Dynamic Data Assimilation
Prof. S Lakshmivarahan
School of Computer Science
Indian Institute of Technology, Madras

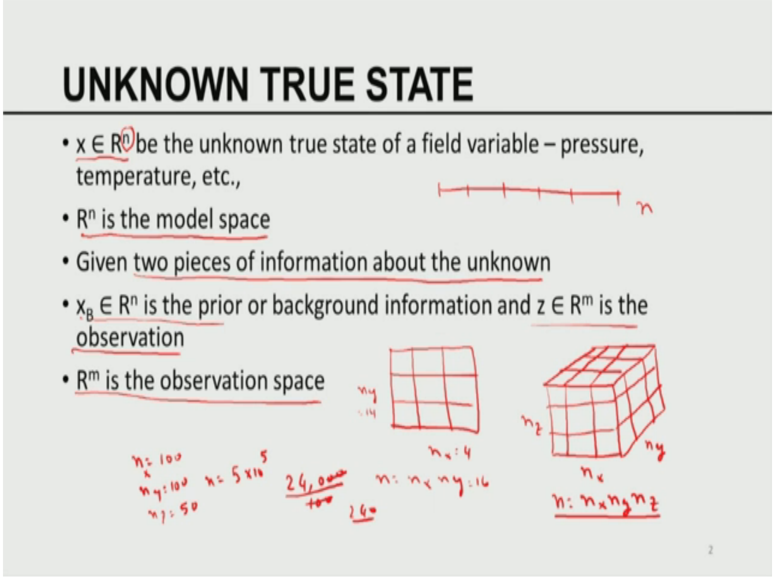
Lecture – 31
A Bayesian-3D-VAR methods

In this lecture, we are going to provide a broad overview of a class of techniques that has come to be called 3D VAR; VAR for variational, 3D essentially tells you we are concerned with problems in the 3 dimensional spatial variables at a given time. So, time is not a factor here. So, the 3D VAR and 4D VAR always represents variational methods these are occurring for variational method 3D essentially no time only space 4D , 3 plus 1, 3 for space 1 for time.

So, you can think of 3D VAR to be a data assimilation scheme done over the spatial domain at a given time 4D VAR is a data assimilation scheme that is done as a system evolves in time in the 3 dimensional space. Again, why are we talking about 3, 4 and so on? Most of the; why most, all the problems of interest in geophysical sciences happens on and around the earth. So, the space of interest is fundamentally a 3 dimensional space x, y, z and many of the physical processes evolve in time.

So, meteorologist in the early days when data assimilation schemes were developed in order to be able to bring out the distinction between when time is a factor, when time is not a factor they concocted this notion of 3 and 4, the way VAR essentially stems from the fact some of the earliest thought processes that led to the data assimilation scheme arose from the variational approach which we have indicated then we did forward sensitive method 4D VAR method or adjoin method and that is the door by which these acronyms crept into the vocabulary. So, you understand my thought is the rose smells the same by whatever name you call. So, you can if they want to call 3D VAR, so they.

(Refer Slide Time: 02:28)



UNKNOWN TRUE STATE

- $x \in R^n$ be the unknown true state of a field variable – pressure, temperature, etc.,
- R^n is the model space
- Given two pieces of information about the unknown
- $x_B \in R^n$ is the prior or background information and $z \in R^m$ is the observation
- R^m is the observation space

What is the idea? The idea is the underlying idea the 3D data assimilation is always I would like to be able to estimate the unknown true state of a field variable. So, I have a spatial domain. I have a grid embed in the spatial domain. I have numbered the grid 1 through n. At each of the grid points I have a pressure value or temperature value; at each grid value I am interested in a scalar variable. I am going to collate all the values and create a vector reaction. In the case of 1D the grids are along the 1 dimension, I would like to talk about for a minute.

So, in here the grid is like this. In case of 2D, you have the grid. So, in here we have n points along one line. In the case of 2D, I have the x dimension, n x points along the x, n y points along y, so, n is equal to n x times ny. In this example 1, 2, 3, 4 n x is equal to 4 ny is also equal to 4. So, n is equal to 16. In the case of a 3D, I have a grid. So, this is nx, this is ny, this is nz, therefore, n is equal to nx times ny times nz.

So, you can see as you go from 1 dimension to 2 dimension to 3 dimension the size explodes. Let me give a quick example; let us assume nx is equal to 100, ny is also equal to 100 and nz is equal to 50. So, n is equal to 1, 2, 3, 4, 5 times 10 to the power of 5 that is half a million. So, the whole question is this. So, what is the nx; the number of points along the x direction. So, let us now consider a spherical coordinate system.

So, I would like to be able to embed a spherical coordinate system built on the earth. The earth diameter is roughly 8000 miles. So, the circumference of the earth at the equator is roughly 24000 miles if you consider a grid of 100 points covering 24000 miles the distance between 2 grid points is 24000 divided by 100 that is a large distance. 24000 miles divided by 100, so, that is 240 miles. So, the distance between 2 grid points is 240 miles. Now, let us take the distance from Bangalore to Madras, for example; it is roughly 200 miles. So, there will be 1 grid point in Madras, 1 grid point in Bangalore, 1 grid point in probably Mangalore. Can that accurately capture the variations? No.

So, if you now change the number. So, if you want better accuracy you have to increase nx, you have to increase ny, you have to increase nz. What is nz? nz is the number of vertical levels generally in meteorology the vertical levels is limited to 50 or 100, that is all. But, what is the general problem for global models nx, ny, the product of nx ny with the nz, the number of points. So, that is why in atmospheric sciences in oceanographic sciences the value of n quickly in a hurry becomes a million, become tens of millions.

So, there are 2 things; one is how to solve conceptually a given problem. Secondly, knowing conceptually how to solve a given problem? How do we solve the problem for which we know that there is an algorithm, but a problem of huge size not every conceptual algorithm will be applicable directly to problems of large size. So, the course of dimensionality often times limits what one can do. We have already alluded to these things.

So, how do we decide what is the value of n? I want to do many things. Whatever I want to do I must be able to do in my lifetime. Earlier, we saw to multiply 2 matrices of million by million on a machine with a teraflop power it will take about 12 days to multiply one matrix, to multiply one pair of matrices. Can we afford to solve a problem for 12 days? Can we afford to solve a problem for one day? Can we afford to solve a problem in 6 hours? No, we want to be able to make forecasts every 6 hours, every 12 hours, every 24 hours. So, what is the time period within which you have to generate forecasts that provides the time horizon? The computer provides you the computing power.

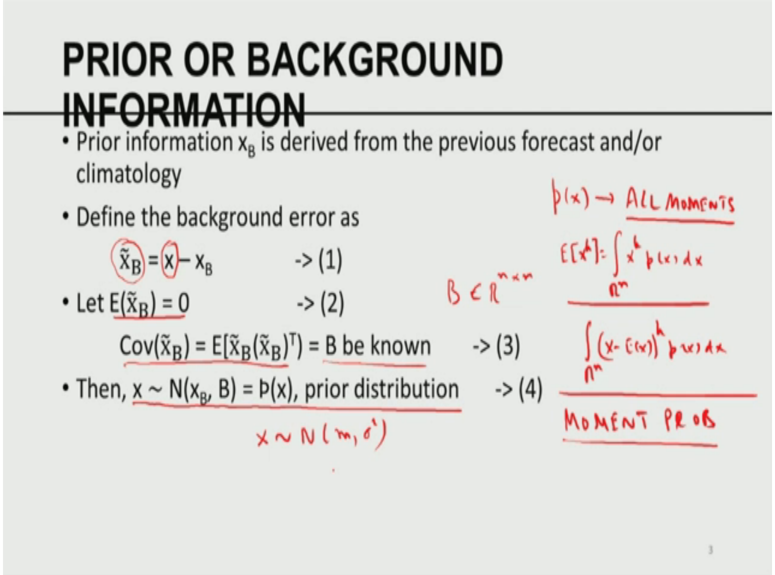
So, the time horizon and the computing power together limit the value of n that you can consider in solving problems. So, whenever we talk about problems of size n, I want you to keep track of this n. Mathematically, I can make this n as large as possible, but physically to be able to put it through the computer, rather notation. So, let us pretend I have a global domain, I have a field variable of interest, I have collated all the field variables of interest in a vector, vector is of size n. n, if it is the 1 dimensional problem, 2 dimensional problem, 3 dimensional problem, how it relates to 1D, 2D, 3D. That is what we have talked about so far.

So, R n is called the model space as we have been talking about all through. There are our notation is pretty common through the whole thing. There are 2 pieces of information about the unknown; one is the background or a prior and z the new observation. Again, this is nothing new; we have talked about it within the base in context at great length.

Let me give an example. Suppose, you are planning a trip to Paris, in the middle of January; you want to be able to pack the right kind of dress, so, what is it that you do? You would like to look at the almanacs that talk about the average low temperature average high temperature in the middle of January in downtown Paris. And, that information is available this day and age the almanac is essentially Google, Wikipedia, they all provide you lots of information. How are these information collected? These are summaries over a long time and that is the information that is called background. The background essentially comes from a summary of the prior information. It is also called climatology.

So, the background can come from several different directions from previous knowledge, climatology previous forecasts and so on and so forth. So, all the goblet go mix of everything I know up until this time, if I can spread it over the grid and embed this as a vector that I call as backgrounds. z is the new information R m is called the observation space. Again, we have talked about it at great length in several occasions.

(Refer Slide Time: 11:05)



PRIOR OR BACKGROUND INFORMATION

- Prior information x_B is derived from the previous forecast and/or climatology
- Define the background error as

  $\tilde{x}_B = x - x_B$ -> (1)
- Let $\underline{E(\tilde{x}_B) = 0}$ -> (2)

  $\underline{Cov(\tilde{x}_B) = E[\tilde{x}_B(\tilde{x}_B)^T)} = \underline{B \text{ be known}}$ -> (3)
- Then, $\underline{x \sim N(x_B, B) = Þ(x), \text{ prior distribution}}$ -> (4)

$x \sim N(m, \sigma^2)$

$B \in R^{n \times n}$

$p(x) \rightarrow \underline{\text{ALL MOMENTS}}$

$E[x^k] = \int_{R^n} x^k p(x) dx$

$\int_{R^n} (x - E(x))^k p(x) dx$

MOMENT PROB

3

.

The prior info, so, I am going to call x B as a prior. So, you can see the Bayesian framework comes into play, sneaks in. The x B has been derived from the previous forecast on the climatology. x is the unknown. I want to be able to estimate. I am interested in trying to estimate x or find x. x B is the prior knowledge about x. So, x minus x B is X tilde B. X tilde B is called the difference between the unknown and the background.

The expected value of X B is 0, the covariance of X tilde B is B. B is the matrix. B is the n by n matrix. So, B is a n by n matrix that encapsulates the spatial covariance that underlie the current knowledge of x B and that should be very clear to us from the previous slides and optimal interpolation as well. What was c there? c was a covariance matrix that is derived from the knowledge of the longtime time series.

So, if you look into Google you can talk about what is the maximum temperature in downtown Paris noon every day, there is a prediction. So, you can from this data you can compute B by doing simple statistical analysis. So, I am assuming I know x B, I know B. What is B? B is represent the spatial variability of x B. but, when you and I look at Google to got Paris in the middle of January we don't worry about the variance, we simply say, hey, the maximum temperature. What it say? The maximum temperature in downtown Paris in middle of January could be 5 degrees Celsius plus or minus 1. The plus or minus 1 is indication of x B. 5 degrees is x B. So, we have plus or minus gives you the variation at a given location, I also know the main volume.

So, I am going to now assume p x is the prior. Prior has the mean x B, the covariance is B, I am going to pretend it is normally distributed. Let us consider the, let us discuss that for a moment. Normal distribution, what is the beauty of a normal distribution? Normal distribution is probably the only distribution; I think it is the only distribution that is uniquely determined by the mean in the variance.

So, there is a general result in probability theory, given a density function p of x, I can compute all the moments. What are the moments? First moment, second moment, third moment, central moment and non central moments; so, what is the expression for the kth moment? The expression for the kth moment is expected value of x to the power k, which is equal to integral. Let us assume scalar variable that is equal to x to the power of k px dx. I can also consider central moments or centered moment; X minus E of x to the power k, p of x dx. I can compute this for every k. So, given p, I can compute all moments.

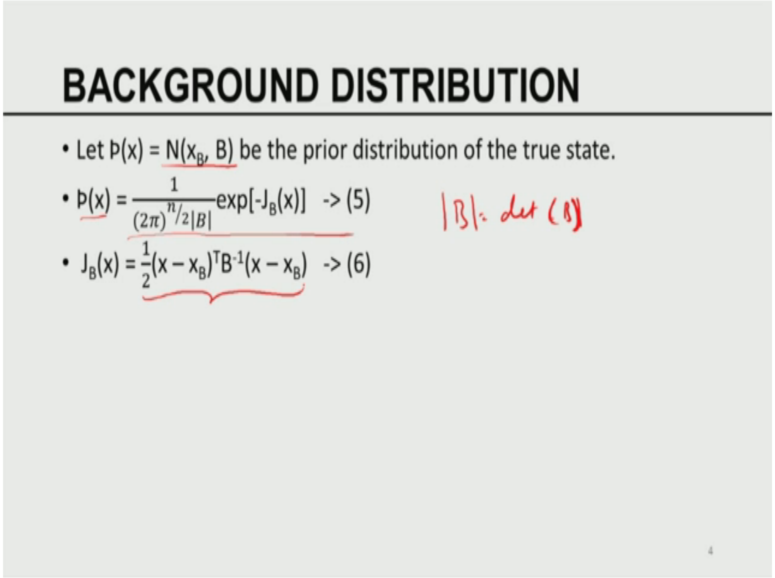
Now, let us consider the converse. How many moments should I know in order that I can build the probability density function. So, given the probability density function I can compute all moments simply evaluation of the integral. Some integrals can be evaluated explicitly, some integral almost all the integrals can be evaluated to a very high degree of approximation numerically. So, the passage from distribution to moment, I can do anything.

The converse question is a fundamental question. That question has come to be known as moment problem. What is the statement of the moment problem? The moment problem tells as the following question, how many moments should I know in order to be able to build the underlying probability density function. There is a very good proof in the first volume of Fellers book that essentially tells you one need to know infinitely many moments to be able to reconstruct p of x. So, the converse problem of going from moments to distribution is a much tougher problem.

So, within the context of this moment problem, now, I want to mention that normal distribution, Gaussian distribution, is the only unique distribution that is determined by the mean and the variance. Why? If I say x is a Gaussian distribution x is uniquely decided by the mean and the covariance or the variance. Therefore, in many meteorological application, many statistical application if I am given a bunch of data, from the data I can grind the mean, from the data I can grind the sample mean and the sample variance, sample covariance. If I have a sample mean and sample covariance I am going to pretend that underlying random variable is distributed with a normally with the sample mean as the mean and sample covariance is the also covariance.

So, that is the reason why even though we only know the first moment and second moment we take a big liberty in assuming that the background is normally distributed with the mean x B and the covariance B. It is a bit of a stretch all we know is only x B and B. To go from x B and B to the normality with x B as the mean and the B as the covariance is the leap in faith, but we often times do that.

(Refer Slide Time: 18:05)



BACKGROUND DISTRIBUTION

- Let Þ(x) = N(x_B, B) be the prior distribution of the true state.
- Þ(x) = $\frac{1}{(2\pi)^{n/2}|B|}\exp[-J_B(x)]$ -> (5)
- $J_B(x) = \frac{1}{2}(x - x_B)^T B^{-1}(x - x_B)$ -> (6)

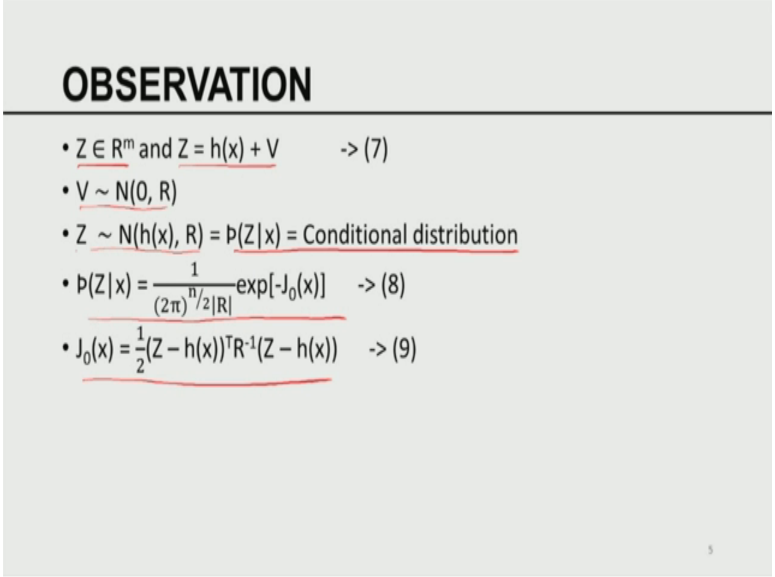
$|B| = \det(B)$

So, for those of us who would like to look at the specificity of the distribution function to say p of x is normal, is to say p of x has this functional form, 1 over 2 pi to the power n by 2 determinant, so, this is the determinant of the matrix B, exponential of minus J B of x, where J B is given by this quadratic function. Now, please understand this kind of quadratic functions we have seen several times over and over and again. So, J B looks like the least square criterion.

This J B which looks like the least square criterion comes as an exponent to an exponential function. Look at this now, least square was invented by Gauss. Gauss in distribution is normal. So, the least square functional form trying to describe the functional form of the normal distribution, it is not an accident. It is all invented by the same person.

So, it is this J B, the mean square type objective function gives raise to the bell shaped curve. So, this is the functional form, it is very important to understand this functional form. So, this is the information about the background.

(Refer Slide Time: 19:42)



OBSERVATION

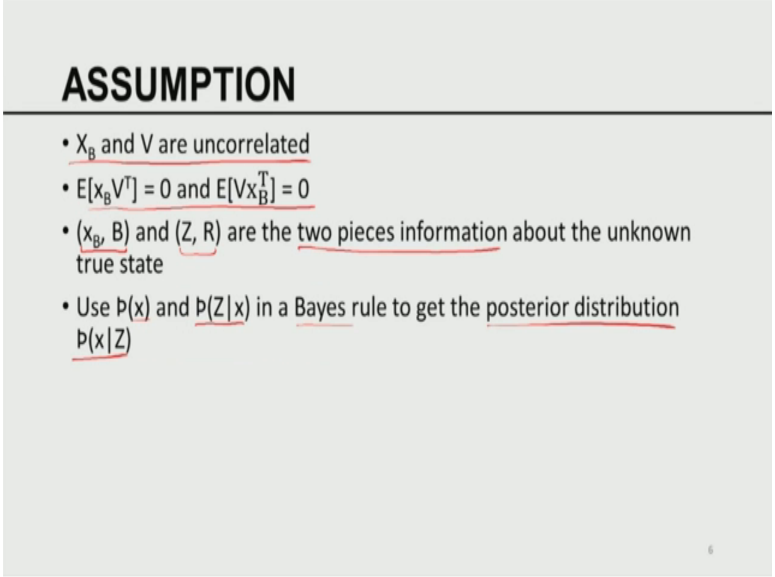
- $Z \in R^m$ and $Z = h(x) + V$ -> (7)
- $V \sim N(0, R)$
- $Z \sim N(h(x), R) =$ Þ$(Z|x)$ = Conditional distribution
- Þ$(Z|x) = \frac{1}{(2\pi)^{n/2}|R|}\exp[-J_0(x)]$ -> (8)
- $J_0(x) = \frac{1}{2}(Z - h(x))^T R^{-1}(Z - h(x))$ -> (9)

Now, let us go to the observation Z is the observation. Z is equal to h of x plus V. Z contains information about the state. So, it’s the indirect measurements of the state. I am going to assume V is normally distributed. So, Z because of the additivity, because of the uncorrelated nature of x and V, the conditional distribution p of Z given x is given by this distribution where J O is given by this. You can see the relation between 5 and 6 and the relation between 8 and 9.

So, 5 and 6 relates to prior, the background is taken as the prior. So, what is the assumption here? The prior information is normally distributed with a known mean and the known covariance. The conditional distribution given x remains given x the conditional distribution is given by 8 and 9. So, I want you to take a good look at J B in 6 and J O in 5. In particular, I would like to draw attention to that J O in 9 J B, B refers to background, J O, O refers to the observation.

So, Z minus h of x, I am sure you will recall this, Z minus h of x is the residual, this is the residual. So, this is essentially least square criterion we have already used, but we did not say anything about least squares. We simply talked about exponent of underlying normal distribution. So, I would like you to develop an appreciation, least squares, normality they are all intimately associated with each other.

(Refer Slide Time: 21:45)



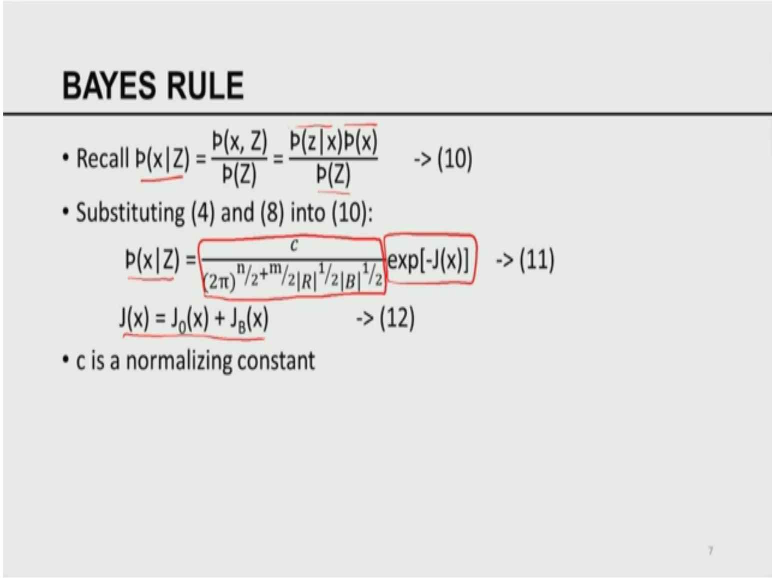
ASSUMPTION

- X_B and V are uncorrelated
- $E[x_B V^T] = 0$ and $E[V x_B^T] = 0$
- (x_B, B) and (Z, R) are the two pieces information about the unknown true state
- Use Þ(x) and Þ(Z|x) in a Bayes rule to get the posterior distribution Þ(x|Z)

So, what are the assumptions? X B and V are uncorrelated. X B, what is X B? X B is the background. The background and the observation noise uncorrelated. These are standard assumptions. Why are these standard assumptions? They are very meaningful. X B and V and 0. V and X B; that means, there is no correlation between V and X B. So, X B, B refers to the prior, Z, R refers to the new information. These are 2 pieces of information about the unknown 2 state.

Please, understand. When we didn't have prior, we only had observation. We know how to do data assimilations for static problem, we used least squares. Now, we are going to invoke to the Bayesian type of analysis, I have prior, I have new information. We are going to use p of x, the prior conditional distribution in a Bayes rule, to get the posterior distribution and that is the underlying philosophy of 3D VAR.

(Refer Slide Time: 22:54)



BAYES RULE

- Recall $\underline{\text{Þ}(x|Z)} = \frac{\text{Þ}(x, Z)}{\text{Þ}(Z)} = \frac{\text{Þ}(z|x)\text{Þ}(x)}{\text{Þ}(Z)}$ -> (10)
- Substituting (4) and (8) into (10):

$$\underline{\text{Þ}(x|Z)} = \boxed{\frac{c}{(2\pi)^{n/2+m/2}|R|^{1/2}|B|^{1/2}}}\boxed{\exp[-J(x)]} \quad \text{-> (11)}$$

$$\underline{J(x) = J_0(x) + J_B(x)} \quad \text{-> (12)}$$

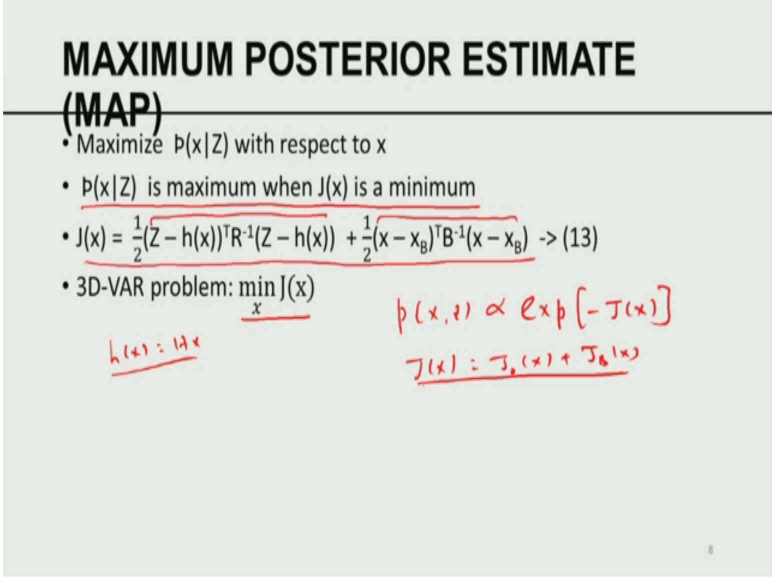
- c is a normalizing constant

So, recall the Bayes rule. The posterior, this is the posterior, that is the prior, that is a conditional distribution; p of Z is a normalizing constant. I can now substitute, look at this, now I have expression for p of x. I have expression for conditional distribution, both are normal, I can substitute everything. If I did that I get this multiplying constant times, exponential minus J of x; minus J of x is simply J naught of x plus J B of x, where c is some normalizing constant. We have looked at this kind of ratio and under normal distribution we have done several exercises in the chapter on Bayes least squares estimation. So, I am trying to do very much similar to what I did in the past.

Now, what is that I would like to do? I would like to be able to. So, now look at this now before I go further, what is that a posterior? You can think of as a likelihood a (Refer Time: 24:09) Fisher. Are you with me, please? So, fisher did not have background. So, he only had the conditional distribution. So, he considered the notion of likelihood and maximizes the likelihood.

Here, we have posterior within the base in context we have already seen the posterior mean is the best estimate. So, I would like to be able to maximize, minimize, appropriate quantities given this setup. That is the optimization problems of interest. So, I hope the expression 11 and 12 are very clear. Now, 12, what is 12? I want to draw your attention to 12. 12 is simply sum of 2 quadratic forms J naught arising from observation J B arising out of background.

(Refer Slide Time: 25:01)



MAXIMUM POSTERIOR ESTIMATE (MAP)

- Maximize $p(x|Z)$ with respect to x
- $\underline{p(x|Z) \text{ is maximum when } J(x) \text{ is a minimum}}$
- $\underline{J(x) = \frac{1}{2}\overbrace{(Z - h(x))^T R^{-1}(Z - h(x))} + \frac{1}{2}\overbrace{(x - x_B)^T B^{-1}(x - x_B)}}$ -> (13)
- 3D-VAR problem: $\underline{\min_x J(x)}$

$\underline{h(x) = Hx}$

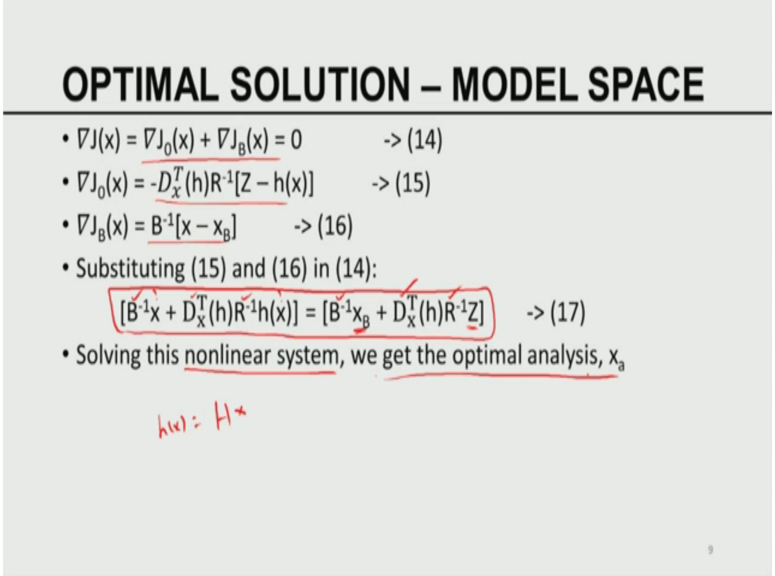
$p(x,z) \propto exp[-J(x)]$

$\underline{J(x) = J_0(x) + J_B(x)}$

So, I would like to be able to maximize p x of Z with respect to x we can see p x of Z; that means, I want to find the location where the p x of Z is maximum. When p x of Z please remember p x of Z is proportional to barring a constant exponential minus J of x and J of x is equal to J naught of x plus J B of x.

So, maximizing exponential of minus J of x is equal to minimizing J of x. Therefore, the algorithm reduces to minimizing this term, this is the background term, this is the observation term, when h of x is equal to linear h of x, both the terms are quadratic. Even h of x is not equal to h linear the first term could be non-linear, but second term is still quadratic. I am interested in minimizing J of x which is given in 13; this minimization problem has come to be called the 3D VAR problem. The 3 dimensional variation assimilation problem is essentially minimizing the 2 quadratic forms; one coming from background another coming from observation and so, minimize J of x with respect to x is an unconstrained minimization problem lot easier, there is no constraint.

(Refer Slide Time: 26:48)



OPTIMAL SOLUTION – MODEL SPACE

- $\nabla J(x) = \underline{\nabla J_0(x) + \nabla J_B(x)} = 0$ -> (14)
- $\nabla J_0(x) = \underline{-D_x^T(h)R^{-1}[Z - h(x)]}$ -> (15)
- $\nabla J_B(x) = \underline{B^{-1}[x - x_B]}$ -> (16)
- Substituting (15) and (16) in (14):

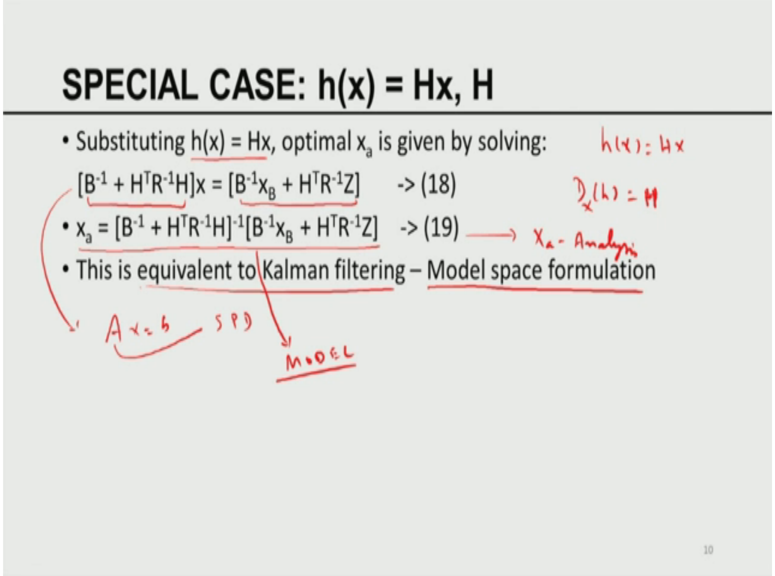
$$\boxed{[B^{-1}x + D_x^T(h)R^{-1}h(x)] = [B^{-1}x_B + D_x^T(h)R^{-1}Z]} \quad \text{-> (17)}$$

- Solving this nonlinear system, we get the optimal analysis, x_a

$h(x) = Hx$

We can compute the gradient. Gradient of the sum is some other gradients Hessian of the sum is some of the Hessians. So, gradient of J naught gradient of J B substituting a summing and equating that to 0, I get the general equation, now look at this now right hand side involves x B, Z, B Jacobian of h and R, everybody is known and the left hand side B is known, R inverse is known, Jacobian of h is known, x I do not know, when h of x is equal to H of x, this reduces to a linear relation which can be solved by any one of the methods. In principle 17 is a non-linear system, is a non-linear algebraic system whose solution gives you the optimal analysis. If it is a non-linear algebraic system the only way to solve the non-linear algebraic system is by numerical methods. So, there are tons of numerical methods one knows to solve 17.

(Refer Slide Time: 28:12)



SPECIAL CASE: h(x) = Hx, H

- Substituting h(x) = Hx, optimal x_a is given by solving:

  $[B^{-1} + H^T R^{-1} H]x = [B^{-1} x_B + H^T R^{-1} Z]$ -> (18)
- $x_a = [B^{-1} + H^T R^{-1} H]^{-1} [B^{-1} x_B + H^T R^{-1} Z]$ -> (19) → x_a - Analysis
- This is equivalent to Kalman filtering – Model space formulation

h(x) = Hx
$D_x(h) = H$
Ax = b
SPD
MODEL
10

In the special case I would like to reemphasize that, in the special case when h of x is equal to H of x in this in. So, then when h of x is equal to H of x, the jacobian of h is equal to H, everything simplifies; therefore, the matrix like this times x times this matrix and that is the optimal solution is given by this. This system I have to solve this system is like A x is equal to b, again A b is SPD. We already know how to solve a SPD systems and it is no accident that this is exactly the same relation that we derived in the model space formulation in the Bayesian scheme.

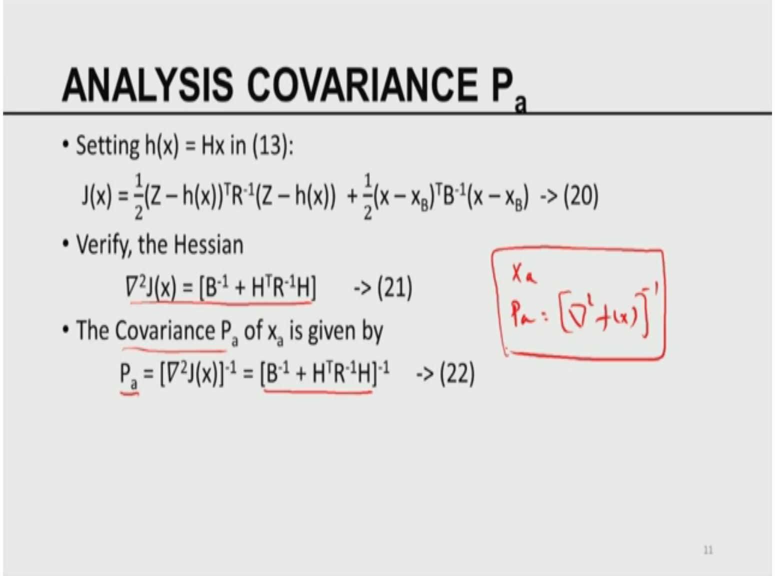
So, this, see the model space formulation, so, this is equivalent to Kalman filter equation. So, the model space formulation. So, what does this tell you as we will see when we come to Kalman filter equation; Kalman filtering essentially consists of the following step namely, the background is replaced by a forecast, a new observation comes in. How do you combine forecast and the observation that is the Kalman filter equations? So, you can readily see what happens at the filtering stage of the Kalman filter. Kalman filter has multiple stages.

One of the stages in Kalman filter is called the filtering phase or a filtering stage. The filtering stage of the Kalman filter is equivalent to 3D VAR, that is a story and 3D VAR is very much related to the Bayesian framework and it is no surprise here we are coming back to Kalman, because we already established the relation between Bayesian least square and linear minimum variance estimate, one did in the model space another did in the observational space we built the bridge between the 2 using the matrix identity called

Sherman-Morrison-Woodbury formula.

So, now I believe the whole thing is falling implies Bayesian, linear minimum variance, 3D VAR, Kalman filter, all these things are all close cousins of each other and they even though the idea is come in different with the different labels and different names and the underlying mathematics are very nearly similar or same. So, the equation to this gives you in an. So, the equation to 19 gives you the analysis. The optimal solution for this is called the analysis. So, that is called X a analysis. X a is the analysis. Please realize analysis is the fancy term that geophysicist use for the posterior estimate. So, what statistician's calls as posterior estimate is essentially analysis in geophysical sciences.

(Refer Slide Time: 31:29)



ANALYSIS COVARIANCE P_a

- Setting h(x) = Hx in (13):

$$J(x) = \frac{1}{2}(Z - h(x))^T R^{-1}(Z - h(x)) + \frac{1}{2}(x - x_B)^T B^{-1}(x - x_B) \quad \text{->} (20)$$

- Verify, the Hessian

$$\underline{\nabla^2 J(x) = [B^{-1} + H^T R^{-1} H]} \quad \text{->} (21)$$

- The Covariance P_a of x_a is given by

$$\underline{P_a} = [\nabla^2 J(x)]^{-1} = \underline{[B^{-1} + H^T R^{-1} H]}^{-1} \quad \text{->} (22)$$

$$\boxed{\begin{array}{l} X_a \\ P_a = [\nabla^2 J(x)]^{-1} \end{array}}$$

11

So, we have already seen this. The Hessian of the J function again I would like you to verify. We have already seen when we did statistical least squares the inverse of the Hessian is indeed the analysis, covariance of the analysis. So, I have X a, I have P a, is essentially x inverse, that is the Hessians. So, this is the analysis covariance. So, this is P analysis covariance. So, I have analysis X a from the previous slide given by equation 19.

The analysis covariance is given by 12. Why am I interested in analysis covariance, because, we are interested we have been talking about statistical methods. In a statistical method mean alone will not cut I need to be able to provide information about the underlying variability as well.

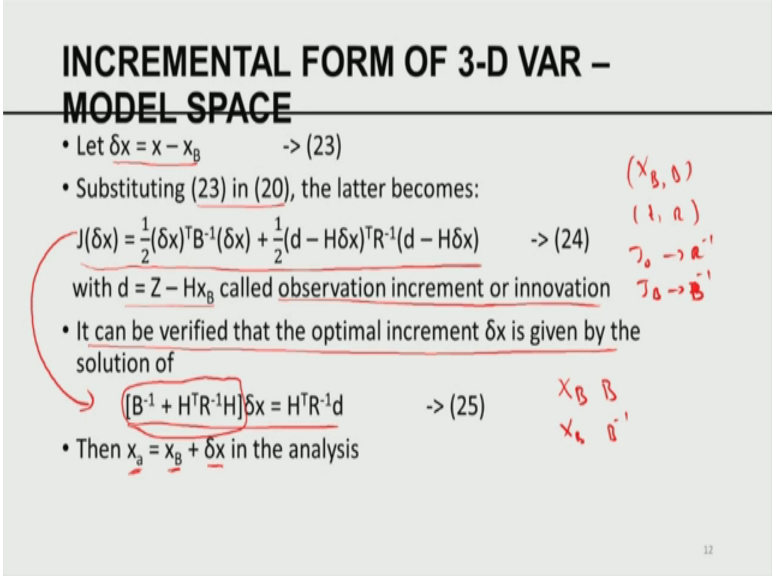
So, in most of the stochastic predictions we are not only interested in the level given by the mean, but also we are interested in getting some feel for the underlying covariance. So, I have both the mean and the covariance. So, this is the best we could do. You gave me the first 2 moments of both the background as well as the observation.

Now; I have combined the first 2 moments of the background of the observation to create the first moment and the second moment of the analysis, the mean of the analysis, the covariance of the analysis. How do I use the covariance of the analysis? Analysis of covariance matrix is a matrix. The diagonal elements are all the individual variance. The sum, the trace, the sum of the diagonal elements which is equal to trace, the total amount of variance across all the components that provides you a good measure, the quality, of the analysis, if that variance is large you cannot put too much faith on your analysis, if that variance is smaller analysis has lot more credibility lesser variation.

So, that is how we would like to be able to whenever we generate an analysis. What is an analysis? Analysis is kind of a prediction. What good is a prediction, if I do not know how good, what is the degree of variability in my prediction? With respect to a lunar eclipse and solar eclipse variability is 0. We know precisely. Can we predict the temperature on noon on March 15th at downtown London? We can't. We can, I can of course, I can generate a number, but what good is that number, unless I know what is going to be underlying variance associated with it.

So, prediction, deterministic prediction, stochastic prediction; deterministic prediction, we are content with one number. In stochastic prediction we also need to have a number that gives us a level also we need to give the variability. So, if you look at current predictions of local weathers they will say, tomorrow there is a chance of 1 inch rain 80 percent probability. So, 1 inch rain is the measure of the mean level of the rain, 80 percent, wow! That is a very high percentage. So, we need to be able to not only give the level, but also an associated variance which is the measure of the variability in my or confidence in my prediction. So, in stochastic predictions I need both mean as well as the variance or the covariance.

(Refer Slide Time: 35:18)



INCREMENTAL FORM OF 3-D VAR – MODEL SPACE

- Let $\delta x = x - x_B$ -> (23)
- Substituting (23) in (20), the latter becomes:

$$J(\delta x) = \frac{1}{2}(\delta x)^T B^{-1}(\delta x) + \frac{1}{2}(d - H\delta x)^T R^{-1}(d - H\delta x) \quad \text{-> (24)}$$

with $d = Z - Hx_B$ called observation increment or innovation

- It can be verified that the optimal increment δx is given by the solution of

$$\boxed{[B^{-1} + H^T R^{-1} H]}\delta x = H^T R^{-1} d \quad \text{-> (25)}$$

- Then $x_a = x_B + \delta x$ in the analysis

(X_B, B)

(Z, R)

$J_o \rightarrow R^{-1}$

$J_B \rightarrow B^{-1}$

$X_B \; B$

$X_B \; B^{-1}$

12

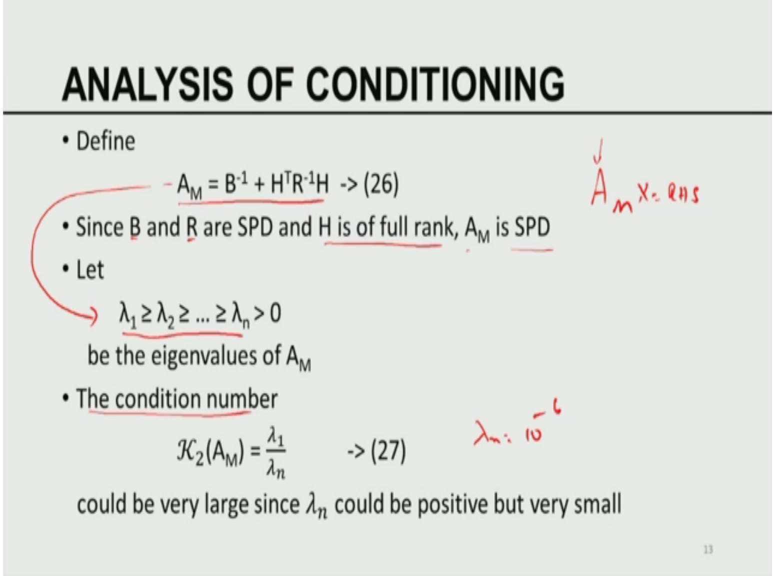
Once we get this, that is, a quick exercise I can talk about the incremental form of 3D VAR in the model space. So, let me run through this very quickly. This is the largely related to computational advantages and disadvantages of whichever formulation we want to follow let this be the increment, this is called incremental formulation substituting this 23 in 20, my J becomes this that you can readily see, where d is equal to Z minus H of x B that is what is called the observation increment or the innovation.

So, we can work with the incremental variable. It can be verified that the optimal increment delta x given by the solution of this in other words if I minimize this, the minimizer is given by the solution of this. What is the idea now? The analysis is given by background plus the increment. The increment comes from this analysis that is another way of looking at it.

One of the difficulties in here in solving this is, this system as well as the previous system let me go back we need to know B inverse, we need to know R inverse. So, there is a question, one can ask oneself, from computational point of view, do I know X B and B do I know X B and R B inverse. Where B inverse comes in? B inverse comes in as the weight in my objective function in the least squares. So, while I know B my data assimilation scheme uses B inverse. B is a very large matrix. n by n, n is a million, is the inverse of a million be million matrix. So, computationally, now one has to ask oneself a question; hey, are we going to really invert these matrices? Is there a way to reformulate the problem without worrying about B inverse, but using B only? I hope you appreciate this problem now.

We know X B, so, what is that may have been given. I know the background information X B and B, I know the observation Z and R, but my J naught uses R inverse. My J B uses B inverse. B inverse, R inverse, who is going to give you? The theory is beautiful, but you need to be able to compute B inverse, R inverse. Look at this now, the left hand side matrix in here involves these inverses. So, where in the world are you going to get these? So, the question here is that is there a way to re-formula, I like the formulation, but I do not like the computational demands of it. Is there a way to reformulate it to an equivalent formulation that does not depend on some of the inverses? So, that is the question that drives some of these.

(Refer Slide Time: 38:15)



ANALYSIS OF CONDITIONING

- Define

$$A_M = B^{-1} + H^TR^{-1}H \quad \text{-> (26)}$$

- Since B and R are SPD and H is of full rank, A_M is SPD
- Let

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0$$

be the eigenvalues of A_M

- The condition number

$$\mathcal{K}_2(A_M) = \frac{\lambda_1}{\lambda_n} \quad \text{-> (27)}$$

could be very large since λ_n could be positive but very small

$A_m x = RHS$

$\lambda_n: 10^{-6}$

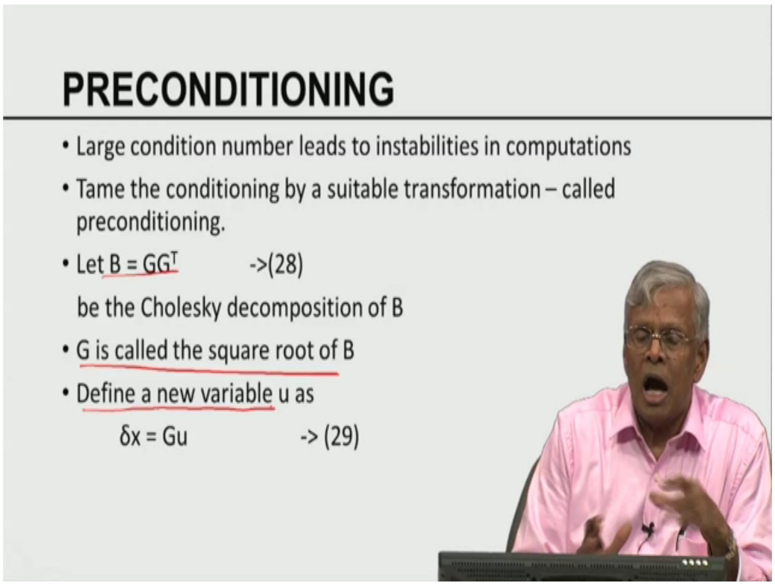
13

So, look at this now A M. So, I am going to now call B inverse plus H transpose R inverse H. As A M, B and R are SPD I am assuming H is a full rank, therefore, A M is SPD. If A M is SPD this combined matrix A M has n Eigen values, the least of the Eigen values is positive. The condition number for this matrix is given by lambda 1 by lambda n. Please understand this is the matrix with which I have to solve A M x is equal to the right hand side, right hand is the vector.

So, the quality of the solution depends on the condition number of this matrix A M, the condition number please recall lambda 1 by lambda n. While A M could be symmetric and positive definite since lambda n positive but lambda n can be very small.

So, the condition number could be very large since lambda n is positive, but small. So, lambda n can be small, but still positive. So, let us assume lambda n as an example is 10 to the power of minus 6, is still positive; but, it comes to the denominator, condition number will be very large. So, condition number for SPD matrices becomes very large because the smallest Eigen value while remaining positive may become very small in those cases. I will have headache computationally.

(Refer Slide Time: 39:59)



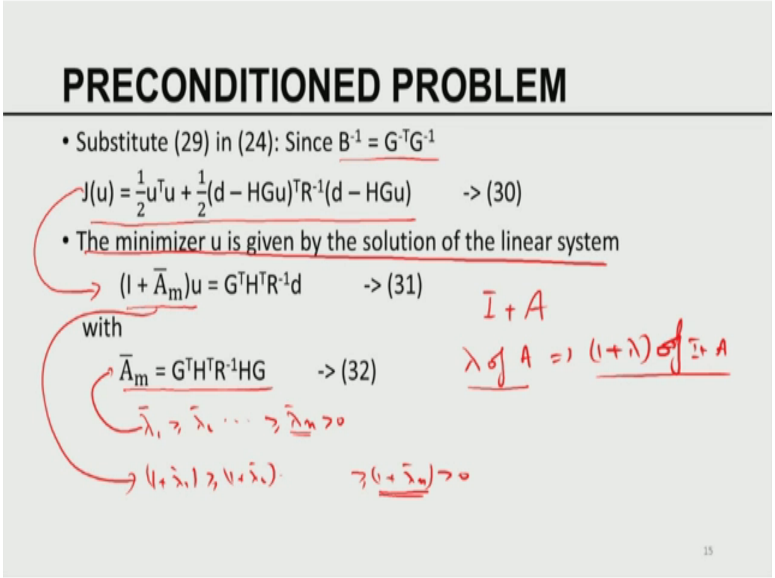
PRECONDITIONING

- Large condition number leads to instabilities in computations
- Tame the conditioning by a suitable transformation – called preconditioning.
- Let $B = GG^T$ ->(28)
  be the Cholesky decomposition of B
- G is called the square root of B
- Define a new variable u as
  $\delta x = Gu$ -> (29)

So, large condition number leads to instabilities and computation. To tame the conditioning we use the suitable transformation called preconditioning. So, preconditioning is the methodology by which I am going to get around the challenges posed by an ill conditioned matrix or a matrix with large condition number. What is the matrix of interest here? A M. What is A M? A M is the sum of B inverse plus H transpose R inverse H. These are all mathematical considerations.

So, how my going to do develop a preconditioning to avoid the challenges of ill conditioned matrices or nearly ill condition matrices, that is what we are going to discuss now. So, we have assumed B is known, we are assumed R is known. If B is known let us pretend I know the Cholesky factors, GG transpose. Every SPD has Cholesky decomposition. So, G is also called Cholesky factor or a square root of B, we have already seen that when we did that. So, I am now trying to define a new variable u in terms of delta x, you can think of this as a linear transformation or you can think of this as a preconditioning transformation whichever way you want to call it.

(Refer Slide Time: 41:18)



PRECONDITIONED PROBLEM

- Substitute (29) in (24): Since $\underline{B^{-1} = G^{-T}G^{-1}}$

$$\underline{J(u) = \frac{1}{2}u^Tu + \frac{1}{2}(d - HGu)^TR^{-1}(d - HGu)} \quad \text{-> (30)}$$

- The minimizer u is given by the solution of the linear system

$$(I + \bar{A}_m)u = G^TH^TR^{-1}d \quad \text{-> (31)}$$

with

$$\underline{\bar{A}_m = G^TH^TR^{-1}HG} \quad \text{-> (32)}$$

$\bar{\lambda}_1 \geq \bar{\lambda}_2 \cdots \geq \underline{\bar{\lambda}_n} > 0$

$(1+\bar{\lambda}_1) \geq (1+\bar{\lambda}_2) \cdot \quad \geq \underline{(1+\bar{\lambda}_n)} > 0$

$I + A$

$\underline{\lambda \text{ of } A} \Rightarrow \underline{(1+\lambda) \text{ of } I + A}$

15

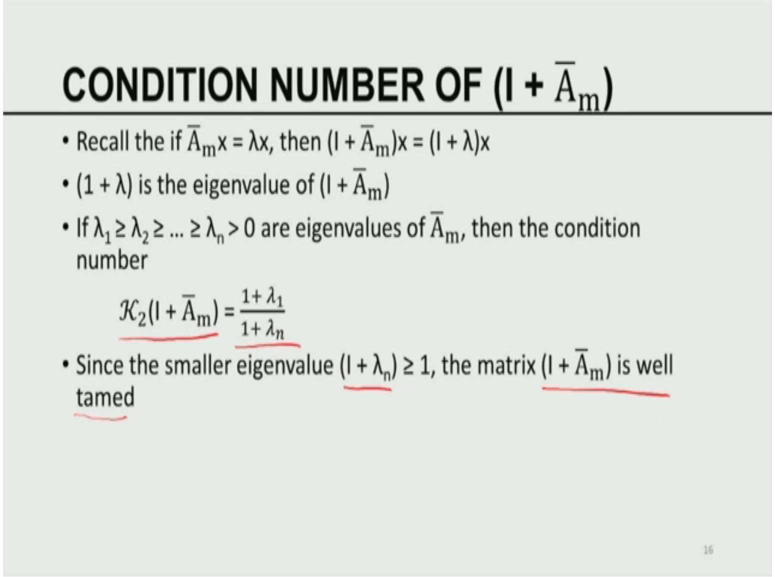
So, with this delta x is equal to Gu; look at this now, if B is equal to GG transpose in 28, B inverse is going to be G minus T G inverse minus T means transpose of the inverse. So, if I substitute this linear transformation my J u becomes this. Look at this now in the first term there is no B. B is gone in the new variable and what is the minimizer? The minimizer of this is given by this. So, I am not trying to show you all the steps, this 30 is a quadratic function, compute the gradient, equate the gradient to 0, equate it to 0. You get the optimal solution in the new variable u to be given by 31. I hope you keep track of everything.

Now, what is the matrix A m? A m is given by G transpose H transpose R inverse HG. Now, let us look at this matrix I A m. So, I have the matrix I, so, let us instead of A m I will consider the matrix A. What is the general resulting matrix theory? If lambda is an Eigen value of A implies 1 plus lambda is an Eigen value of I plus A, that is a fact. So, if let us assume now if lambda bar 1 is greater than equal to lambda bar 2 greater than equal to lambda bar n greater than equal to 0. So, if you A m is a symmetric positive definite its Eigen value. The least Eigen value is lambda bar n, please understand, I use lambda 1 lambda 2 lambda 3 for A, I am using lambda bar 1 bar 2 bar n for A bar m.

So, the Eigen values let this be the Eigen values of this. The Eigen values of I plus m are 1 plus lambda 1 bar greater than equal to 1 plus lambda 2 bar greater than equal to 1 plus lambda n bar this greater than 0. Now, look at this now the least Eigen values of 1 plus lambda n bar 1 plus lambda n bar can never be close a 0. Lambda n bar is positive, but 1 plus that mean I am bounding the value of the least Eigen value; that means, my condition number is not going to explore in this new space therefore, I have tamed the condition number by a useful transformation.

So, such transformations are called preconditioning methods. Preconditioning methods have been known for a long time that is essentially motivated by to tame the instabilities that may arise because of large are a large condition number ill conditioned cases.

(Refer Slide Time: 44:25)



CONDITION NUMBER OF $(I + \bar{A}_m)$

- Recall the if $\bar{A}_m x = \lambda x$, then $(I + \bar{A}_m)x = (I + \lambda)x$
- $(1 + \lambda)$ is the eigenvalue of $(I + \bar{A}_m)$
- If $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0$ are eigenvalues of $\bar{A}_m$, then the condition number

$$\underline{\mathcal{K}_2(I + \bar{A}_m)} = \frac{1+\lambda_1}{\underline{1+\lambda_n}}$$

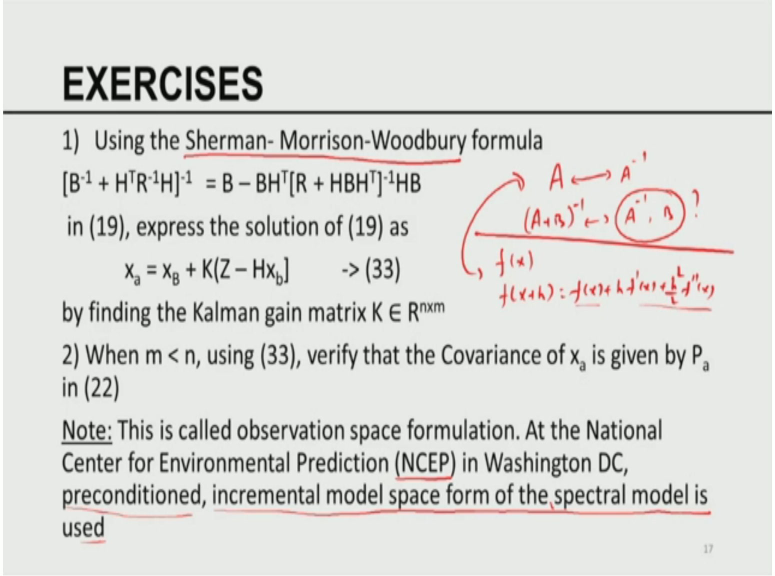
- Since the smaller eigenvalue $\underline{(I + \lambda_n)} \geq 1$, the matrix $\underline{(I + \bar{A}_m)}$ is well tamed

16

So, I have already talked about all this. So, the condition number for the new matrix is this and this is always the least Eigen value is always bounded away from 0 instead of being bounded away from 0. So, this matrix is well tamed. If this matrix is well tamed I should not expect any difficulty in numerical analysis in numerically trying to compute the solution when I use any method any meaningful method to solve.

So, with that we come to the end of the discussion of the so called 3D VAR. So, in summary 3D VAR is another way of looking at the Bayesian formulation which we have already talked about within the context of statistical estimation theory, especially development of Bayes least squares theory. We are trying to reinforce it by using a new language that is prevalent in atmospheric sciences in data assimilation called 3D VAR and we talked about the nature of the solution we talked about the potential for ill conditioning, we also talked about how do we tame using a preconditioning transformation to be able to induce better conditioning for matrices so that I can have reasonably good stable numerical results.

(Refer Slide Time: 45:44)



EXERCISES

1) Using the Sherman- Morrison-Woodbury formula

$[B^{-1} + H^TR^{-1}H]^{-1} = B - BH^T[R + HBH^T]^{-1}HB$

in (19), express the solution of (19) as

$x_a = x_B + K(Z - Hx_b]$ -> (33)

by finding the Kalman gain matrix $K \in R^{nxm}$

2) When m < n, using (33), verify that the Covariance of x_a is given by P_a in (22)

Note: This is called observation space formulation. At the National Center for Environmental Prediction (NCEP) in Washington DC, preconditioned, incremental model space form of the spectral model is used

$A \longleftrightarrow A^{-1}$

$(A+B)^{-1} \longleftrightarrow \boxed{A^{-1}, B}$?

$f(x)$

$f(x+h) : f(x) + hf'(x) + \frac{h^2}{2}f''(x)$

I would like to quickly bring to your attention to the Sherman-Morrison-Woodbury formula. You can see the importance of this formula. Again, I want to tell you something Sherman-Morrison-Woodbury formula was developed in mathematics for the sake of interest in matrix identity. What is that they were interested in? They were interested in the following; if A is the matrix, if I know A inverse, if I perturbed that matrix by adding another matrix if I am interested in a inverse can I find the inverse of A plus B which is the perturbed matrix, simply as a function of A inverse and B. That is the question they were interested in.

In other words, if I know how if I have a routine to compute inverse of a matrix I can plow the entire new matrix A plus B through the algorithm and can compute it, but that is not that what they were interested in. I know the inverse of A has been updated by B, how do I update a inverse to get the inverse of A plus B. That is the idea. The idea is very similar to what we do in Taylor series. What is the example of in Taylor series? I know the value of the function at f of x, I would like to be able to know the value of the function at f of x plus h and what do we do we simply say this is f of x plus h times f prime of x plus h square divided by 2 times f double prime of x.

So, what is the idea in general both here as well as in here knowing what I know, how can I extrapolate my knowledge in a neighborhood of around what I know. So, I know A inverse from A, A plus B is a matrix which is close to A. So, if I know somebody close to somebody who is invertible can I express my inverse based on the inverse of a fellow whom I know is invertible and I know his inverse; that is idea. Same thing in here, I know the value function at a point x, I am considering neighborhood of a point f of x plus h how do I extrapolate the value of f of x to f of x plus h I can do it by knowing the value of the function and it is derivative.

Here, I simply want to be able to use A inverse and B and that is the beauty of Sherman-Morrison-Woodbury formula. When they developed it they had no idea whether it will be used at all. They did it for the sake of mathematic, for the sake of beauty. In fact, Sherman-Morrison formula plays such a crucial role in data assimilation, especially in trying to relate the data assimilation formulation from model space to observation space. Why is this relation needed? If the model space is the n dimensional observation space is m dimensional we result them m is equal to n. So, either m is greater than n or m is less than n.

So, what is that we would like to take advantage of? It is always cheaper to perform computation in a smaller dimensional space. So, n is smaller you do the operations and up and model space if m is smaller do the operations in the observation space who provides this freedom to go between these 2 world, Sherman-Morrison-Woodbury formula, that is beauty.

It has no physics, it does not meteorology does nothing it is yes yeah beautiful mathematical enterprise that helps you to build a bridge between these 2 world and that relieves us of the pain of having to do all the computations only in one world irrespective whether m is greater of m is less. It is this power of the Sherman-Morrison-Woodbury formula, I would like you to understand appreciate and see how repeatedly we have used in our discussion of data assimilation.

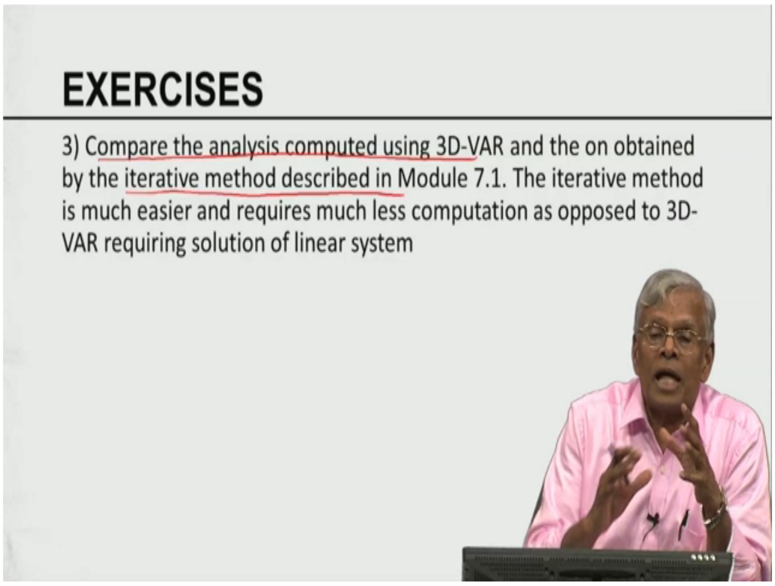
I would like to make one or 2 comments NCEP in the United States, which is called the National Center for Environmental Protection, they use the preconditioned incremental model space for the spectral model. So, it is mouthful. Let me tell you this now. What is the model they use? They use spectral model for the phenomena. What is the spectral model? So, consider a primitive equation in a spherical domain. You express the solution in the spherical harmonics as a Fourier series; you substitute the expansion in the spherical domain in the spherical harmonics into the differential equation. You reduce a system of partial differential equation to a system of ordinary differential equation on the amplitudes of the spherical harmonics.

So, what is the expansion of the field variables such as velocity, such as pressure, such as temperature? So, what is that I am assuming? I can encapsulate the spatial variation by involving spherical harmonics. What I do not know, it is the amplitude. So, we assume the spatial variation can be captured by a cover combination of spherical harmonic functions; amplitude is something I do not know. So, by substituting this Fourier series expansion in spherical domain into the primitive equation model I can reduce the infinite dimensional partial differential equation to a finite dimensional ODE. The resulting finite dimensional ODEs have come to be called either low order models or spectral models.

So, spectral models are essentially resulting from the application the Fourier analysis to the continuous domain. So, they consider the spectral model. They do a model space formulation, they use an incremental version; what is incremental version, delta x, they use the preconditioned incremental model space formulation of the spectral model. I hope you are now appreciate that; that means, there is a background, I want to have the analysis, I want to go from prior to posterior, background is prior, posterior is analysis. I would like to be able to express the analysis, the posterior as background plus an increment, it is that increment in this incremental formulation that is being computed, that is what we have expressed that, because of the computational problem they use the preconditioned version and they also use model space formulation.

So, model space formulation of the spectral model precondition increment as opposed to NCEP in NASA they use the observation a space formulation. So, different places develop programs, systems in different formulation because of their belief for the targeted applications in mind.

(Refer Slide Time: 53:50)



EXERCISES

3) Compare the analysis computed using 3D-VAR and the on obtained by the iterative method described in Module 7.1. The iterative method is much easier and requires much less computation as opposed to 3D-VAR requiring solution of linear system

So, I would like to leave you with a very interesting computational project it could be a very nice master's thesis. In fact, compare, so, what is that? You have a 3D VAR problem, I can do an iterative method customer like iterative method I can do a 3D VAR like method. So, it behooves us to ask a question to compare the results of 3D VAR with iterative methods and also I not only would like you to compare the quality of the result, but also the computational requirements. I think these kinds of projects at the masters level would be very educative, very inspiring people to be able to look at different kinds of methodologies that could be brought to back to solve different types of data assimilation problems of interest in daily life.

Thank you.