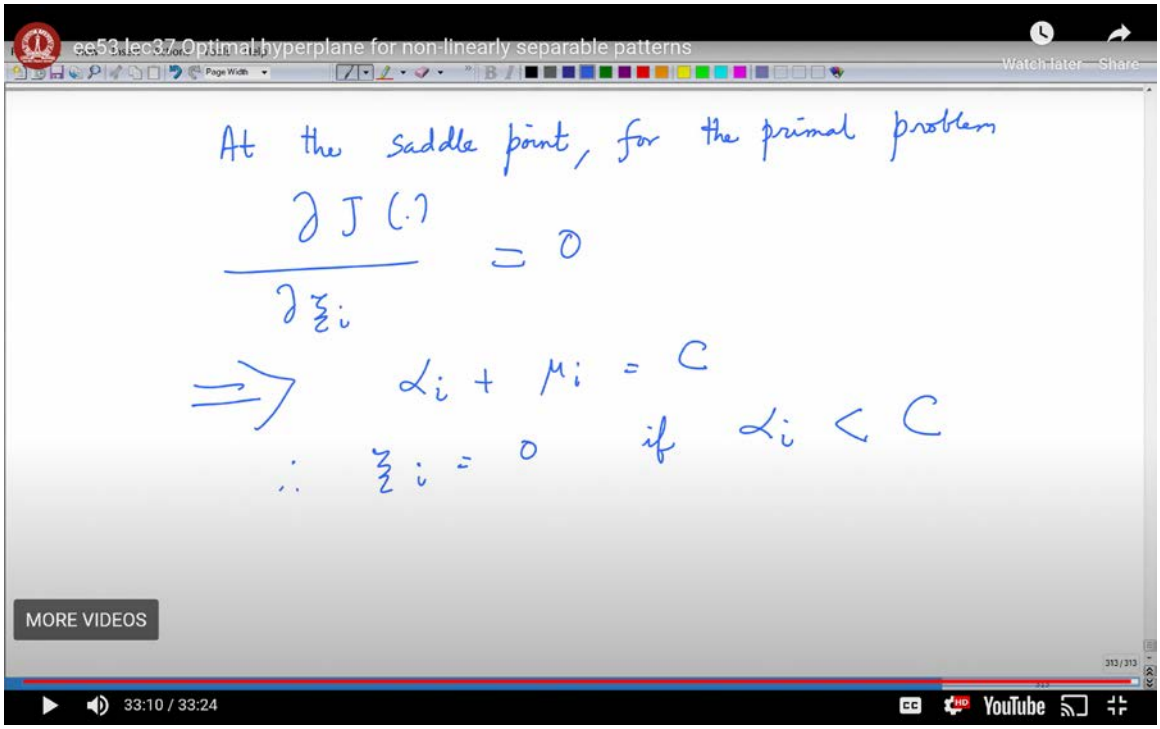
$$\mu_i \zeta_i = 0 \quad \text{for all } i = 1,2,\dots,n,$$

where μ_i are the Lagrange multipliers introduced to ensure that the slack variables ζ_i remain non-negative. This condition is vital as it ensures that the slack variables, which account for misclassification, stay within acceptable bounds.

In the primal problem, at this critical point, the derivative of the Lagrangian with respect to ζ_i is zero. This leads us to the important relationship:

$$\alpha_i + \mu_i = C \quad \text{and} \quad \zeta_i = 0 \text{ if } \alpha_i < C.$$

(Refer Slide Time: 33:10)



These relationships can be explored further in your homework exercises. Conceptually, the key takeaway here is the introduction of the slack variables ζ_i. These variables allow us to accommodate misclassified points and focus on minimizing the average cost associated with these misclassifications. Consequently, the cost functional we aim to minimize is formulated as:

$$\frac{1}{2}W^{\mathrm{T}}W + C\sum_{i=1}^{n}\zeta_i,$$

subject to the constraints. This differs from the linearly separable case, where the objective was simply $\frac{1}{2}W^{\mathrm{T}}W$ without the slack variables.

Thus, our objective function now has two components: a convex function of W ($\frac{1}{2}W^{\mathrm{T}}W$) and a term that depends on the slack variables ζ_i. This additional term effectively minimizes the number of misclassified points averaged over the entire dataset. From here, we formulate the dual problem following the same steps as in the linearly separable case. With this understanding, we can compute the optimum hyperplane using the set of support vectors that satisfy the equality conditions outlined above.

This completes our discussion on non-linearly separable patterns. The detailed derivations of the dual problem setup, including the formulation of the Lagrangian for the primal problem and setting the partial derivatives to zero, have been omitted here but will be part of your homework exercises. Interested readers are encouraged to work through these details independently, as they are essential for a deep understanding. Lastly, note that the partial derivative with respect to ζ_i (denoted as $\frac{\partial J}{\partial \zeta_i}$) should be zero, which is a crucial detail to keep in mind. We will stop this lecture at this point.