Neural Networks for Signal Processing-I

Prof. Shayan Srinivasa Garani

Department of Electronic System Engineering

Indian Institute of Science, Bengaluru

Lecture – 45
Generalized Applicability of the Representer Theorem

In our last discussion, we explored how any function in a reproducing kernel Hilbert space can be expressed as a linear combination of Mercer kernel functions. This key insight stems from the representer theorem. Now, a pertinent question arises: is this result useful within the regularization framework of a regression problem? Specifically, when we aim to optimize a standard error in a regression problem, subject to certain regulatory conditions that obey monotonicity properties, can the representation provided by the representer theorem serve as a solution to this optimization problem? In other words, can we minimize our standard error in the regression problem while adhering to these regulatory conditions? Let's delve into this and examine it as a practical application when dealing with data.

I'll state the theorem first, and then we'll go through the proof in detail. As discussed earlier, f(x_j) can be expressed as a linear combination of Mercer kernel functions, meaning that this function in the reproducing kernel Hilbert space can be written as a sum:

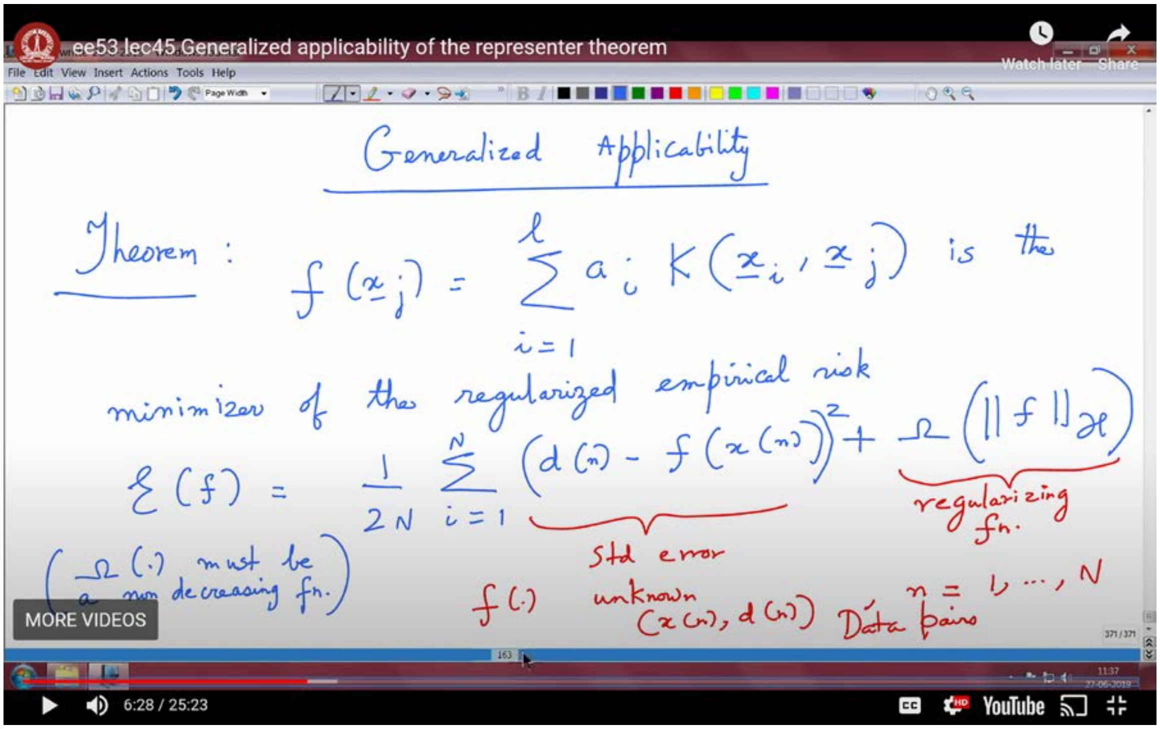
$$f(x_j) = \sum_{i=1}^{l} a_i \; k(x_i, x_j)$$

This is our starting point, and according to the representer theorem, this representation minimizes the regularized empirical risk, which is given by E(f).

The empirical risk E(f) is expressed as:

$$E(f) = \frac{1}{2N} \sum_{i=1}^{N} \left(d_i - f(x_i)\right)^2$$

(Refer Slide Time: 06:28)


ee53 lec45 Generalized applicability of the representer theorem
Watch later
Share
File Edit View Insert Actions Tools Help
MORE VIDEOS
6:28 / 25:23
YouTube

Generalized Applicability

Theorem : $f(\underline{x}_j) = \sum_{i=1}^{l} a_i K(\underline{x}_i, \underline{x}_j)$ is the minimizer of the regularized empirical risk

$$\mathcal{E}(f) = \frac{1}{2N} \sum_{i=1}^{N} \underbrace{\left(d(n) - f(x(n))\right)^2}_{\text{Std error}} + \underbrace{\Omega\left(\|f\|_{\mathcal{H}}\right)}_{\text{regularizing fn.}}$$

($\Omega(\cdot)$ must be a non decreasing fn.)

$f(\cdot)$ unknown

$(x(n), d(n))$ Data pairs $n = 1, \dots, N$

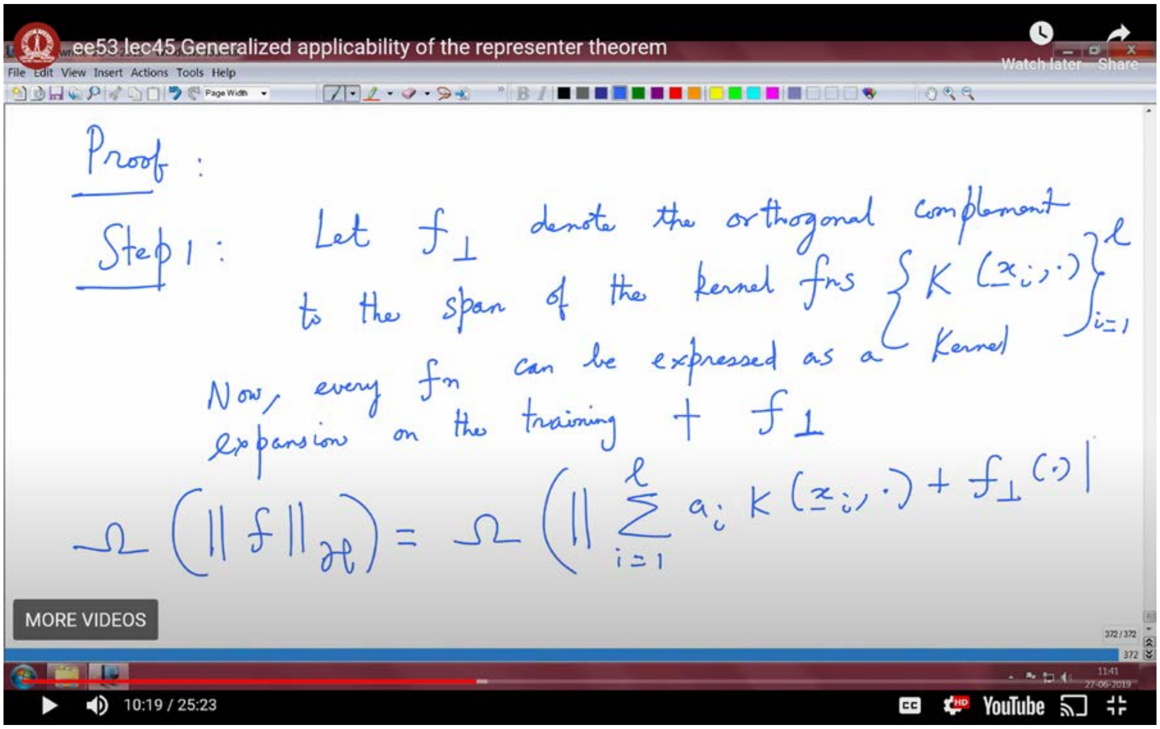
Here, I take the squared error, where the error is essentially the deviation between my desired response, d_i, and the unknown function f that needs to be estimated. The desired response is a function of the data, and this is how I formulate my error criterion. This E(f) is my standard error, where f is unknown, and (x_n, d_n) are my data pairs. In addition to the error term, I also impose a regularity condition given by Ω(|f|), which is a norm of the function f in this Hilbert space. This norm could be an integral norm, among others.

An important detail to note is that the index n runs from 1 to N, where N represents my training samples. The function Ω serves as my regularizing function, and it is crucial that this regularizing function Ω satisfies certain properties, specifically, it must be a non-decreasing function.

With these conditions in place, we begin the proof of this result. Let's proceed step by step.

Step 1 of the proof proceeds as follows: Let $f_\perp$ denote the orthogonal complement to the span of the kernel functions. That is, consider the set of kernel functions indexed from i = 1 to L, and define $f_\perp$ as the function orthogonal to this span. This implies that the inner product of $f_\perp$ with any kernel function k will be zero. Essentially, any function can be represented as a kernel expansion over the training data. This kernel expansion is computed at specific data points, and it also includes the component $f_\perp$.

(Refer Slide Time: 10:19)


ee53 lec45 Generalized applicability of the representer theorem
File Edit View Insert Actions Tools Help
Watch later
Share
MORE VIDEOS
10:19 / 25:23

Proof :

Step 1 : Let $f_\perp$ denote the orthogonal complement to the span of the kernel fns $\{K(x_i,\cdot)\}_{i=1}^{l}$

Now, every fn can be expressed as a Kernel expansion on the training + $f_\perp$

$$\Omega(\|f\|_{\mathcal{H}}) = \Omega\left(\| \sum_{i=1}^{l} a_i K(x_i,\cdot) + f_\perp(\cdot) \|\right.$$

In the representer theorem, we decomposed every function into a part that lies within the span of the kernel functions and another component that is orthogonal to these kernel functions. Here, the norm Ω on this function in the Hilbert space plays a critical role. By constructing Ω as a regularizing function, which operates on the norm of the function within this reproducing kernel Hilbert space, we introduce regulatory conditions. These conditions are enforced by applying the norm within the Hilbert space. We can, therefore, express f as a combination of $f_\perp$ and the kernel expansion.

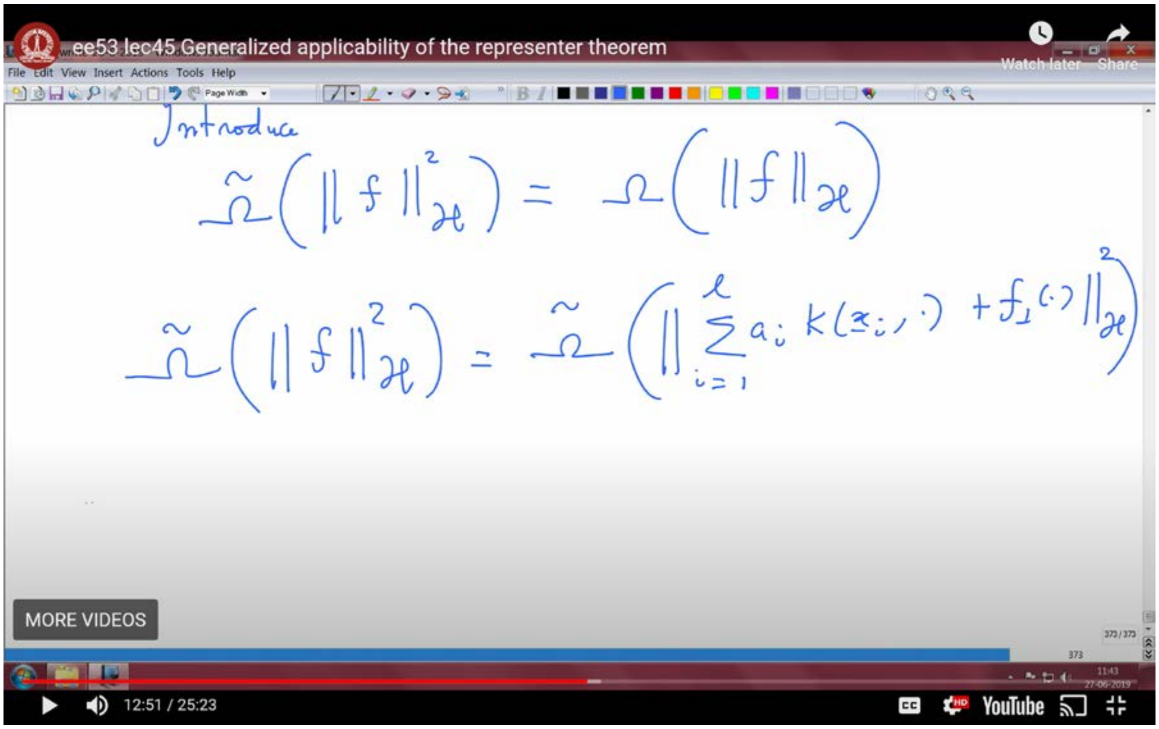
Now, let’s expand this function:

$$\Omega(|f|) = \Omega\left(\sum_{i=1}^{L} a_i k(x_i,\cdot) + f_\perp\right)$$

Here, Ω acts on the norm of the function, which has been decomposed into the sum of a kernel expansion and the orthogonal component $f_\perp$. Next, we introduce another function, $\tilde{\Omega}$, where:

$$\tilde{\Omega}(|f|^2) = \Omega(|f|)$$

(Refer Slide Time: 12:51)



ee53 lec45 Generalized applicability of the representer theorem
Watch later
Share
File Edit View Insert Actions Tools Help
Page Width
MORE VIDEOS
12:51 / 25:23
YouTube

Introduce

$$\tilde{\Omega}\left(\|f\|_{\mathcal{H}}^{2}\right) = \Omega\left(\|f\|_{\mathcal{H}}\right)$$

$$\tilde{\Omega}\left(\|f\|_{\mathcal{H}}^{2}\right) = \tilde{\Omega}\left(\Big\|\sum_{i=1}^{l} a_i k(x_i,\cdot) + f_\perp(\cdot)\Big\|_{\mathcal{H}}^{2}\right)$$

This is merely a redefinition, where $\tilde{\Omega}$ operates on the squared norm over this space. Therefore, we can express $\tilde{\Omega}(|f|^2)$ as:

$$\tilde{\Omega}\left(\sum_{i=1}^{L} a_i k(x_i,\cdot) + f_\perp\right)$$

This leads us to express the norm as follows:

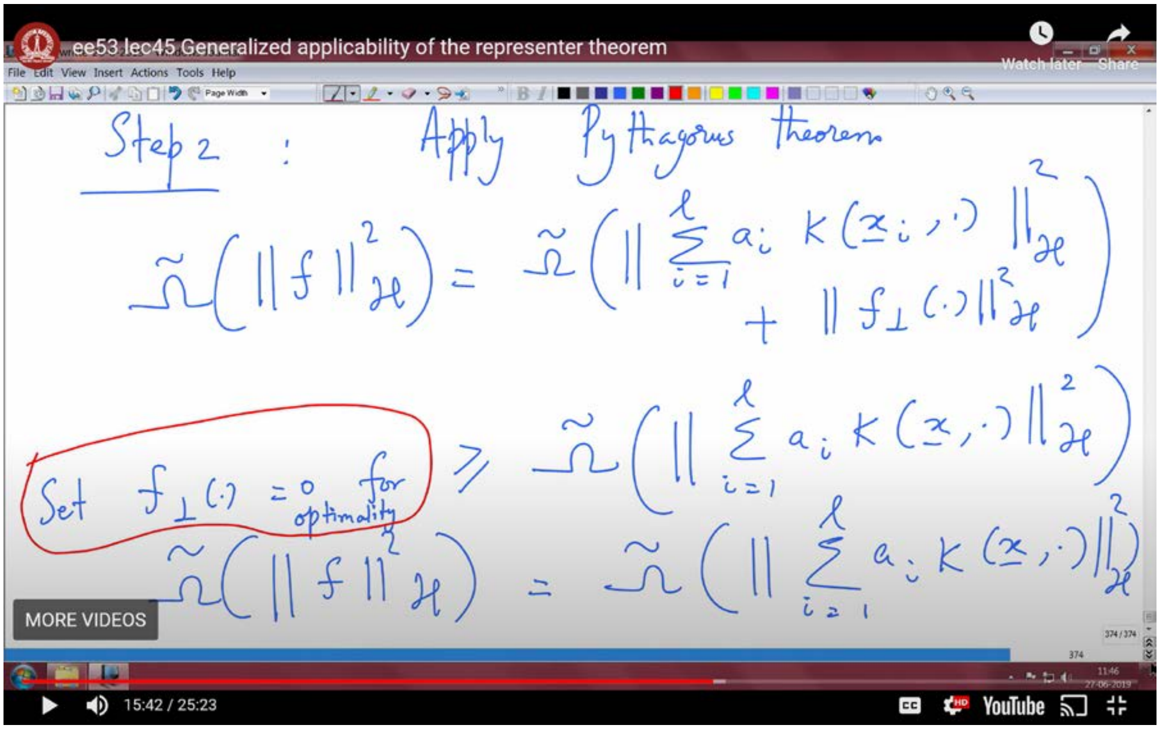
$$\tilde{\Omega}(|f|^2) = \tilde{\Omega}\left(\sum_{i=1}^{L} a_i k(x_i,\cdot) + f_\perp\right)^2$$

Moving to Step 2: Here, we invoke the Pythagorean theorem property. Since we are dealing with the norm of a sum of two functions, we can express this norm as the sum of the squared norms of the individual functions. Given that $f_\perp$ is orthogonal to the kernel expansion, we can simplify the expression.

Thus, we have:

$$\widetilde{\Omega}(|f|^2) = \widetilde{\Omega}\left(\left(\sum_{i=1}^{L} a_i k(x_i,\cdot)\right)^2 + |f_\perp|^2\right)$$

(Refer Slide Time: 15:42)


ee53 lec45 Generalized applicability of the representer theorem
File Edit View Insert Actions Tools Help
Page Width
Watch later
Share
MORE VIDEOS
374 / 374
15:42 / 25:23

Step 2 : Apply Pythagorus theorem

$$\tilde{\Omega}\left(\|f\|^2_{\mathcal{H}}\right) = \tilde{\Omega}\left(\left\|\sum_{i=1}^{l} a_i\, k(x_i,\cdot)\right\|^2_{\mathcal{H}} + \|f_\perp(\cdot)\|^2_{\mathcal{H}}\right)$$

$$\geq \tilde{\Omega}\left(\left\|\sum_{i=1}^{l} a_i\, k(x,\cdot)\right\|^2_{\mathcal{H}}\right)$$

Set $f_\perp(\cdot) = 0$ for optimality

$$\tilde{\Omega}\left(\|f\|^2_{\mathcal{H}}\right) = \tilde{\Omega}\left(\left\|\sum_{i=1}^{l} a_i\, k(x,\cdot)\right\|^2_{\mathcal{H}}\right)$$

This equality holds strictly because the projection of $f_\perp$ onto the kernel functions is zero. As a result, we find that:

$$\widetilde{\Omega}(|f|^2) \geq \widetilde{\Omega}\left(\left(\sum_{i=1}^{L} a_i k(x_i,\cdot)\right)^2\right)$$

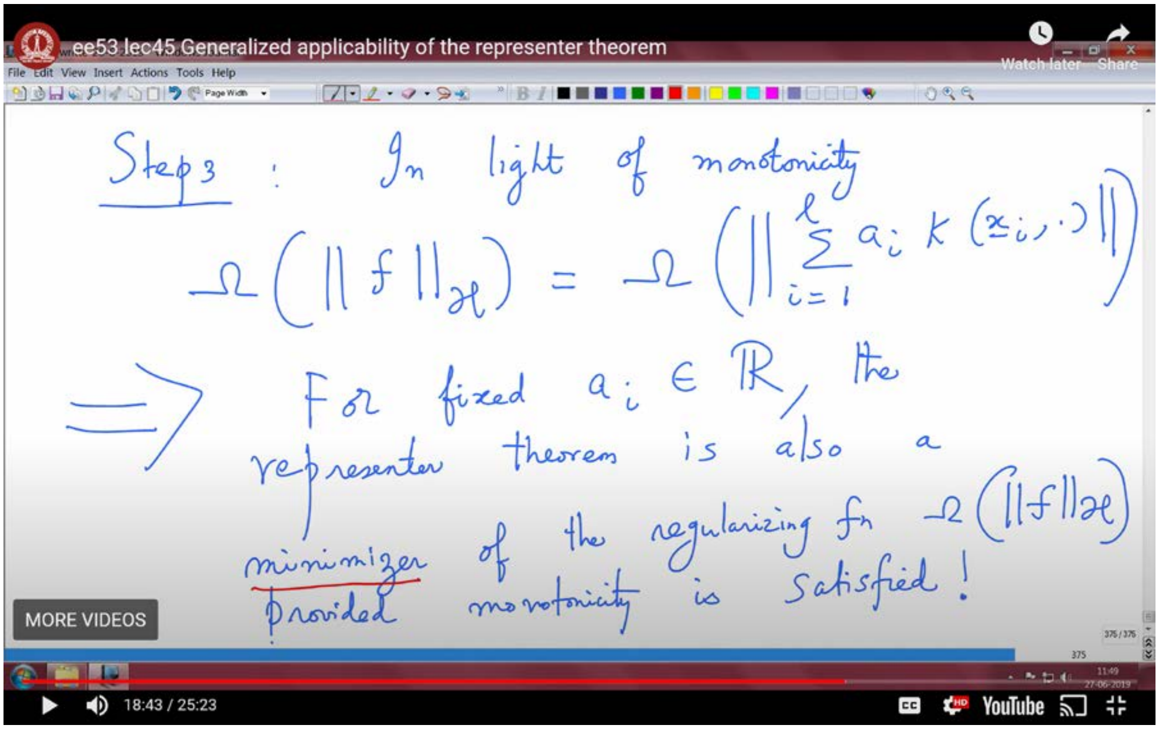
Here, the inequality holds with equality when we set $f_\perp$ to zero for optimality. When $f_\perp = 0$, the expression simplifies further:

$$\widetilde{\Omega}(|f|^2) = \widetilde{\Omega}\left(\left(\sum_{i=1}^{L} a_i k(x_i,\cdot)\right)^2\right)$$

Finally, let's explore how the monotonicity property of the regularizing function Ω can further assist us in optimizing this expression.

In light of the monotonicity property, the expression $\Omega(|f|)$ simplifies to $\Omega(|\sum_{i=1}^{L} a_i k(x_i,\cdot)|)$. This is because earlier, we defined $\widetilde{\Omega}(|f|^2)$ as $\Omega(|f|)$. Now, having expressed everything in terms of $\widetilde{\Omega}(|f|^2)$ up until our application of the Pythagorean theorem, we simply substitute it back under the monotonicity condition. Consequently, $\Omega(|f|)$ in the Hilbert space becomes $\Omega(|\sum_{i=1}^{L} a_i k(x_i,\cdot)|)$.

(Refer Slide Time: 18:43)


ee53 lec45 Generalized applicability of the representer theorem
Watch later
Share
File Edit View Insert Actions Tools Help
Page Width
MORE VIDEOS
375 / 375
375
11:49
27-06-2019
18:43 / 25:23

Step 3 : In light of monotonicity

$$\Omega(\|f\|_{\mathcal{H}}) = \Omega\left(\left\|\sum_{i=1}^{\ell} a_i\, k(x_i,\cdot)\right\|\right)$$

$\Longrightarrow$ For fixed $a_i \in \mathbb{R}$, the representer theorem is also a minimizer of the regularizing fn $\Omega(\|f\|_{\mathcal{H}})$ provided monotonicity is satisfied !

What this implies is that, for fixed coefficients a_i belonging to R, the representer theorem also serves as a minimizer of the regularizing function $\Omega(|f|)$ in the Hilbert space, provided the monotonicity condition is met. This is a critical result as it connects our representative function with the objective of minimizing the standard error along with the regularizing term.

To recap, let's briefly walk through the steps of the theorem. We started with the representer theorem, which states that any function in a reproducing kernel Hilbert space can be expressed as a linear combination of Mercer kernels. This concept, as outlined in the representer theorem, also minimizes the regularized empirical risk, which is governed by this expression. Additionally, the regularizing function must adhere to monotonicity properties.

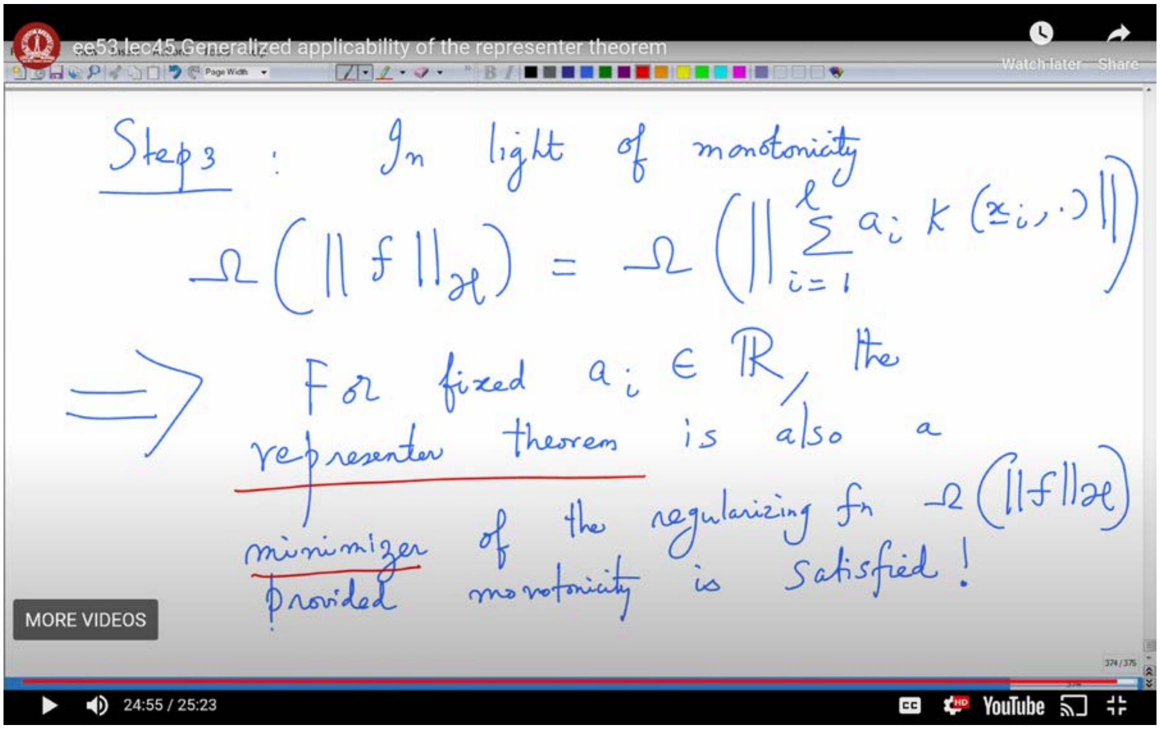
Given the data points (x_n, d_n), where n = 1 to N, our proof unfolds in three key steps. The first step, similar to the proof of the representer theorem, involves decomposing the function f into two components: one that lies within the span of the kernel functions and another that is orthogonal to them.

Next, we introduce the regularizing function $\Omega(|f|)$ operating in the Hilbert space and decompose f in terms of the kernel span and its orthogonal complement. We then define another function $\widetilde{\Omega}$ that acts on the squared norm.

Finally, the monotonicity property is invoked by defining $\widetilde{\Omega}(|f|^2)$ as $\Omega(|f|)$. We then decompose $\Omega(|f|)$ using the perpendicular and parallel components of f. The purpose of separating these components is to apply the Pythagorean theorem in the next step, which naturally fits because the projection of one component onto the other is zero. Thus, the norm of the sum of two functions, a + b, can be expressed as the sum of the norms squared: $|a + b|^2 = |a|^2 + |b|^2$. This is exactly what we did here, relying on the fact that the inner product of a and b is zero.

So, where a and b are our functions, we now substitute this into our equation. This gives us one term that depends on the span of the kernel functions, and the other term is essentially the perpendicular component. For optimality, we set $f_\perp$ (the perpendicular component) to zero. As a result, we are left with $\widetilde{\Omega}$ acting on the norm of the linear combination of these kernel functions, which is exactly what we need here. Given the monotonicity assumption we started with, we can state that $\Omega(|f|)$ is essentially Ω applied to the norm of a function expressed as a linear combination of the kernel functions.

(Refer Slide Time: 24:55)



Step 3 : In light of monotonicity

$$\Omega\left(\|f\|_{\mathcal{H}}\right) = \Omega\left(\left\|\sum_{i=1}^{\ell} a_i \, k(x_i, \cdot)\right\|\right)$$

$\Longrightarrow$ For fixed $a_i \in \mathbb{R}$, the representer theorem is also a minimizer of the regularizing fn $\Omega\left(\|f\|_{\mathcal{H}}\right)$ provided monotonicity is satisfied !

This means that the representer theorem also acts as a minimizer of the regularizing function, provided the monotonicity property is satisfied. A small detail was pointed out by one of my students: in the statement of this theorem, I mentioned that the representer theorem minimizes the regularized empirical risk. However, in the proof, we concluded that for a fixed set of constants a_i (where $a_i \in R$), the representer theorem minimizes the regularizing function. So, while this was proven, the question remains—how do we choose the a_i values that minimize the overall empirical risk?

To address this, you need to substitute the a_i values into the empirical risk expression. The a_i coefficients appear in both terms of the equation E(f). When you plug in $f(x_j)$, the a_i values influence both the standard error term and the regularizing term. You then need to optimize these variables a_i to minimize the overall empirical risk. This is a subtle but important detail when setting up this optimization framework.

So, while I proved that for fixed a_i values the representer theorem minimizes the regularizing function, taking it one step further requires optimizing a_i within both the standard error and regularizing terms to find the values that minimize the empirical risk. This completes the details. This is a very useful result because it connects the standard error with the regulatory conditions to the solution obtained from the representer theorem. I hope this theoretical insight can be effectively applied in practical scenarios.