Neural Networks for Signal Processing-I

Prof. Shayan Srinivasa Garani

Department of Electronic System Engineering

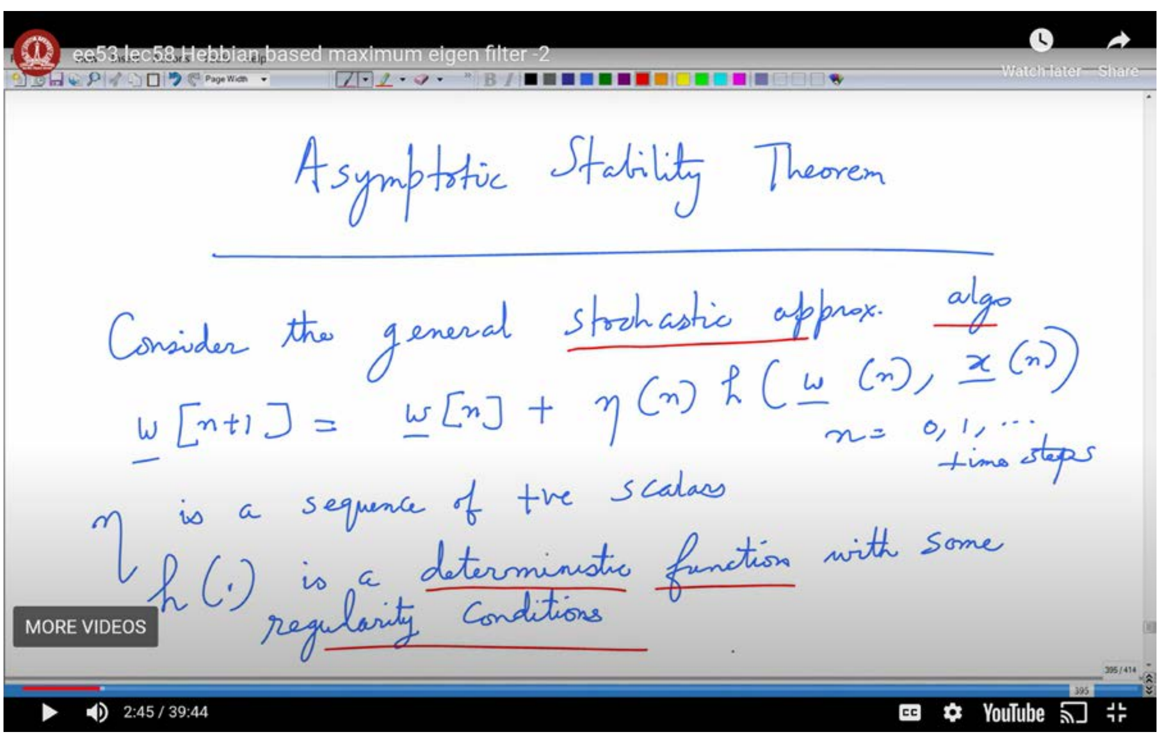
Indian Institute of Science, Bengaluru

Lecture – 58

Hebbian based Maximum Eigen Filter -2

In this module, we will explore the Asymptotic Stability Theorem, which is essential for analyzing non-linear stochastic difference equations. While normalization has proven effective in practice, ensuring stability and boundedness, the real challenge lies in analyzing this stability rigorously. For linear systems, stability analysis is relatively straightforward. You simply examine the governing differential or difference equations, compute the eigenvalues, and check if they are less than 1 to determine stability. However, non-linear differential equations introduce complexity that requires a more nuanced approach.

(Refer Slide Time: 02:45)

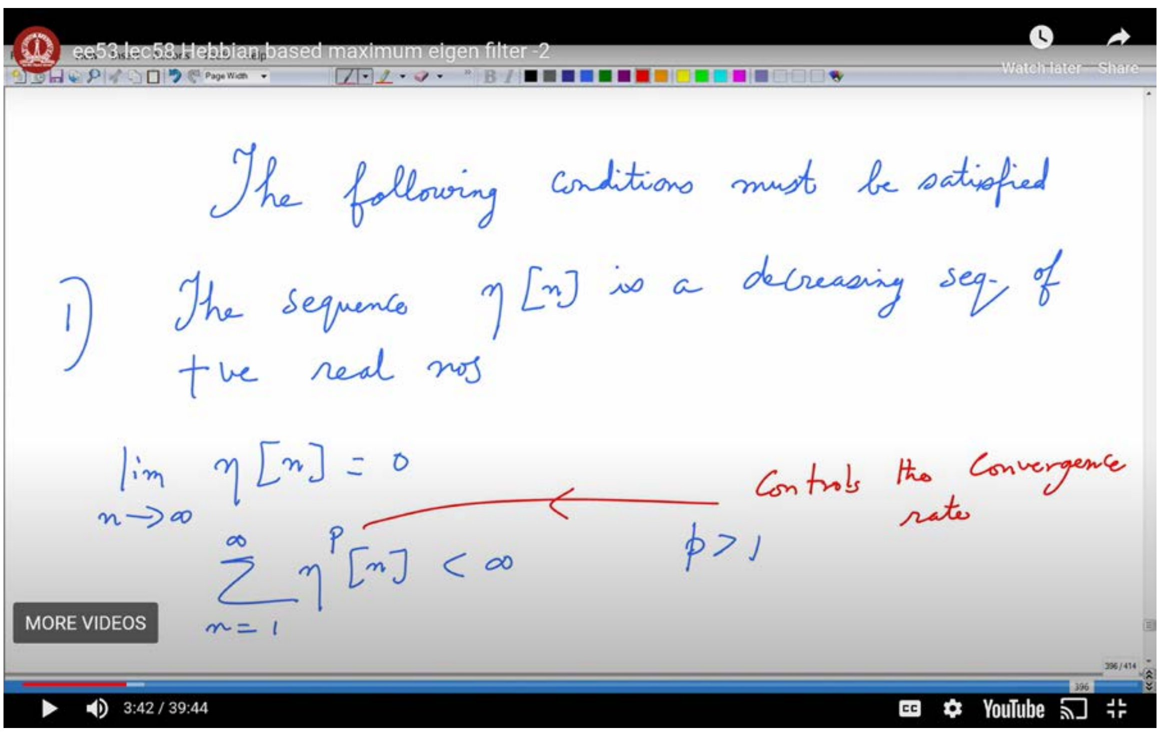


To address this, we need to delve into the Asymptotic Stability Theorem. This theorem provides the conditions necessary for assessing stability in such systems. Let's start by considering a general stochastic approximation algorithm, which can be represented by the update rule:

$$W_{n+1} = W_n + \eta_n \cdot h(W_n, x_n)$$

Here, W_n is the weight vector at time step n, η_n is the learning rate that may vary with n, and $h(W_n, x_n)$ is a function that depends on the weight and input at time n. The index n ranges from 0, 1, and so forth, until convergence is achieved. In this context, η_n is a sequence of positive scalars (learning rates), and h is a deterministic function with certain regularity conditions.

(Refer Slide Time: 03:42)

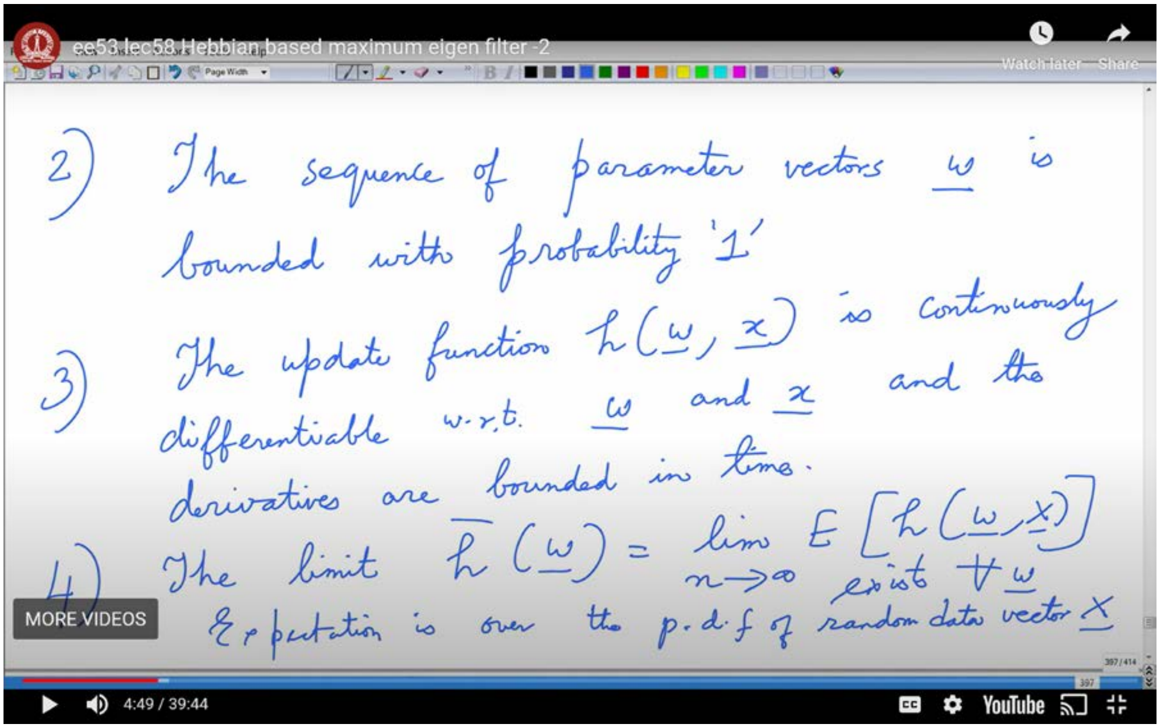


Now, let's outline the conditions required for the Asymptotic Stability Theorem:

1. Decreasing Learning Rates: The sequence η_n must be a decreasing sequence of positive real numbers, converging to 0 as n approaches infinity. Additionally, the sum of η_n^p (where

p is a power greater than 1) should be finite. This sum must be less than infinity. The power p essentially controls the convergence rate of the algorithm, which is crucial for stability.

(Refer Slide Time: 04:49)

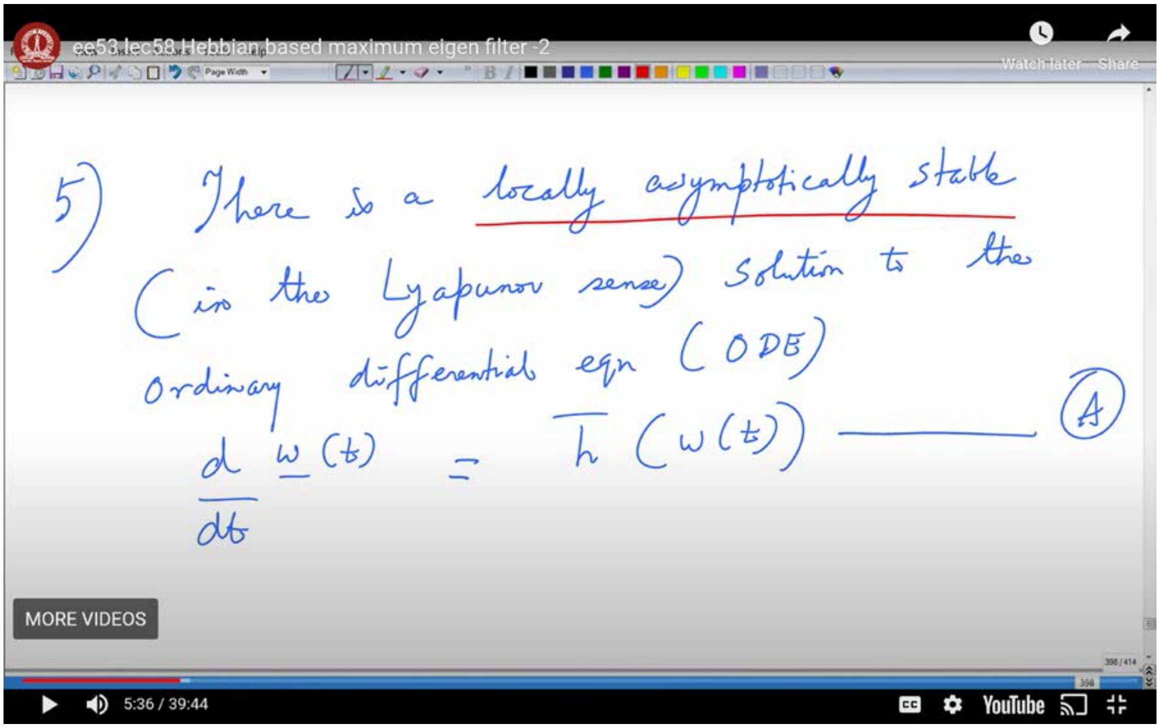


2. Bounded Parameter Sequence: The sequence of parameter vectors W_n must be bounded with probability 1. Since we are normalizing the weight vector at every time step, this condition ensures that the sequence remains bounded.

3. Continuous Differentiability: The update function h, which depends on both the weight W and the input x, must be continuously differentiable with respect to both W and x. Moreover, the derivatives of this function should be bounded over time. This regularity condition is important for ensuring the smoothness and predictability of the update process.

4. Existence of the Limit: The limit of h(W), denoted as $\bar{h}(W)$, which is the expectation of the function h as n approaches infinity, must exist for all weight vectors W. The expectation is taken with respect to the probability density function of the random vector x.

(Refer Slide Time: 05:36)

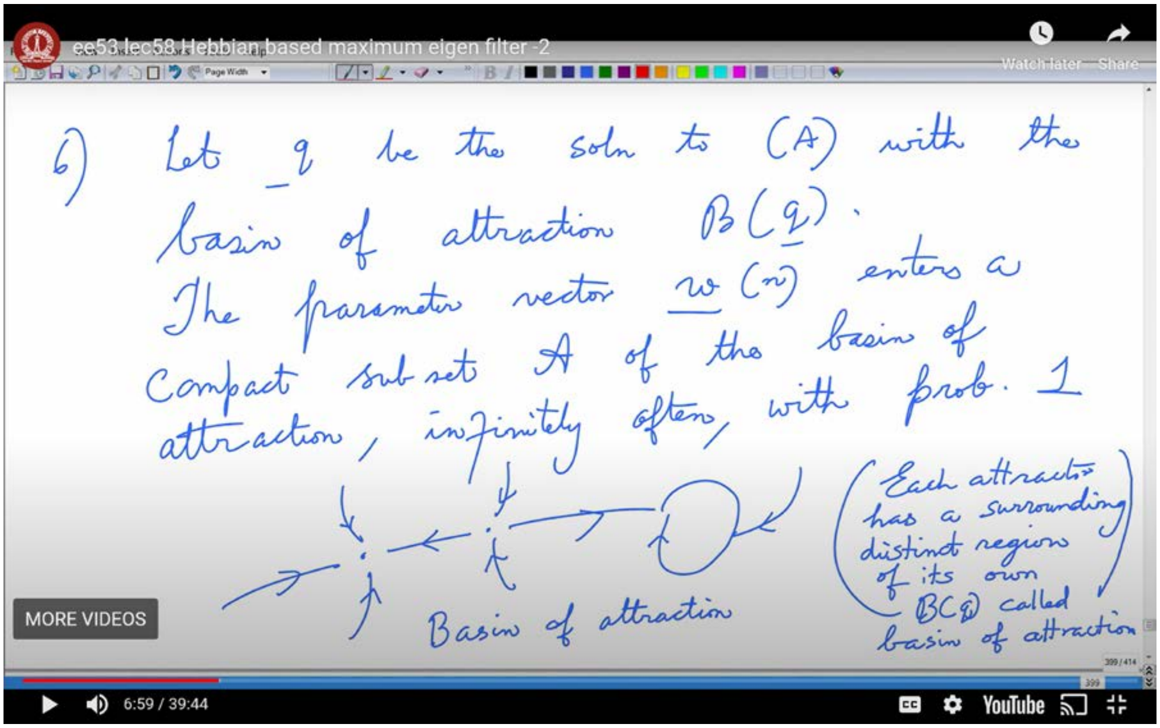


By satisfying these conditions, we can rigorously analyze the stability of the system under the Asymptotic Stability Theorem, and we will see how these conditions apply to Oya's update rule.

In neural networks, we are given data vectors x, and when we compute statistics, we are essentially dealing with the statistics of these input data vectors. This means that we need to understand the concept of local asymptotic stability in the Lyapunov sense for the solution to the ordinary differential equation described by equation A. The derivative of W(t) approaches $\bar{h}(W(t))$ in the limit, indicating that there is a locally asymptotically stable solution to this differential equation.

If condition φ is satisfied, it forms one of the requirements for the asymptotic stability theorem. Let Q be the solution to equation A with a corresponding basin of attraction B. To explain the concept of a basin of attraction, consider a non-linear dynamical system that has multiple attractors and repellers.

(Refer Slide Time: 06:59)

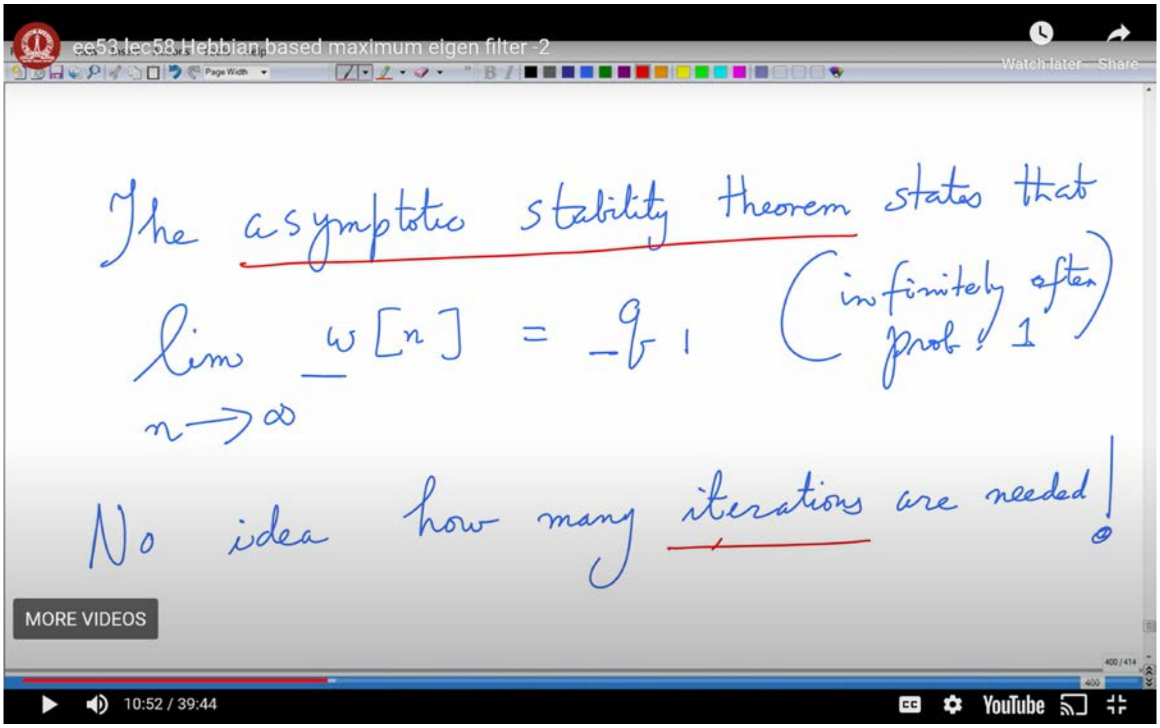


Imagine a river with several whirlpools. As you approach a whirlpool, you are drawn into it. If there are multiple whirlpools in the river, your trajectory depends on your initial conditions. You might end up being caught by one of the whirlpools, and the area around each whirlpool where this happens is known as the basin of attraction. The whirlpool itself acts as an attractor.

This concept is illustrated in a diagram where the attractors are marked as blue dots, with distinct regions around them representing their basins of attraction. The parameter vector W(n) will enter a compact subset defined by $\mathcal{A}$ (the basin of attraction) infinitely often with probability 1. This implies that, despite various random realizations, the trajectory will almost certainly land in the basin of attraction, approaching the attractor.

In contrast to linear dynamical systems, where stability can be assessed by eigenvalues (with values less than 1 indicating stability and greater than 1 indicating instability), non-linear dynamical systems require more nuanced analysis. For linear systems, eigenvalues provide straightforward insights into stability, oscillatory behavior, or unbounded growth.

(Refer Slide Time: 10:52)

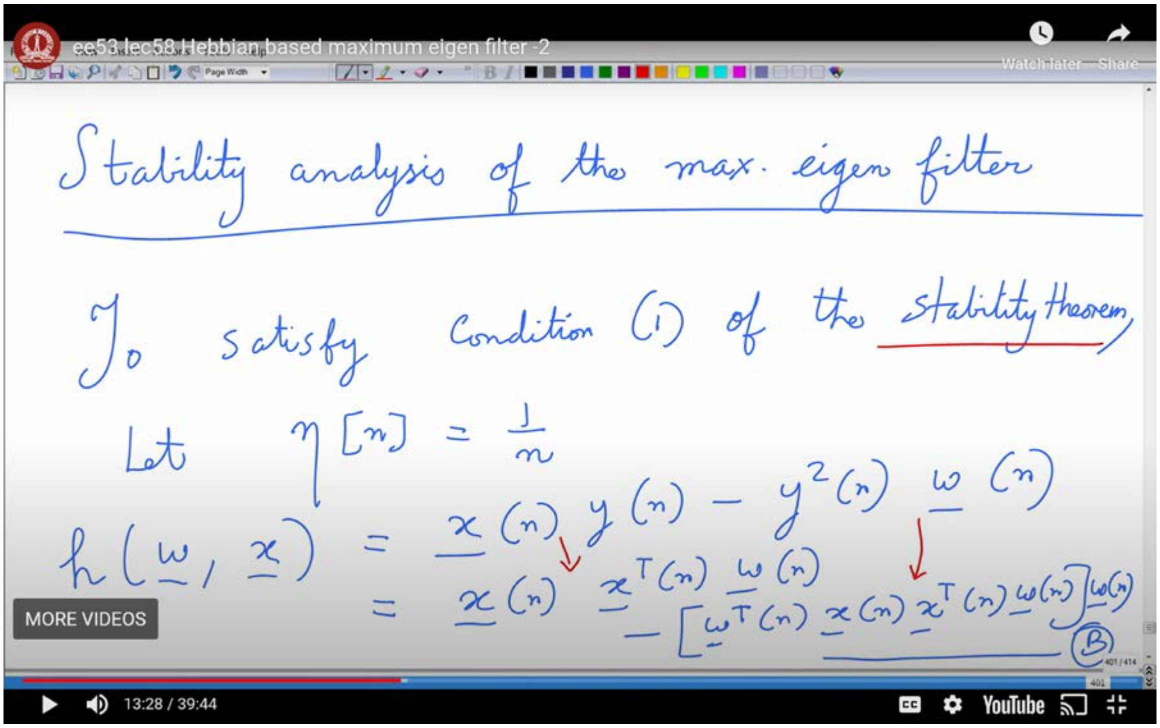


The asymptotic stability theorem states that as n approaches infinity, the weight vector W(n) converges to a vector q_1 with probability 1. However, this theorem does not specify the exact number of iterations required to achieve this convergence. While it guarantees that W(n) will asymptotically approach q_1, the finite number of iterations, denoted as N, needed to ensure this convergence is not directly addressed by the theorem.

So, with this understanding, we are now ready to dive into the stability analysis of the maximum eigen filter. We have approached this in three distinct parts. First, we formulated a non-linear stochastic difference equation based on the normalization applied to the Hebbian update. Second, we examined the conditions required for the asymptotic stability theorem, focusing on the behavior of the weight vector in the limit. Now, given our update equation derived from Oya's rule, we need to analyze its stability with respect to these conditions.

To satisfy the first condition of the stability theorem, we can choose the weights η(n) to follow a sequence like 1/n. This choice ensures that the governing function h(w, x) is defined as $x(n) \cdot y(n) - y^2(n) \cdot w(n)$. Here, y(n) is a scalar, x is a vector, and y(n) is expressed as the inner product of x and w. Consequently, $x(n) \cdot x^T(n) \cdot w(n)$ becomes our first term, while the second term, $y^2(n) \cdot w(n)$, simplifies to the expression $w^T(n) \cdot R \cdot w(n)$, where R is the correlation matrix.

(Refer Slide Time: 13:28)



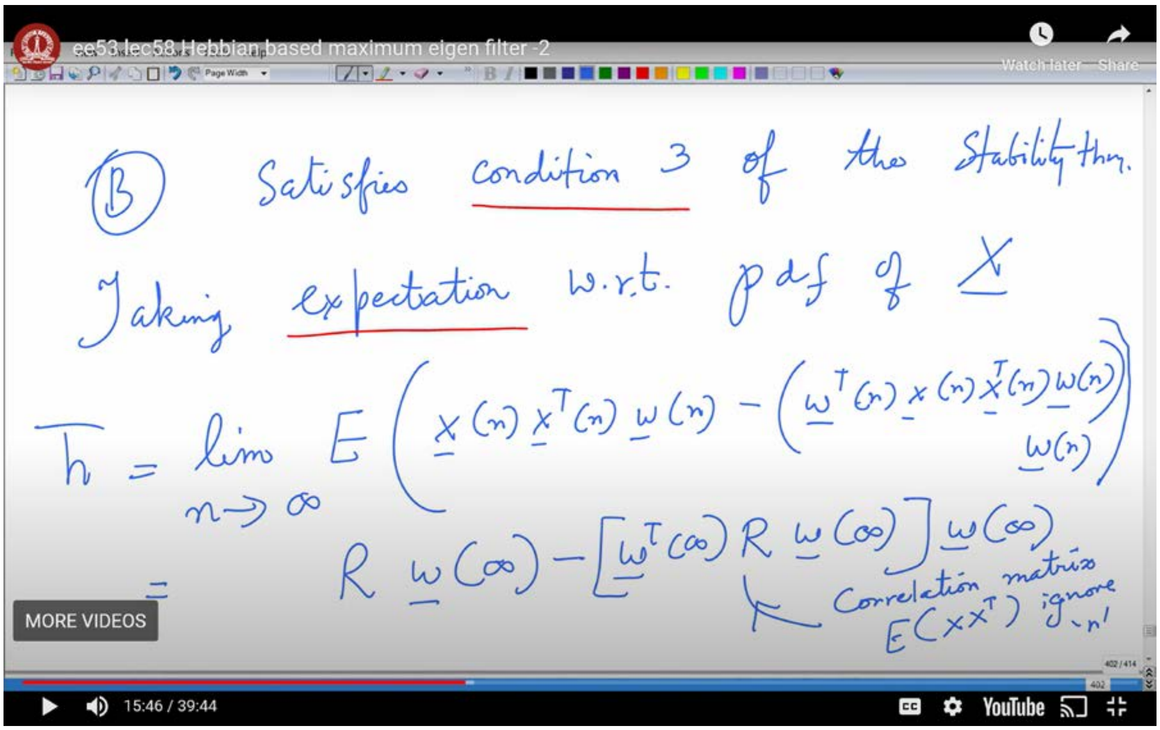
It is essential to recall that y(n) can be expressed as $w^T \cdot x$ and also as $x^T \cdot w$, reflecting a symmetry in our formulation. Let us denote this equation as B. This equation satisfies the third condition of the stability theorem, as it is continuously differentiable with respect to w and x, and its derivatives are bounded over time.

Next, we take the expectation of this equation with respect to the probability density function (PDF) of x. It is crucial to note that the input is a stationary process, meaning the PDF of the data vector x remains constant over time. This stationarity simplifies the analysis significantly.

We can express $\bar{h}$ as the limit of the expectation of the equation involving x(n), x^T(n), w(n), and w^T(n) as n approaches infinity. The expectation is taken over x, not w. This allows us to pull w out of the expectation, resulting in:

$$\bar{h} = \lim_{n \to \infty} [x(n) \cdot x^T(n) \cdot w(n) \cdot w^T(n) - y^2(n) \cdot w(n) \cdot w^T(n)]$$

(Refer Slide Time: 15:46)



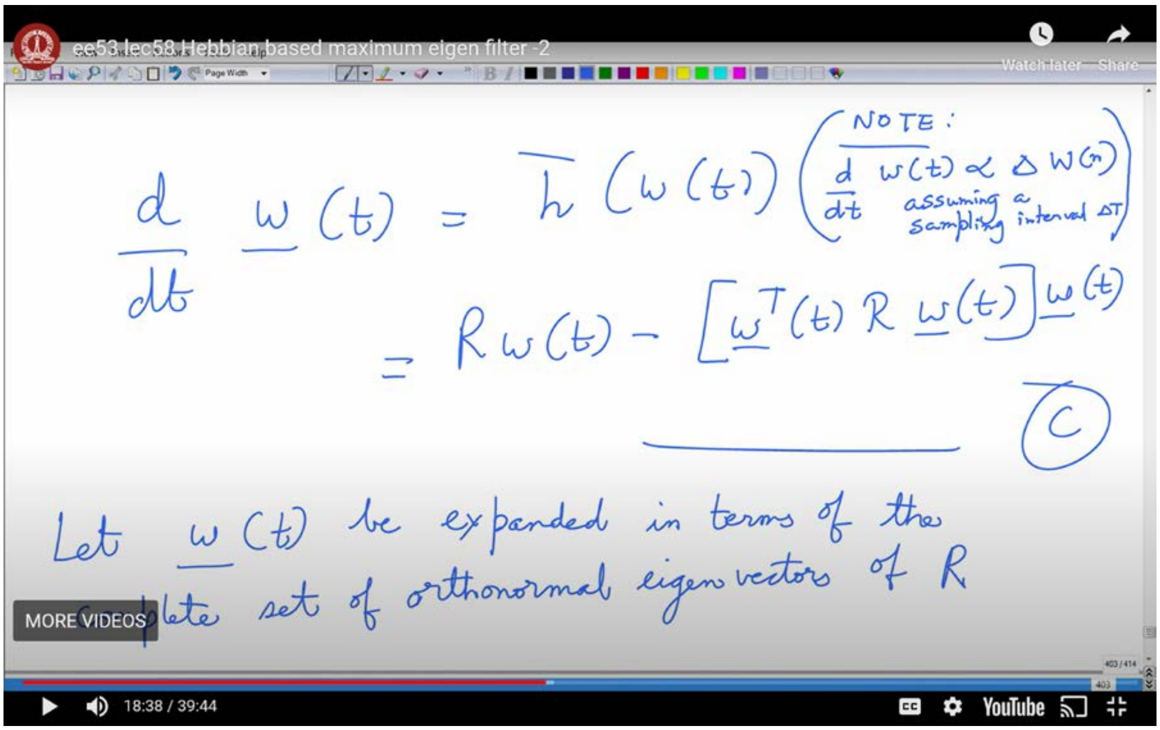
Given that the input is stationary, the expectation simplifies to the correlation matrix R multiplied by w(n) evaluated at infinity:

$$\bar{h} = w(\infty) \cdot R \cdot w(\infty)$$

This simplification is possible because the correlation matrix is constant over time. If the input were not stationary, we would have to account for time-dependent correlation matrices, adding complexity to the analysis.

Moving on, consider the differential equation $\frac{dW(t)}{dt} = \bar{h}(W(t))$. This differential operator $\frac{d}{dt}$ is a linear differential operator, assuming a sampling interval Δt. In the discrete case, the difference vector Δ W(n) is W(n) - W(n-1). This discrete difference operation relates to the continuous derivative $\frac{d}{dt}$, assuming the sampling interval. Hence, $\frac{dW(t)}{dt}$ is proportional to Δ W(n), where Δ W(n) represents the change in the weight vector over one time step. This detail, while subtle, is crucial for linking discrete updates to continuous dynamics.

(Refer Slide Time: 18:38)



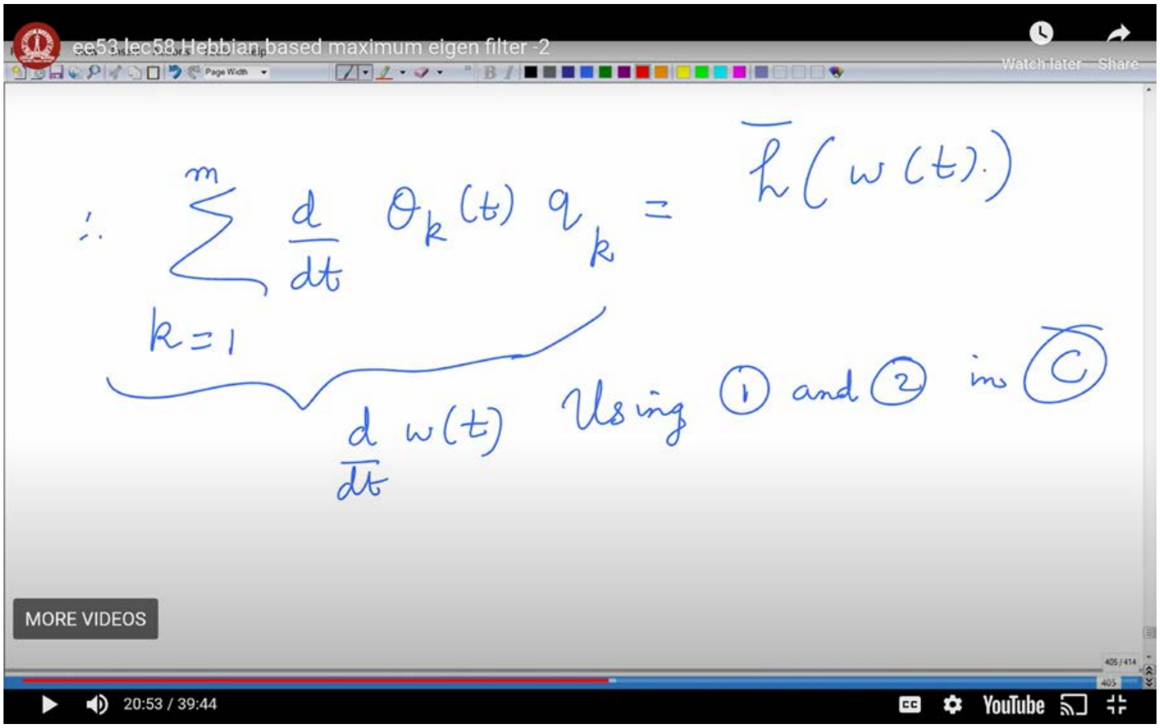
Therefore, transitioning from a difference equation to a differential equation simplifies our analysis. I can write the differential equation as $\frac{d}{dt}W(t) = \bar{h}\big(W(t)\big)$. I previously mentioned that moving from a difference equation to a differential equation, particularly in this case involving a non-linear differential equation, is often more manageable. Non-linear differential equations can be easier to tackle when approached from this perspective.

Now, let's express $\frac{d}{dt}W(t)$ in terms of the correlation matrix R. Specifically, we have:

$$\frac{d}{dt}W(t) = RW(t) - \big(W^T(t)RW(t)\big)W(t)$$

To simplify this, let's perform an eigen decomposition of the correlation matrix R. We can expand W(t) using a complete set of orthonormal eigenvectors of R. This means expressing W(t) in terms of its eigenvectors q_k and the corresponding time-varying coefficients $\theta_k(t)$.

(Refer Slide Time: 20:53)



So, W(t) can be written as:

$$W(t) = \sum_{k=1}^{m} \theta_k(t) q_k$$

where $\theta_k(t)$ represents the time-varying projection of W(t) onto the eigenvector q_k. The eigen decomposition involves the eigenvalue equation:

$$Rq_k = \lambda_k q_k$$

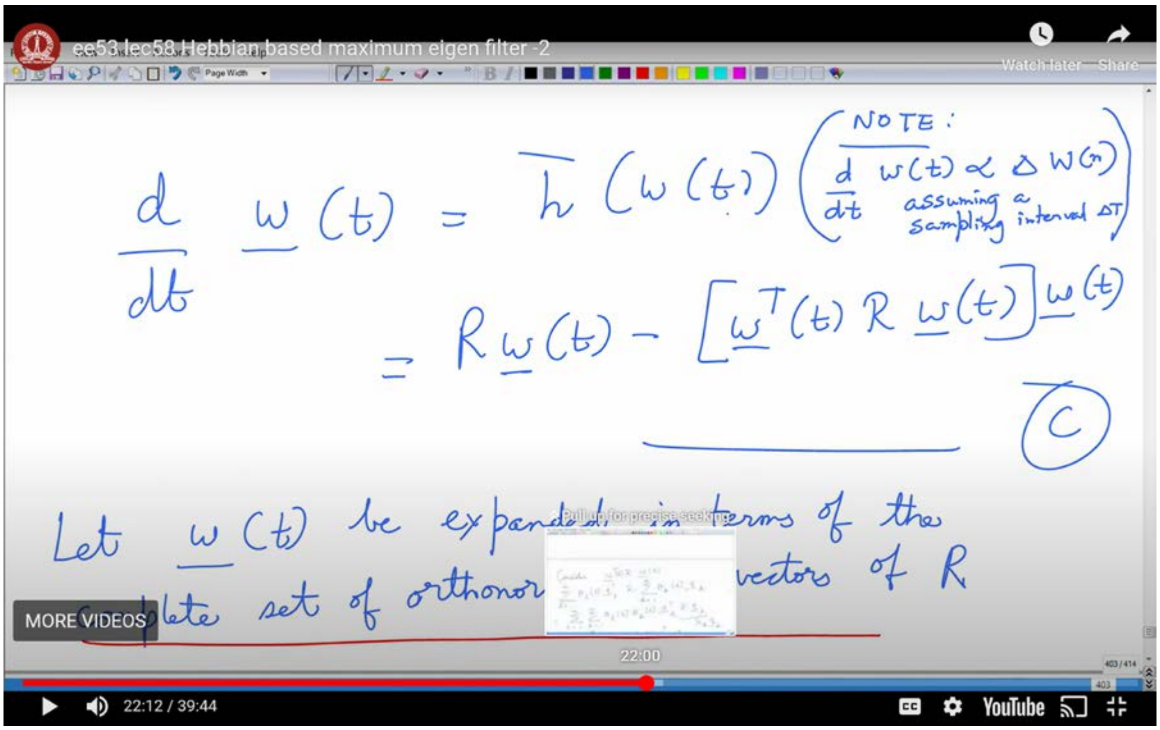
where q_k is the eigenvector and λ_k is the eigenvalue. The eigenvalue λ_k can be expressed as:

$$\lambda_k = q_k^T R q_k$$

Now, returning to our differential equation, since W(t) is expanded in terms of eigenvectors q_k and time-varying coefficients $\theta_k(t)$, we can simplify the left side of our equation:

$$\frac{d}{dt}W(t) = \sum_{k=1}^{m} \frac{d}{dt}\theta_k(t) q_k$$

(Refer Slide Time: 22:12)



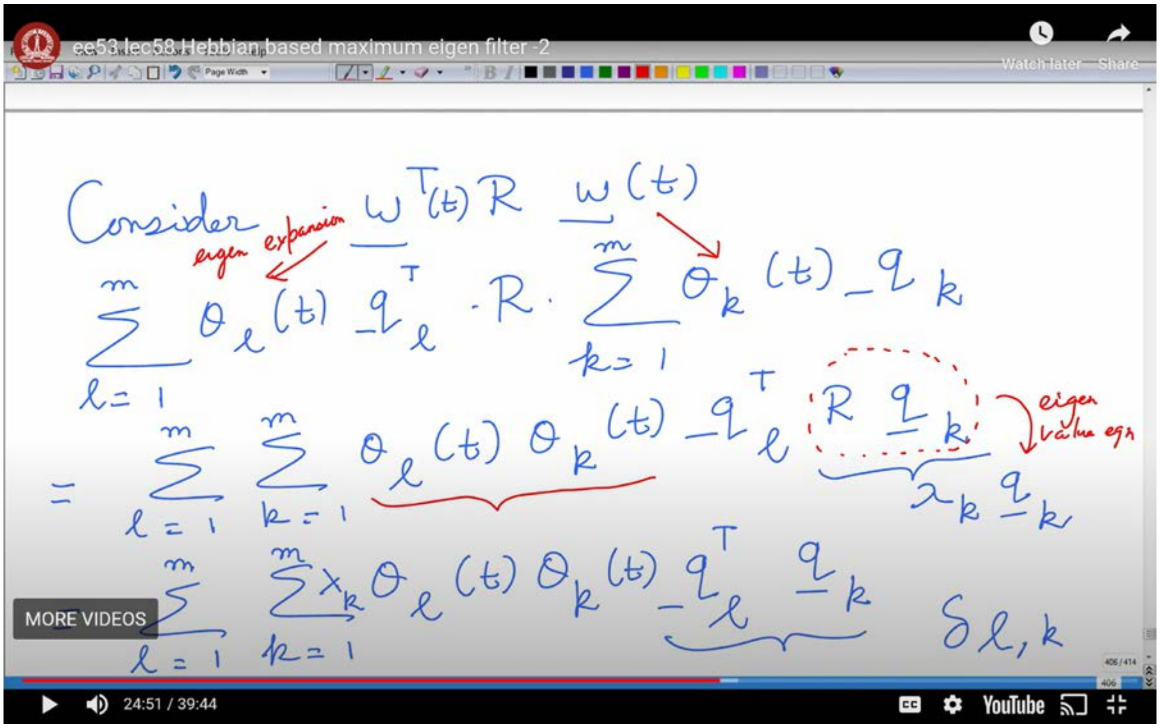
For the right side, $\bar{h}(W)$, we need to consider both terms:

$$\bar{h}\big(W(t)\big) = RW(t) - \big(W^T(t)RW(t)\big)W(t)$$

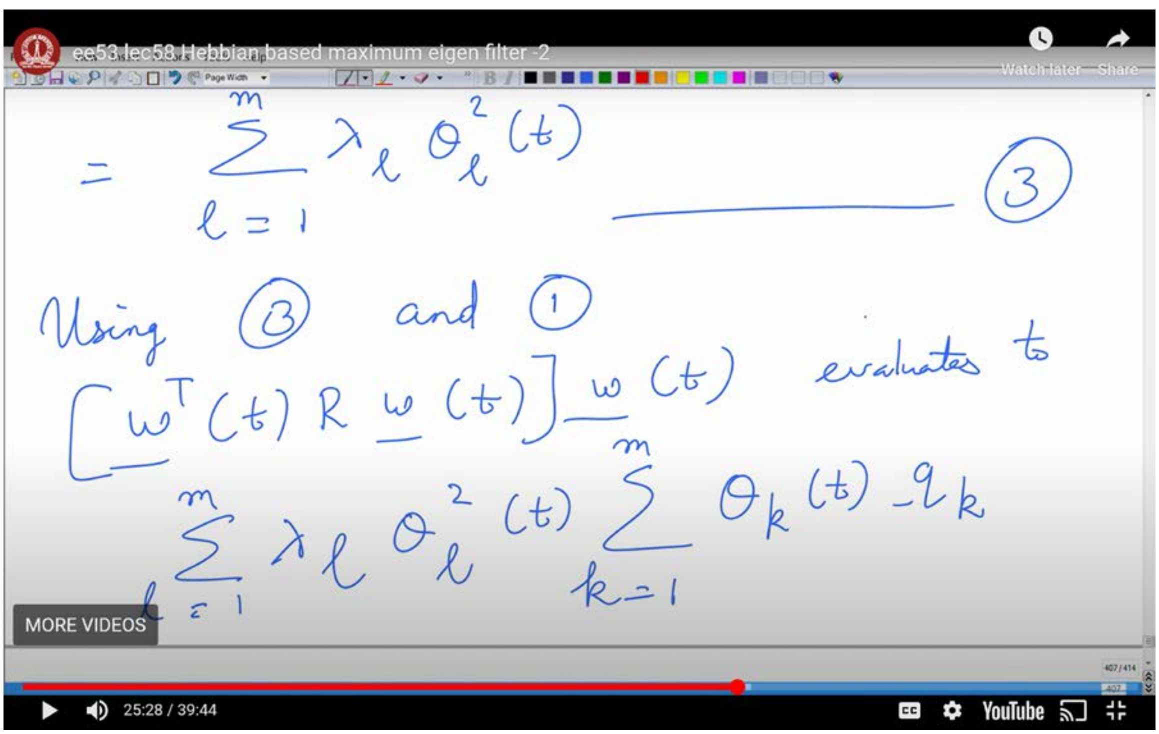
Substitute the eigen expansion for W(t):

$$W(t) = \sum_{l=1}^{m} \theta_l(t) q_l$$

(Refer Slide Time: 24:51)



(Refer Slide Time: 25:28)



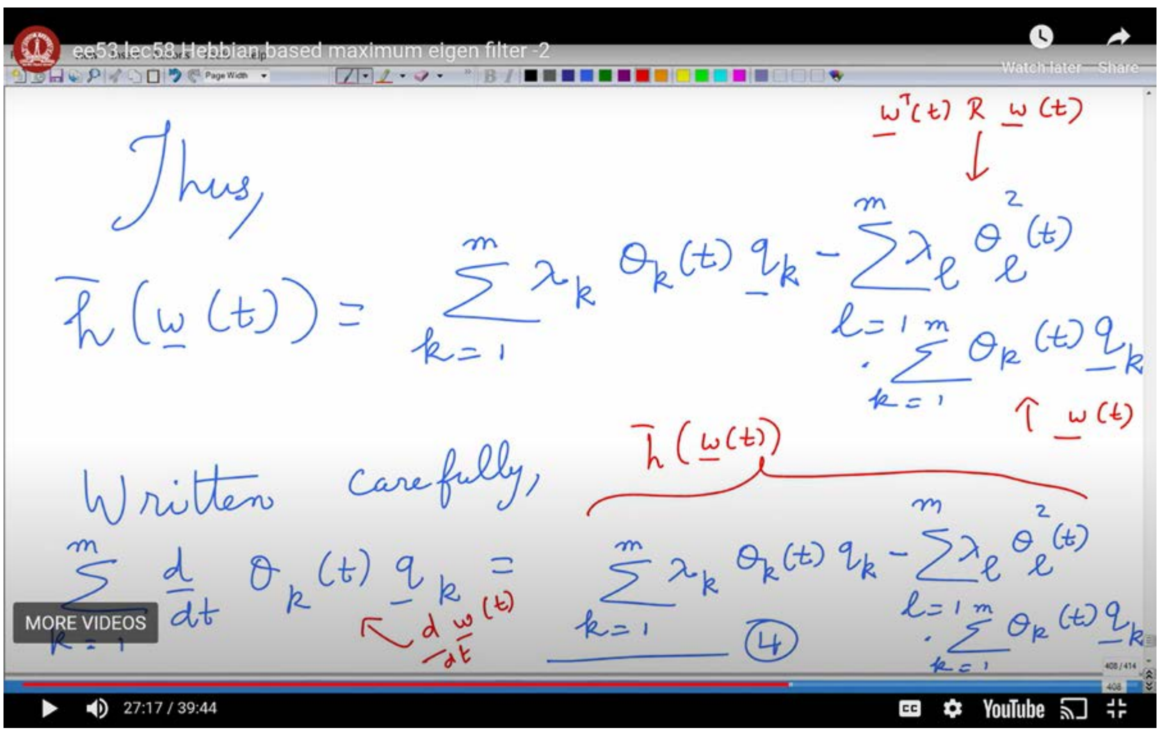
Expanding $\bar{h}(W)$, we have:

$$\bar{h}(W(t)) = R\left(\sum_{l=1}^{m} \theta_l(t) q_l\right) - \left(\left(\sum_{l=1}^{m} \theta_l(t) q_l\right)^T R\left(\sum_{k=1}^{m} \theta_k(t) q_k\right)\right)\left(\sum_{k=1}^{m} \theta_k(t) q_k\right)$$

Now, let's simplify the terms. For the first term, we get:

$$R\left(\sum_{l=1}^{m} \theta_l(t) q_l\right) = \sum_{l=1}^{m} \theta_l(t) R\, q_l = \sum_{l=1}^{m} \theta_l(t) \lambda_l q_l$$

(Refer Slide Time: 27:17)



For the second term, the inner product simplifies to:

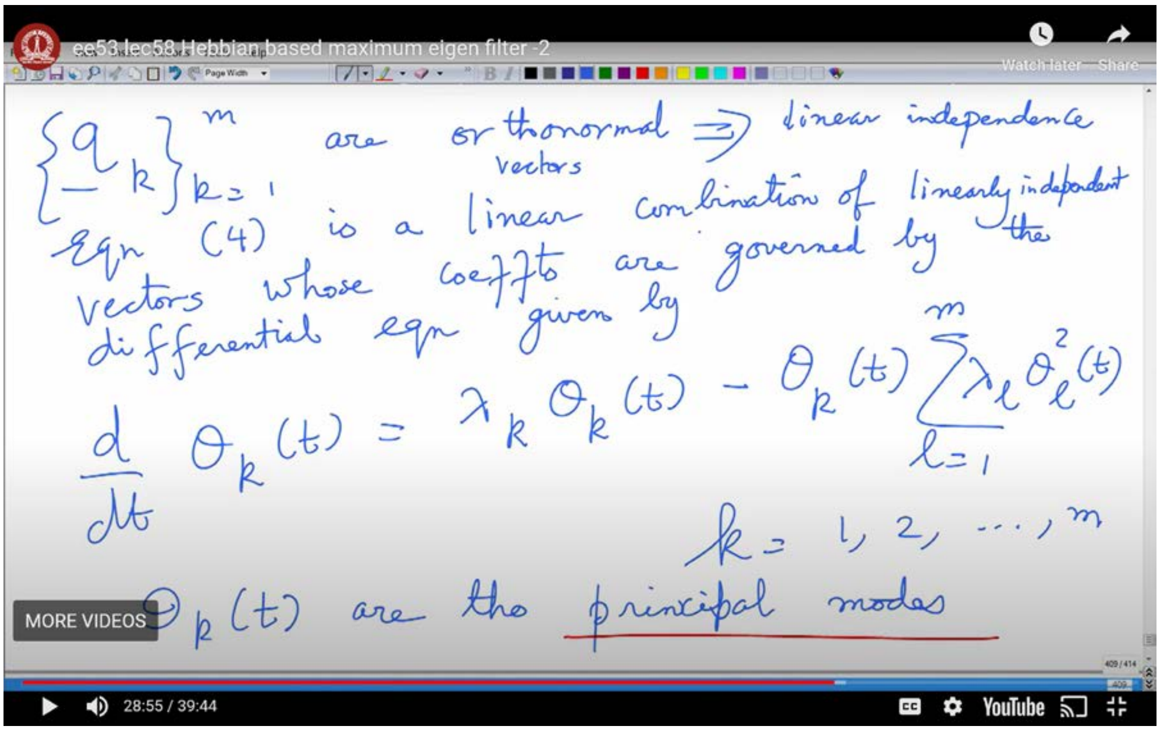
$$\left(\sum_{l=1}^{m} \theta_l(t) q_l\right)^T R\left(\sum_{k=1}^{m} \theta_k(t) q_k\right) = \sum_{l=1}^{m}\sum_{k=1}^{m} \theta_l(t)\theta_k(t) q_l^T R\, q_k$$

$$= \sum_{l=1}^{m}\sum_{k=1}^{m} \theta_l(t)\theta_k(t)\lambda_k q_l^T q_k$$

Because the eigenvectors are orthonormal, $q_l^T q_k$ equals 1 if l = k and 0 otherwise. Therefore:

$$\sum_{l=1}^{m}\sum_{k=1}^{m}\theta_l(t)\theta_k(t)\lambda_k q_l^T q_k = \sum_{k=1}^{m}\theta_k(t)^2\lambda_k$$

(Refer Slide Time: 28:55)



So, the expression simplifies to:

$$\left(\sum_{k=1}^{m}\theta_k(t)^2\lambda_k\right)\left(\sum_{k=1}^{m}\theta_k(t)q_k\right)$$

Thus, our differential equation becomes:

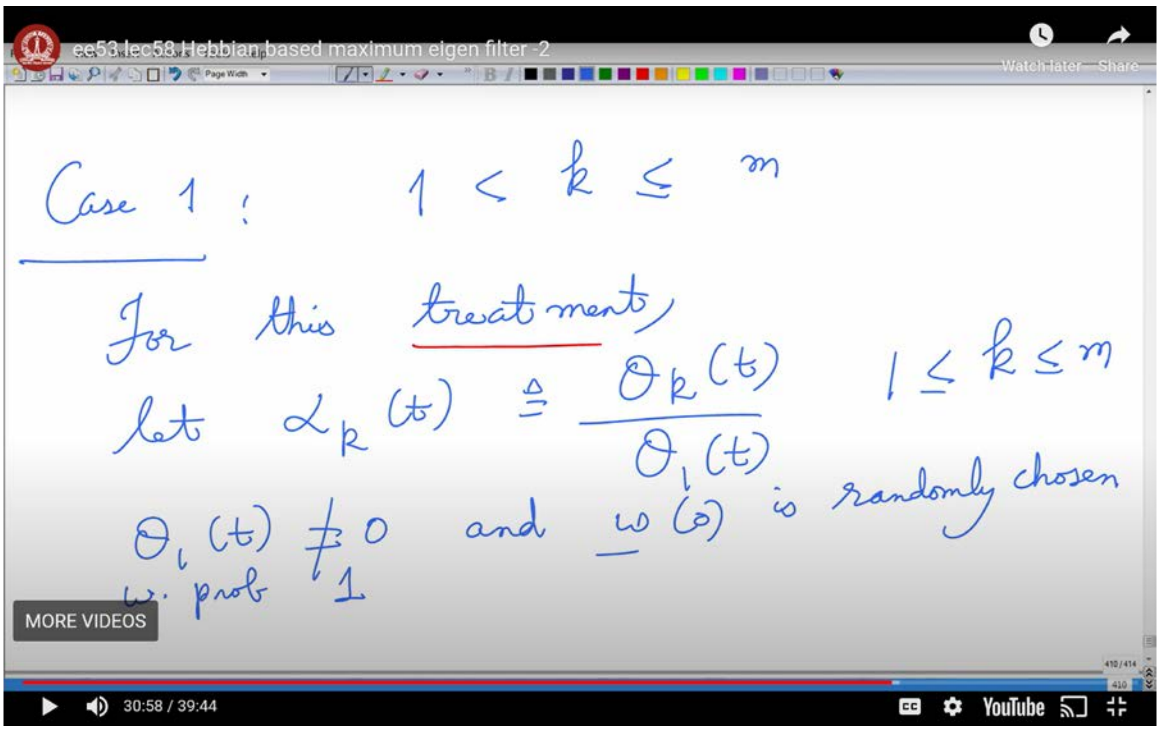
$$\frac{d}{dt}W(t) = \sum_{k=1}^{m}\frac{d}{dt}\theta_k(t)q_k$$

And for the function $\bar{h}(W)$, we account for the contributions from the eigenvalues and eigenvectors, allowing us to understand the dynamics of the system more clearly.

We can simplify the expression $Q_L^T Q_k$ using the Kronecker delta function Δ_{Lk}. This function equals 1 when L is equal to k and equals 0 otherwise. This simplification allows us to reduce the double summation to a single summation, as the terms where $L \neq k$ vanish. Consequently, the expression simplifies to:

$$\sum_{L=1}^{m} \lambda_L \theta_L^2(t)$$

(Refer Slide Time: 30:58)



Let's refer to this simplified expression as Equation 3. With this result and using our expansion for W(t) in terms of the eigenbasis, we can compactly simplify the function $\bar{h}(W(t))$. As shown in the expression above, $\bar{h}(W(t))$ can be written as:

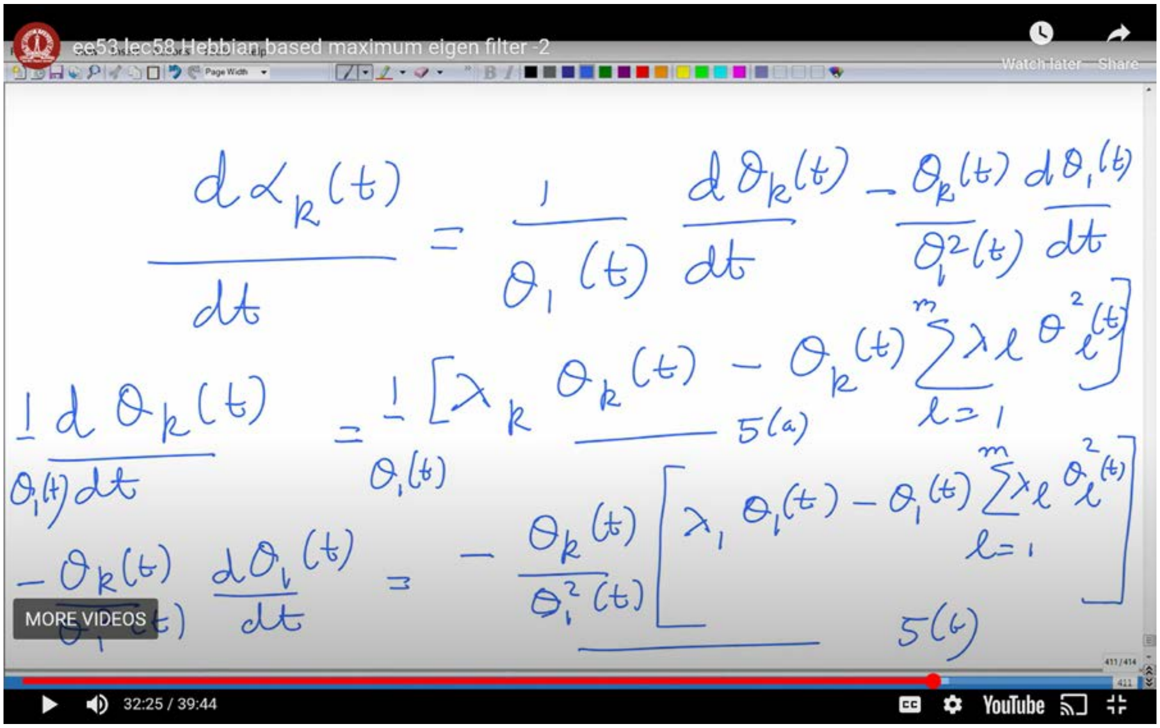
$$\sum_{k=1}^{m} \lambda_k \theta_k q_k - \left(\sum_{l=1}^{m} \sum_{k=1}^{m} \theta_l \theta_k q_l^T R\, q_k - 2 \text{ terms} \right)$$

Here, the first term represents W(t) and the second term corresponds to $W^T(t)$ R W(t). With this, we define Equation 4:

$$\frac{d}{dt}W(t) = \bar{h}\big(W(t)\big)$$

where the left side of Equation 4 is what we seek, namely $\frac{d}{dt}W(t)$, and the right side is the simplified form of $\bar{h}\big(W(t)\big)$.

(Refer Slide Time: 32:25)



The eigenvectors q_k for k = 1, …, m are orthonormal, which implies they are also linearly independent. Hence, Equation 4 represents a linear combination of these linearly independent vectors. This indicates that the coefficients of these eigenvectors in Equation 4 must satisfy a differential equation.

By equating the coefficients of these eigenvectors, we find that each coefficient must be zero. Thus, the differential equation simplifies to:

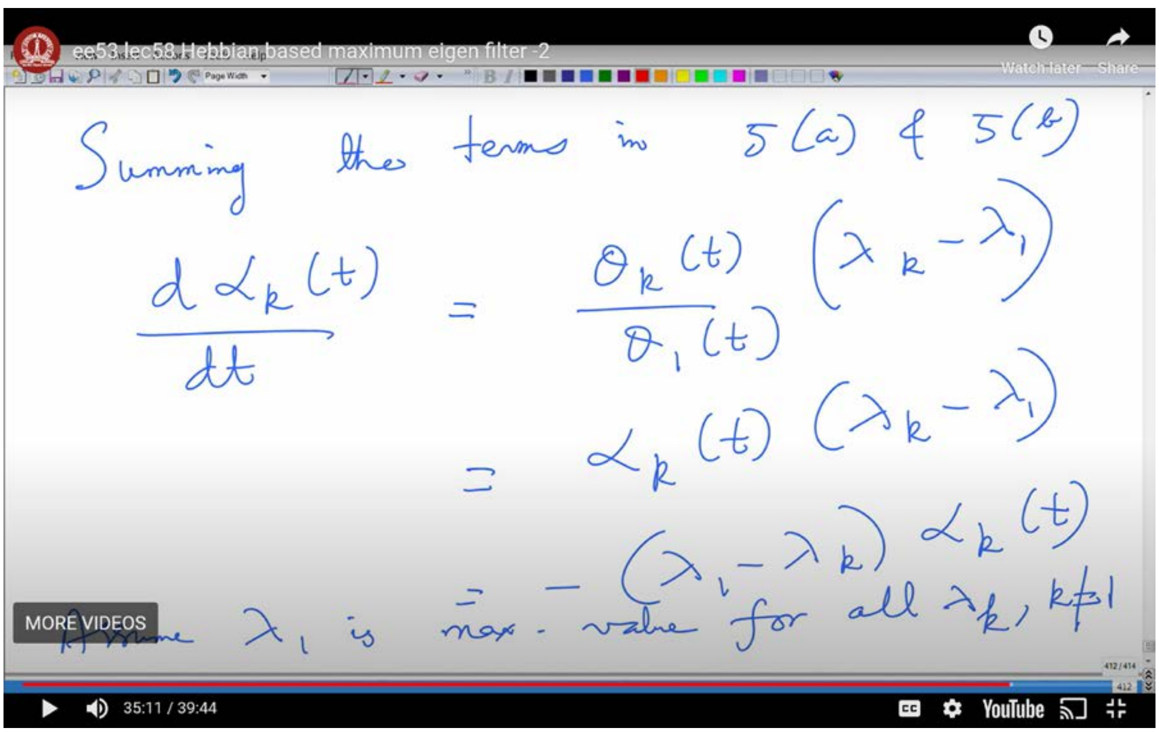
$$\frac{d}{dt}\theta_k(t) = \lambda_k \theta_k(t) - \theta_k(t) \sum_{l=1}^{m} \lambda_l \theta_l^2(t)$$

This equation holds for k = 1, 2, …, m, and $\theta_k(t)$ represents the principal modes of the non-linear dynamical system. Our goal is to solve for these principal modes. To tackle this, we break the problem into two cases:

1. Case 1: $1 < k \leq m$

2. Case 2: k = 1

(Refer Slide Time: 35:11)



For Case 1, define $\alpha_k(t)$ as:

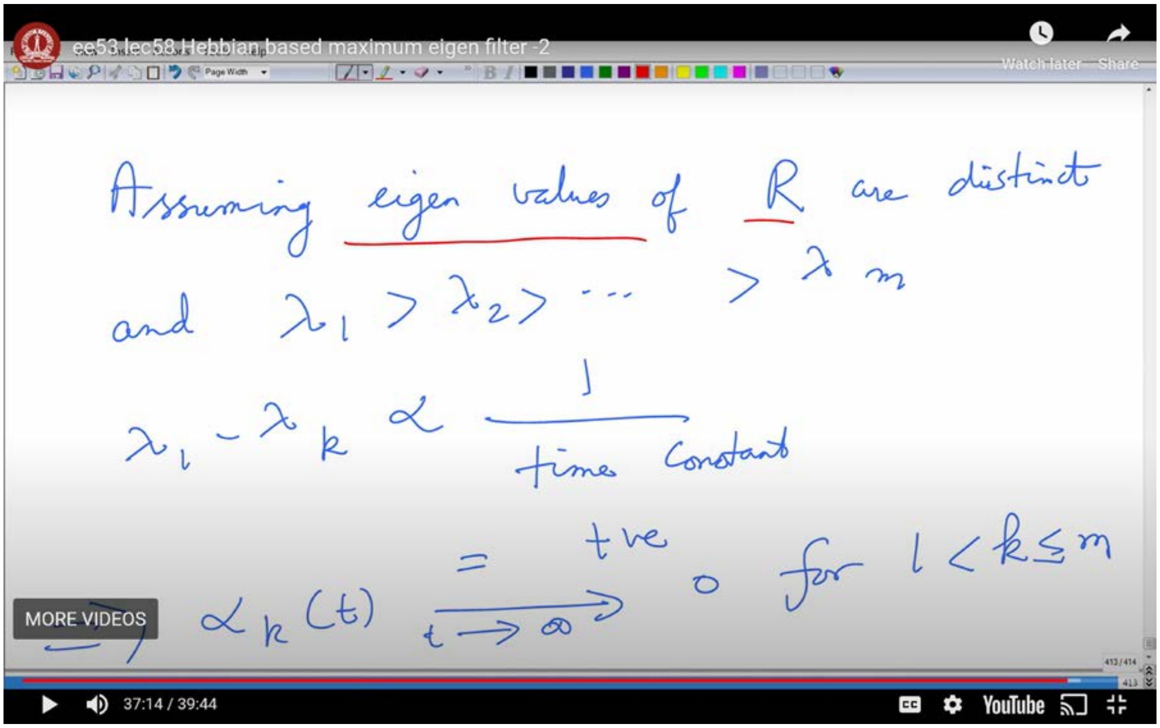
$$\alpha_k(t) = \frac{\theta_k(t)}{\theta_1(t)}$$

where $1 \leq k \leq m$. In this case, $\alpha_k(t)$ is a function of time and holds for all $k \neq 1$, with $\theta_1(t) \neq 0$ with probability 1. The weight at time step 0 is chosen randomly from a distribution with small random values.

Now, let's examine our governing equation for $\alpha_k(t)$. Here, $\alpha_k(t)$ is defined as $\frac{\theta_k(t)}{\theta_1(t)}$. To find the derivative of $\alpha_k(t)$ with respect to time, we use the chain rule. Specifically, we need to differentiate $\frac{\theta_k(t)}{\theta_1(t)}$, which can be expressed as:

$$\frac{d\alpha_k(t)}{dt} = \frac{d}{dt}\left(\frac{\theta_k(t)}{\theta_1(t)}\right)$$

(Refer Slide Time: 37:14)



Applying the quotient rule, this derivative becomes:

$$\frac{d\alpha_k(t)}{dt} = \frac{\theta_1(t)\frac{d\theta_k(t)}{dt} - \theta_k(t)\frac{d\theta_1(t)}{dt}}{\theta_1^2(t)}$$

Let's break this down into two terms for clarity. The first term in the numerator is $\frac{1}{\theta_1(t)}\frac{d\theta_k(t)}{dt}$. From the differential equation we previously derived, we know:

$$\frac{d\theta_k(t)}{dt} = \lambda_k\,\theta_k(t) - \theta_k(t)\sum_{l=1}^{m}\lambda_l\,\theta_l^2(t)$$

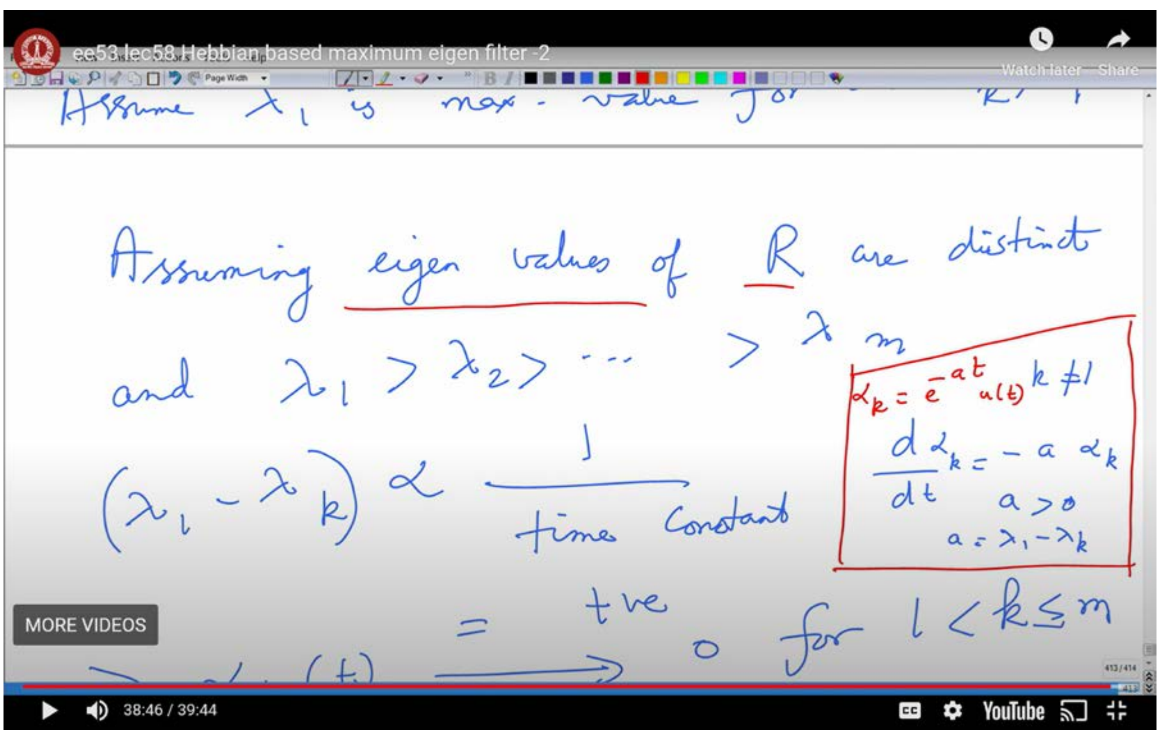
Substituting this into our first term, we get:

$$\frac{1}{\theta_1(t)}\left(\lambda_k\theta_k(t) - \theta_k(t)\sum_{l=1}^{m}\lambda_l\theta_l^2(t)\right)$$

For the second term in the numerator, we pull out $\frac{\theta_k(t)}{\theta_1^2(t)}$ and focus on $\frac{d\theta_1(t)}{dt}$. To find $\frac{d\theta_1(t)}{dt}$, we substitute k = 1 into our differential equation. This gives us:

$$\frac{d\theta_1(t)}{dt} = \lambda_1\theta_1(t) - \theta_1(t)\sum_{l=1}^{m}\lambda_l\theta_l^2(t)$$

(Refer Slide Time: 38:46)



Putting this into the second term, we get:

$$\frac{\theta_k(t)}{\theta_1^2(t)}\left(\lambda_1\theta_1(t) - \theta_1(t)\sum_{l=1}^{m}\lambda_l\theta_l^2(t)\right)$$

Combining these terms, we simplify the expression. The summation terms involving $\theta_k(t)$ and $\theta_1(t)$ cancel each other out. Therefore, the simplified expression for $\frac{d\alpha_k(t)}{dt}$ is:

$$\frac{d\alpha_k(t)}{dt} = \alpha_k(t)(\lambda_k - \lambda_1)$$

The interpretation of this result is as follows: We have evaluated the time derivative of the ratio of the components $\theta_k(t)$ and $\theta_1(t)$. This simplification helps us assess how the relative strength of each component changes with respect to the first component. The key takeaway here is that $\alpha_k(t)$ evolves according to the difference in eigenvalues λ_k and λ_1, which provides insight into the dynamics of the system's principal modes.

This means that I am examining the function θ as a function of time t for the first eigencomponent, which is denoted by q_1. My goal is to assess how this component scales relative to any other component k where $k \neq 1$. Specifically, I want to determine whether this ratio trends toward zero or remains non-zero. This is the essence of the study and the reason behind the analysis.

Let α_k be defined as $\frac{\theta_k}{\theta_1}$. We have derived that:

$$\frac{d\alpha_k(t)}{dt} = (\lambda_k - \lambda_1)\alpha_k(t)$$

Since λ_1 is the largest eigenvalue, and $\lambda_1 > \lambda_2 > \lambda_3$, it is useful to express $\lambda_k - \lambda_1$ as $-(\lambda_1 - \lambda_k)$. This reformulation makes it clear that $\lambda_1 - \lambda_k$ is a positive quantity, meaning $\alpha_k(t)$ evolves as:

$$\frac{d\alpha_k(t)}{dt} = -(\lambda_1 - \lambda_k)\alpha_k(t)$$

Given that the eigenvalues of the correlation matrix are distinct, the term $\lambda_1 - \lambda_k$ is positive. This term is proportional to $\frac{1}{\text{time constant}}$, meaning we can write:

$$\frac{d\alpha_k(t)}{dt} = -a\alpha_k(t)$$

where a is a positive constant equal to $\lambda_1 - \lambda_k$.

This results in a simple linear differential equation with the solution:

$$\alpha_k(t) = \alpha_k(0)e^{-at}$$

where $\alpha_k(0)$ is the initial condition. As time t approaches infinity, $\alpha_k(t)$ decays to zero. This implies that for all principal modes $k \neq 1$, $\alpha_k(t)$ will approach zero. Therefore, the only mode that requires further investigation is the mode where k = 1.

We will pause here and continue our analysis for the case when k = 1 in the next lecture. This case is more complex and will involve Lyapunov analysis.