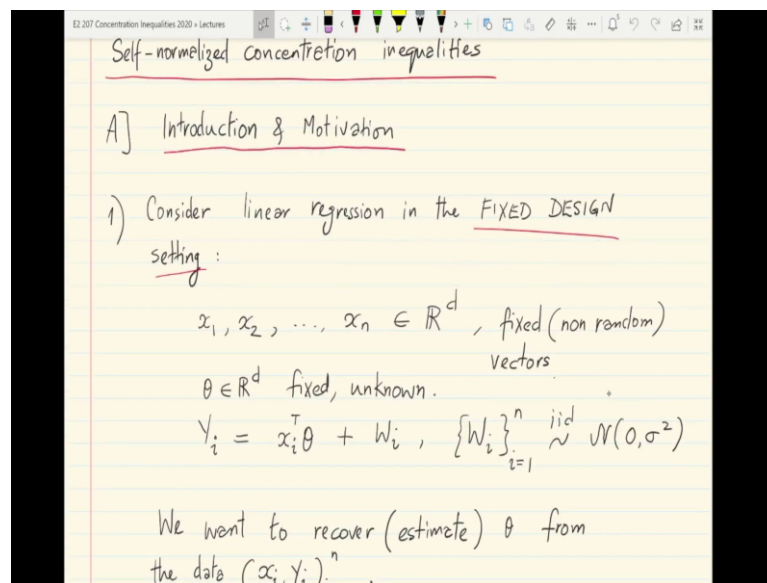


Concentration Inequalities
Prof. Aditya Gopalan
Prof. Himanshu Tyagi
Department of Electrical Communication Engineering
Indian Institute of Science, Bengaluru

Lecture - 26
Self normalized concentration inequalities and application to online regression

(Refer Slide Time: 00:20)



Self-normalized Concentration Inequalities

A] Introduction & Motivation

1) Consider linear regression in the FIXED DESIGN setting:

$x_1, x_2, \dots, x_n \in \mathbb{R}^d$, fixed (non random) vectors.

$\theta \in \mathbb{R}^d$ fixed, unknown.

$y_i = x_i^T \theta + W_i$, $\{W_i\}_{i=1}^n \sim \text{iid } \mathcal{N}(0, \sigma^2)$

We want to recover (estimate) θ from the data $(x_i, y_i)_{i=1}^n$.

Hi all. This talk is going to about what are called Self normalized concentration inequalities and these are again this is a method to prove concentration inequalities for functions of sequential processes, where current random variables depend on previous random variables and so, there is a lot of time correlation.

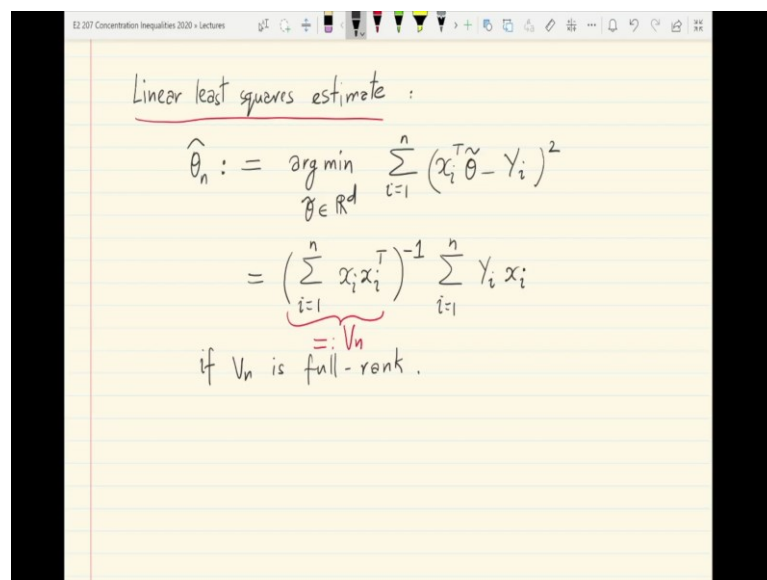
So, let us introduce the necessity for these kinds of concentration inequalities in the context of motivating example involving estimation a parametric estimation ok. So, what is the so, what is the arguably simplest kind of parametric estimation problem? Well, it is often a regression problem.

So, let us consider a standard let us start off considering a standard least squares regression problem linear regression problem in what we call the fixed design setting ok. So, what is the fixed design setting? You have n d dimensional vectors x_1, x_2 up to x_n these are fixed or non random vectors also called your design ok.

So, these are essentially vectors that represent your choice the experimenter or data collector's choice of directions to measure upfront. So, the design vectors x_1 through x_n are fixed upfront. Let θ be a d dimensional unknown parameter and let Y_i be a random variable which is the inner product of x_i with θ with an additive Gaussian noise independent Gaussian noise variable W_i which has mean 0 and standard deviation σ .

And the goal is as usual. We want to recover or estimate θ from our data x_i, Y_i ok. So, this is the standard regression problem with Gaussian noise additive Gaussian noise. And the most natural way to estimate θ is what is called least squares regression or linear least squares estimation.

(Refer Slide Time: 02:37)



Linear least squares estimate :

$$\hat{\theta}_n := \arg \min_{\theta \in \mathbb{R}^d} \sum_{i=1}^n (x_i^T \theta - Y_i)^2$$

$$= \left(\sum_{i=1}^n x_i x_i^T \right)^{-1} \sum_{i=1}^n Y_i x_i$$

if V_n is full-rank .

(Note: In the original image, a red bracket under the sum $\sum_{i=1}^n x_i x_i^T$ is labeled V_n .)

So, the linear least squares estimation is solving the following optimization problem. So, the estimate for θ from these n data is defined to be $\hat{\theta}_n$ which is the solution to a minimization problem over all possible θ in \mathbb{R}^d of the sum over your θ of the fitting error.

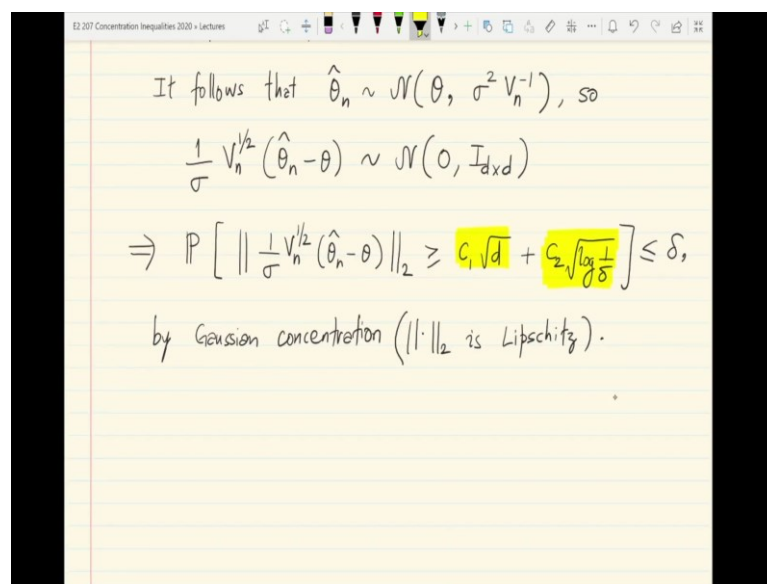
So, $x_i^T \theta - Y_i$ and if the x_i 's are chosen let us say. So, that they span the entire space \mathbb{R}^d then it is not difficult to show that this has an explicit closed form solution as sum over i going from 1 to n $x_i x_i^T$. So, this is a matrix inverse. So, if the x_i 's all span θ then this matrix is going to be full ranks and hence invertible into

this particular vector which is the weighted sum of the x_i 's with the respective scalar measurements y_i ok.

So, this is true if so, let us denote this matrix whose inverse is being taken as V_n . So, this is true if V_n is invertible or equivalently full rank ok. So, this is the standard linear least squares estimate for inference about the unknown parameter θ of the of this linear model and in fact, it turns out that since we have assumed that the x_i 's are all fixed the V_n is a fixed or non random matrix.

So, $\hat{\theta}_n$ is just a fixed linear function of the sequence of Gaussian random variables y_1 through y_n and it follows that since all the random variables y_1 through y_n are multivariate normal together.

(Refer Slide Time: 04:45)



It follows that $\hat{\theta}_n \sim \mathcal{N}(\theta, \sigma^2 V_n^{-1})$, so

$$\frac{1}{\sigma} V_n^{1/2} (\hat{\theta}_n - \theta) \sim \mathcal{N}(0, I_{d \times d})$$

$$\Rightarrow \mathbb{P} \left[\left\| \frac{1}{\sigma} V_n^{1/2} (\hat{\theta}_n - \theta) \right\|_2 \geq c_1 \sqrt{d} + c_2 \sqrt{\log \frac{1}{\delta}} \right] \leq \delta,$$

by Gaussian concentration ($\|\cdot\|_2$ is Lipschitz).

It follows that $\hat{\theta}_n$ is also the estimate $\hat{\theta}_n$ also has a multivariate normal distribution. So, more precisely it follows that $\hat{\theta}_n$ is distributed as multivariate normal vectors with mean as the original parameter θ and with covariance matrix sigma square V_n inverse.

So, sigma square recall is the variance of the per measurement additive noise and V_n in some sense is the overall design matrix representing how much of each direction in space of d dimensions how much is being explored. So, if we just do some linear algebra it follows that $1/\sigma$ into the $\sqrt{V_n}$ into the difference $\hat{\theta}_n - \theta$ is distributed as a

standard multivariate normal vector ok. This is the left hand side is in some sense just the whitened form of a $\hat{\theta}_n - \theta$.

And this in turn implies by using standard concentration results for multivariate normal vectors that the probability that the ℓ_2 norm of this vector $\frac{1}{\sigma} \sqrt{V_n} (\hat{\theta}_n - \theta)$ ok which is basically the ℓ_2 norm of a multivariate standard multivariate normal in d dimensions exceeds a let us say some constant times the $\sqrt{\text{of dimension}}$ + another constant times $\sqrt{\text{of log } 1/\delta}$ is bounded by δ ok.

This is just by appealing to standard Gaussian concentration since the ℓ_2 norm function is a Lipschitz function with respect to the ℓ_2 norm itself and. So, you can use the Gaussian concentration inequality along with basically this $C \sqrt{d}$ which is actually a bound on the expected value of the ℓ_2 norm of a standard Gaussian vector ok. So, that is where you get this norm.

So, the second term here comes from concentration about the mean the first term is basically by estimating the expected norm of a Gaussian vector itself ok. So, this is standard. So, we basically get this kind of a concentration inequality for the fluctuation of $\hat{\theta}_n$ about θ with respect to the shape V_n as well.

(Refer Slide Time: 07:46)

Denoting $\|x\|_A := \sqrt{x^T A x}$ ("Matrix weighted norm"), this suggests that

$$(\dagger) \quad C_n := \left\{ \beta \in \mathbb{R}^d : \left\| \frac{1}{\sigma} V_n^{1/2} (\hat{\theta}_n - \beta) \right\|_2 < c_1 \sqrt{d} + c_2 \sqrt{\log 1/\delta} \right\}$$

$$= \left\{ \beta : \left\| \hat{\theta}_n - \beta \right\|_{V_n} < \underbrace{\sigma c_1 \sqrt{d}}_{O(\sqrt{d})} + \underbrace{\sigma c_2 \sqrt{\log 1/\delta}}_{O(\sqrt{\log 1/\delta})} \right\}$$

is a probability $(1-\delta)$ -Confidence Set (Ellipsoid)

for the true θ : $\mathbb{P}[\theta \in C_n] \geq 1-\delta$.

Let us introduce let us denote. So, denoting by x subscript A the $\sqrt{\text{of } x^T A x}$. So, when x when A is a positive definite matrix this is often called the matrix weighted

Euclidean norm ok. So, if A is the identity matrix then you get the standard Euclidean norm.

So, this the above suggests that so, if you define this particular d dimensional subset C_n as the set of all β d dimensional vectors such that the l_2 norm of $\frac{1}{\sqrt{n}} \sum_{i=1}^n V_i \beta$ raised to half $\hat{\theta}_n - \beta$ in l_2 norm is less than $C_1 \sqrt{d} + C_2 \sqrt{\log 1/\delta}$ by δ . So, notice that this can equivalently be written by rescaling as the set of all β s such that $\hat{\theta}_n - \beta$.

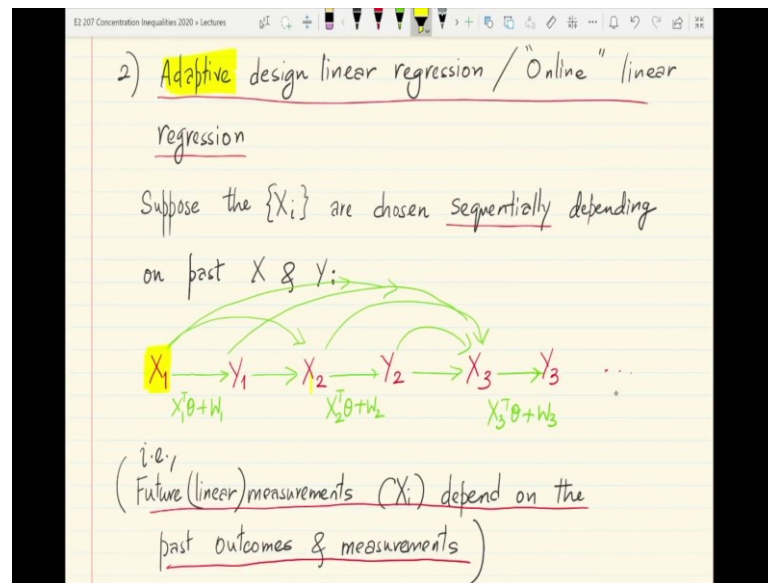
So, the l_2 norm of V raised to half $\hat{\theta}_n - \beta$ is equivalent to the V_n matrix weighted norm of $\hat{\theta}_n - \beta$ by definition being less than C_1 . So, you can multiply throughout by σ . So, this set so, based on your data you could compute $\hat{\theta}_n$ and you could also compute V_n . In fact, V_n is known in advance even before collecting the measurements Y_1 through Y_n because it is a fixed design setting.

But in any case the this set C_n , which is basically an ellipsoidal set in d dimensional space is actually a probability $1 - \delta$. So, what we get from the concentration inequality about is that this particular set C_n is a probability $1 - \delta$ confidence set.

In fact, it is an ellipsoidal confidence it is a you can call it a confidence ellipsoid specifically for the true parameter θ of the linear model ok. In the sense that so, that the sense of the confidence set is that probability that θ belongs to C_n is at least $1 - \delta$ ok.

So, if you had this data $x_1 Y_1$ up to $x_n Y_n$ and if you constructed the set C_n , you can be assured that C_n is a high probability confidence set that will typically contain θ with it ok.

(Refer Slide Time: 11:31)



So, we now next come to a more complicated setting which often arises in practice and this is the setting of not fixed design linear regression, but adaptive design linear regression. It is also called online linear regression. But the setting is basically the as follows. Suppose the measurements X_i are also a random along with the Y_i 's in the sense that suppose the X_i are chosen sequentially and adaptively depending on the previous X 's and Y 's ok.

So, by that what I mean is as follows. So, there is this dependency diagram that we can imagine. So, X_1 is the first linear measurement the first direction along which you want to collect a noisy projection of the unknown parameter θ . So, you get. So, you choose X_1 , let us say using some internal randomness and then you apply X_1 and you get the first measurement as Y_1 which is $X_1^T \theta + w_1$ ok.

So, this arrow indicates that Y_1 depends on X_1 via this function Y_1 is exactly $X_1^T \theta + w_1$ followed by this X_2 is chosen, but X_2 can depend on Y_1 and X_1 . So, that is why X_2 has arrows coming in from X_1 as well as Y_1 , so this arrow is a tangent ok.

So, X_2 depends can depend on X_1 and Y_1 following which Y_2 is obtained only depending on X_2 following the linear model $X_2^T \theta + w_2$. Following Y_2 X_3 is a new direction that can be computed by looking at all the past θ . So, X_1 Y_1 X_2 and Y_2 can be used to decide X_3 and so on and so forth ok. So, the key difference from the

previous fixed design setting is that the future linear measurements that is the X_i 's can actually depend on the outcomes seen so far.

So, upfront there is no way to predict or decide what the trajectory of the measurement X_i 's will be as opposed to the fixed design linear regression setting. So, both the X 's and the Y 's are random here in this sequential or causal manner ok. So, that is what is called adaptive design linear regression.

It often arises in settings where the next measurement is designed or constructed carefully by looking at the history of past measurements ok, when there is the ability to adopt your next sensing decision or measurement decision based on whatever has happened in the past so far.

(Refer Slide Time: 14:21)

In this general case,

- $\hat{\theta}_n$ need not be Gaussian
- $\hat{\theta}_n$ could be biased:
$$\mathbb{E}_{\theta}[\hat{\theta}_n] \neq \theta$$

→ over randomness of X, Y .

Question: Is it still possible to build a high-prob. confidence set for $\hat{\theta}_n$ after n ADAPTIVE measurements X_1^n , for

So, in this more general case it is actually not too hard to see that. In fact, $\hat{\theta}_n$ which was multivariate Gaussian earlier in the fixed design setting need not be Gaussian anymore, its distribution as a vector in d dimensions need not be Gaussian in the general case. There are examples where it is it can be quite far away from the Gaussian distribution in the sense of having tails and so on in for certain kinds of measurement processes.

And moreover $\hat{\theta}_n$ in general can even be a biased estimate. So, in the fixed design setting note that the expected value of $\hat{\theta}_n$ was θ because $\hat{\theta}_n$ was a multivariate

normal centered at θ . But in the adaptive design setting there could be biases because of the data collection mechanism or the measurement mechanism that decides the next X_t depending on the past which would actually introduce which could actually introduce a bias in $\hat{\theta}_n$ ok.

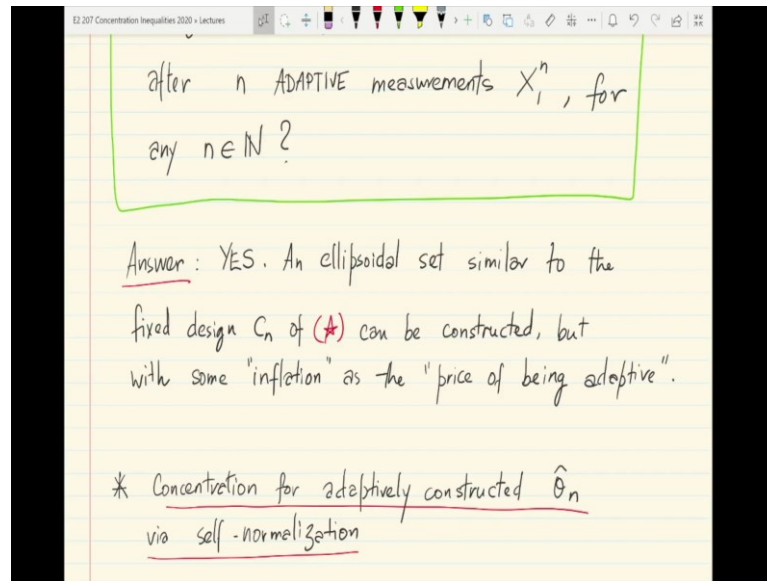
So, expected value of $\hat{\theta}_n$, so, when I say the expectation this is over the randomness or distribution of both the X of the X 's and the Y 's ok. This could this expectation need not be equal to θ anymore ok, but the question still remains.

The interesting question for my inference point of view still remains that is it still possible to build or construct a high probability confidence set for such an adaptively constructed estimate $\hat{\theta}_n$ after n adaptive measurements and at any point in time small n ok. Note that even if you are at a fixed point in time small n the pattern of taking measurements $X_1 X_2$ up to X_n could have a general distribution and be random.

So, one has to worry about that aspect as well ok. So, the same Gaussian theory does not hold because essentially your $\hat{\theta}_n$ which is $V_n^{-1} \sum_{i=1}^n Y_i x_i$ is such that both the V_n part and the Y_i and the x_i 's are all random.

So, it is simply not a fixed linear transformation of a sequence of Gaussian random variables and so on. So, is there any hope is it still possible to build in some sense analogous high probability confident sets for adaptive estimates adaptive sequential estimates?

(Refer Slide Time: 17:08)



And the answer is rather surprisingly yes. In fact, one can argue we will argue now that an ellipsoidal set confidence set very similar to the fixed design confidence set that we called C_n . Let us see in this equation star C_n of star can be constructed, but of course, we have to pay a price for being adaptive and incurring some bias.

So, there will be some inflation that we will have to do to handle the extra bias as the; so, at a high level this is the price of being adaptive over time. So, we will see exactly how much of a price we will roughly have we will have to pay. And so, this is where the idea of a concentration inequalities for adaptively constructed estimates arises via what is called self normalizing inequalities self normalized inequalities.

(Refer Slide Time: 18:52)

via self-normalization

Adaptive setting: X_t depends on $Y_{t-1}, X_{t-1}, Y_{t-2}, X_{t-2}, \dots$

$$\begin{aligned}\hat{\theta}_n &= \left(\sum_{i=1}^n X_i X_i^T \right)^{-1} \sum_{i=1}^n X_i Y_i \\ &= V_n^{-1} \sum_{i=1}^n X_i (X_i^T \theta + w_i) \quad \text{iid } \mathcal{N}(0,1) \\ &= V_n^{-1} \left(V_n \theta + \sum_{i=1}^n w_i X_i \right) \\ &= \theta + V_n^{-1} S_n\end{aligned}$$

So, just to set the stage recall that in the adaptive linear regression setting the next measurement X_t can depend on is some function of the past measurements and sensing actions that is the Y 's and X 's in the past. So, let us start by rewriting $\hat{\theta}_n$ in a slightly different form.

So, recall that $\hat{\theta}_n$ was defined as the summation of i equal to 1 to n $X_i X_i^T$ inverse into the sum of $X_i Y_i$. Note that the X_i 's are now in capital letters because they can be random in the adaptive setting. So, we will still use; we will still call this as V_n , however V_n is a random matrix now.

So, we can write this as V_n inverse and by definition we can expand Y_i as $X_i^T \theta + W_i$ because that is the assumed linear model for the Y_i 's. So, the W_i 's are iid normal $0, 1$. Let us say let us assume we normalize the variance of the noise to 1 in whatever follows and so, this is equal to V_n inverse.

If you expand if you distribute the inner sum you get a V_n times θ by definition + the second sum which is $W_i X_i$. Let us call this sum as S_n ok. And so, this finally, gives you this finally gives you $\theta + V_n$ inverse S_n ok. So, $\hat{\theta}_n$ is exactly equal to the original $\theta + V_n$ inverse S_n . So, θ is the original non random parameter. V_n inverse S_n is a random vector which could have bias because of adaptive data collection.

(Refer Slide Time: 21:13)

$$\begin{aligned} \therefore \|\hat{\theta} - \theta\|_{V_n}^2 &= \|V_n^{-1} S_n\|_{V_n}^2 = S_n^T \cancel{V_n^{-1}} \cancel{V_n} V_n^{-1} S_n \\ &= \|S_n\|_{V_n^{-1}}^2. \end{aligned}$$

* BOTH S_n & V_n are RANDOM quantities,
 but there is a "self-normalization" (S_n grows "linearly" with n ,
 so we can expect $\|S_n\|_{V_n^{-1}}^2$ to be of "constant" size. (V_n grows "linearly" with $n \Rightarrow V_n^{-1}$ decays $\propto 1/n$)
 (as with the fixed design estimate). How?

Nevertheless, if you subtract if you take θ to the other side and find the V_n norm of the V_n weighted matrix norm of $\hat{\theta} - \theta$ then we just get the V_n weighted matrix norm of this $V_n^{-1} S_n$. This by definition is $S_n^T V_n^{-1} V_n^{-1} S_n$ in between.

And then again $V_n^{-1} S_n$ and so, this and this can cancel. So, we finally, get so, its equivalent to the to a certain inverse matrix metered norm of this quantity S_n . Now the important; so, if we want to control $\hat{\theta} - \theta$ let us say in the V_n norm as before as with the fixed design setting this is equivalent to controlling the V_n^{-1} weighted norm of the random vector S_n . S_n itself is the sum of n random vectors each of them being the $W_i X_i$'s.

Now, one thing that is worth noticing here is that both S_n and V_n in the adaptive design setting are random quantities. The former a random matrix in the latter a random vector; the former a random vector and the latter a random matrix, but we notice that there is a kind of self normalization.

The self normalization essentially is that; you know what is I said what do we mean by self normalization? S_n grows let us say it basically linearly this is very roughly linearly with n because each time S_n grows by the addition of a new component $W_i X_i$. And on the other hand V_n grows again linearly with n . So, linearly is in quotes because these

are all rough statements and hence V_n inverse grows slowly decays as $1/n$ roughly ok.

So, S_n grows linearly with n , V_n also grows linearly with n roughly and so, essentially you expect $S_n V_n$ inverse or S_n normalized by V_n inverse or S_n normalized by V_n to be essentially your some kind of constant scale ok. There is a self normalization. So, we can very roughly expect $S_n V_n$ inverse to be of constant size ok.

Could differently sort of you know S_n can grow large, but so, will V_n ok. So, S_n normalized by V_n inverse can probably be expected to not become too large ok just as with the fixed design estimate fixed design estimate. So, in a fixed design setting things were little easier because V_n was a fixed matrix, it was not random. So, it was easy to handle terms involving V_n . So, how is this? So, how do we make this formula ok?

(Refer Slide Time: 25:42)

Theorem (Abbasi-Yadkori et al, 2011)

Fix $\epsilon > 0$, & let $\delta \in (0, 1)$.

$$\mathbb{P}\left[\exists t \in \mathbb{N} : \frac{\|S_t\|^2}{(\epsilon I + V_t)^{-1}} \geq \underbrace{2 \log\left(\frac{1}{\delta}\right)}_{O(1)} + \underbrace{\log\left(\frac{\det(\epsilon I + V_t)}{\epsilon^d}\right)}_{O(d \log t)}\right] \leq \delta.$$

Proof: The Method of mixtures for martingales
(Laplace, Robbins-Siegmund '70, de La Peña, '08).

So, here is a result that we will prove for the adaptive for concentration of the adaptive regression estimate using a new technique which we will call the method of mixture. So, this theorem is essentially inspired by the paper of Abbasi Yadkori et al, it is a paper about linear multi armed bandits from 2011 which essentially establishes the following type of result.

So, let us fix some small epsilon. This is a constant you should think of epsilon is a constant and let δ belong to $(0, 1)$ the probability. So, this theorem says that the probability

when the X_i 's are adaptively designed that exists any time t in the set of all natural numbers at which. So, it is a uniform deviation statement for the entire self normalized process S_t normalized by so, we add a small epsilon $I + V t$ ok.

So, the small epsilon I sort of required because of the. So, it enters through the method of proof, but imagine epsilon I as to be so, small to not significantly affect the scale of $V t$. So, epsilon $I + V t$ inverse is really roughly like $V t$ inverse exceeding a level $2 \log 1$ by $\delta +$ this term which is a log the determinant of this matrix epsilon $I + V t$ divided by epsilon raise to d is at most δ ok.

So, this is what it says the probability that at any time the self normalized quantity S_t inverse weighted by $V t$ square exceeds this particular number is bounded ok. So, to get a sense of the scale of this number, so, if δ is fixed this is basically an order 1 quantity and if δ is fixed this is you can show that this is roughly an order $d \log t$ quantity.

So, this is not too far from the kind of threshold self normalized level in the fixed design case where you again had a order \sqrt{d} term. So, note that here the norm is not being squared. So, you should imagine each of these terms being squared to make a fair comparison. So, order \sqrt{d} would be orders order d if you squared it and this term is order $\sqrt{\log 1}$ by δ ok.

So, the order $\sqrt{\log 1}$ by δ squared could appear here as the $\log 1$ by δ term which is the order 1 term except that we are paying a small price of this extra $\log t$ ok to be uniform in time ok. So, this is the price of sort of; this is sort of in some sense the price of being able to uniformly control an adaptively obtained estimate, a function of a stochastic process and you could you could think of this as not too large a price to pay ok.

So, it just scales logarithmically with the number of time steps here ok. So, but for a small blow up you get exactly uniform deviations.

(Refer Slide Time: 29:34)

Note: $\frac{1}{2} \|S_t\|_{V_t^{-1}}^2 = \max_{\lambda \in \mathbb{R}^d} \left(\lambda^T S_t - \frac{1}{2} \|\lambda\|_{V_t}^2 \right)$

$\forall \lambda \in \mathbb{R}^d$, define

$$M_t^\lambda := \exp \left(\lambda^T S_t - \frac{1}{2} \|\lambda\|_{V_t}^2 \right).$$

CLAIM: $\{M_t^\lambda\}_{t=0}^\infty$ is a martingale, i.e.,

$$\forall t \geq 1: \mathbb{E} [M_t^\lambda | X_1, Y_1, \dots, X_{t-1}, Y_{t-1}] = M_{t-1}^\lambda \text{ a.s.}$$

PROOF OF CLAIM: $X_t \in \sigma(X_1, Y_1, \dots, X_{t-1}, Y_{t-1})$, so

LHS = exp

So, let us see how this kind of result is obtained and the proof is going to expose a very interesting technique called the method of mixtures. So, this is called the method of mixtures for martingale processes which is probably the most interesting take away from this lecture that I would like you to have ok. It is also related very intimately to something called Laplace's method.

So, this was actually developed by Laplace centuries ago and used in modern probability theory by first by Robbins and Siegmund dates back to Robbins and Siegmund in 1970. And there is a survey book by de la Pena of 2008, which actually details a lot of applications of the method of mixtures techniques ok.

So, what is the method of mixtures technique? So, we note first let us start by noting the following that if you; so, the quantity that whose fluctuations we want to control uniformly in time is equivalently the S_t the norm the V_t inverse norm of this random walk type object S_t ok.

So, half of V_t inverse norm of S_t squared. In fact, you can show just by linear algebra that this is the solution to the following variational problem. It is the solution to max over λ in \mathbb{R}^d of $\lambda^T S_t - \frac{1}{2} \|\lambda\|_{V_t}^2$ ok.

So, this is actually an identity that expresses actually a matrix weighted norm as the solution to an optimization problem involving the weighted norm of the inverse matrix

on the other side ok. So, V^t inverse becomes V^t on the other side ok. So, if we transfer our attention to the right hand side here we would like to control the left hand side.

So, equivalently it makes sense to try and control the. So, one sufficient way to do control the left hand side is to be able to control this particular quantity here for every possible value of λ ok that is leading to a to an upper bound on the left hand side ok.

So, towards this is sort of what brings us to the Laplace method or the method of mixtures. So, to this end for every vector λ define let us define the following quantity. Define $M^t(\lambda)$ let this is a random variable which is $e^{\lambda^T S^t - \frac{1}{2} V^t \text{norm square of } \lambda}$ ok right.

So, this $M^t(\lambda)$ if you think of λ as fixed and t evolving in time the it forms a stochastic process $M^1(\lambda)$, $M^2(\lambda)$, $M^3(\lambda)$ and so on. And this is the key claim that makes that allows us to do something useful with the with this process. It turns out that t is the sequence of random variables $M^t(\lambda)$ as t goes from 0 to infinity. By the way if t is 0, this is an M^t sum and so, you just directly get $M^0(\lambda)$ is defined to be 1 triple e .

So in fact, this process over time for a fixed λ turns out to be a martingale ok. More specifically what we mean by this is, that for any time greater than equal to 1 if you take the conditional expectation of $M^t(\lambda)$ given X^1, Y^1 all the previous measurements and the all the previous sensing decisions and the observed measurements then this is equal to $M^{t-1}(\lambda)$ almost surely ok.

(Refer Slide Time: 34:18)

PROOF OF CLAIM: $X_t \in \sigma(X_1, Y_1, \dots, X_{t-1}, Y_{t-1})$, so

$$\text{LHS} = \exp\left(-\frac{1}{2}(\lambda^T X_t)^2\right) \cdot M_{t-1}^\lambda \cdot \mathbb{E}\left[e^{W_t(\lambda^T X_t)} \mid \dots\right]$$

$$= \exp\left(-\frac{1}{2}(\lambda^T X_t)^2\right) \cdot M_{t-1}^\lambda \cdot \exp\left(\frac{1}{2}(\lambda^T X_t)^2\right).$$

$\{\because W_t \stackrel{iid}{\sim} \mathcal{N}(0,1)\}.$

So, what is the proof of this claim? Well, it is the fact that the left hand side so in fact, we note that because of the adaptive sensing X_t is already determined once X_1, Y_1 and all the way up to X_{t-1}, Y_{t-1} are given ok. So, if you have this conditional expectation M_t^λ is the X_t is e raised to $\lambda^T X_t$ - another sum.

So, the left hand side of the expression above in the left hand side you can actually bring out the second term here. You can bring out the term here corresponding to X_t in fact. So, that is the e raised to $-\frac{1}{2} \lambda^T X_t$ the whole square. This is just the last term in $\text{norm } \lambda^T V_t$ the whole square. It comes out because X_t is already measurable with respect to X_1, Y_1 up to X_{t-1}, Y_{t-1} just because of the causal nature of the sensing decision the X_t .

And then you have M_{t-1}^λ that also comes out of the conditional expectation because it is a function of exactly all these things X_1, Y_1 up to X_{t-1}, Y_{t-1} and the only thing that remains inside the conditional expectation is the first term is the last term of $\lambda^T X_t$ which is basically e raised to $W_t \lambda^T X_t$ given everything in the past.

And given everything in the past the only randomness is in this W_t random variable which is an independent standard Gaussian by assumption whose moment generating function is precisely. So, you have all of these terms and the last term the conditional

expectation is exactly the moment generating function of $W^T t$ evaluated at this inner product $\lambda^T X^T t$.

And so, you just directly get this is the moment generating function for Gaussian half $\lambda^T X^T t$ the whole square ok. So, the first term and the third term cancel leaving you with what you require ok. So, this is because W_i are assumed to be iid standard norm ok. So, what does they show? For each fixed λ M_t^λ is actually a martingale process in the sense that it essentially does not change its mean more than average.

(Refer Slide Time: 37:31)

Handwritten notes on a digital notepad:

$\{\because W_i \text{ iid } N(0,1)\}.$

$$\therefore \exp\left(\frac{1}{2} \|S_t\|_{V_t^{-1}}^2\right) = \sup_{\lambda \in \mathbb{R}^d} M_t^\lambda$$

Unfortunately not in general a martingale, so cannot control this using standard results like Azuma's ineq.

So, hence we raise to the quantity that we want to control which is a half $S^T V^{-1} S$ square is actually exactly equal to the supremum of a martingale at time t of an uncountably large collection of martingales ok. So, M_t^λ we have just shown is a martingale for a fixed value of λ and in some sense you know the what we want to control is on the left hand side up to a monotonic transformation of e exponentiating.

And we essentially would like to control the fluctuations of a supremum of many martingales ok. So, unfortunately this supremum of many martingales; so, we know very well there is a lot of martingale theory that tells us about the fluctuations of martingales results like Azuma, Azuma's inequality from martingales and so on, but the maximum of many martingales is unfortunately not in general a martingale.

So, we cannot control this right hand side. So, we cannot control this supremum over many martingales using standard results like let us say Azuma's inequality or its variance ok. So, this leaves us with sort of this we are write this recall where we have not we cannot really proceed at first glance to control this supremum of a really large family of martingales.

(Refer Slide Time: 39:36)

Method of mixtures / Laplace integral approximation

Suppose f is a smooth function on $[a, b] \rightarrow \mathbb{R}$
 with a maximum at $\lambda^* \in (a, b)$.

$$\int_a^b e^{sf(\lambda)} d\lambda \xrightarrow{s \rightarrow \infty} e^{sf(\lambda^*)} \cdot \sqrt{\frac{2\pi}{s \cdot c}}$$

IMPLICATION: "maximization"

depends on smoothness of f about λ^ .*

So, this is exactly where the method of mixtures enters and finds helps us find a way forward or also called; so, you can really view it as an instance of what is called the Laplace integral approximation ok. So, what is this Laplace integral approximation or Laplace integral formula?

So, suppose it is sort of this following phenomenon. Suppose f is a function, you can say f is let us say a smooth function on defined on the interval a cross b ok. So, it is a real to real function with a maximum at an interior point λ^* in a, b ok.

So, f exception argument λ in the interval a, b in the closed interval a, b and its maximum actually occurs at its interior where somewhere in its interior. So, the so, what one can show using basically Taylor approximations up to second order is the following. So, if you integrate the function e raised to S so, fixed S as some real number a large real number e raised to $S f \lambda^* b \lambda^*$ ok.

So, consider integrating $e^{\lambda f}$ over the range a to b as S becomes larger and larger one can argue that this essentially tends to $e^{\lambda f^*}$ times some constant that this constant depends essentially on the smoothness of f about its maximum point λ^* .

It essentially like it is it relates to the Hessian of or the second derivative of f at λ^* . So, the point of this is up to a constant you essentially by integrating you recover $e^{\lambda f^*}$ ok. So, the implication here is that the key implication here is that maximization is roughly equal to approximate integration ok. So, the integral is a convenient approximate way to find the maximum ok in some sense.

(Refer Slide Time: 43:20)

Handwritten notes on a digital notepad:

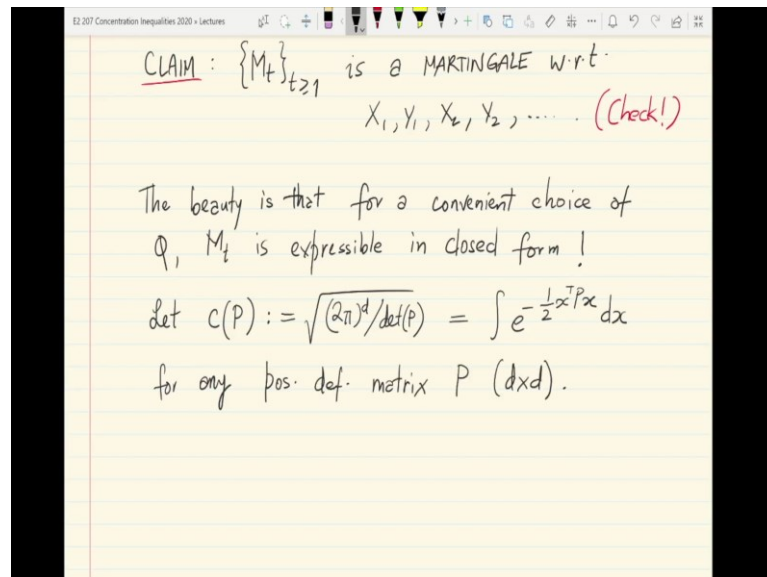
- Top right: $S.C.$ with a red arrow pointing down to "depends on smoothness of f about λ^* ."
- Left side: $\frac{\partial}{\partial \lambda}$
- Center: IMPLICATION : "maximization = approximate integration"
- Below: "We consider the process"
- Equation: $M_t := \mathbb{E}_{\lambda \sim Q} [M_t^\lambda]$, which is a
- Text: "(mixture) martingale for any prob. distribution Q on \mathbb{R}^d ."

And this is very useful in our context because we can consider instead of the sup over λ of $M_t(\lambda)$ what we can do is, we can set up the stochastic process M_t define this as M_t as the expected value over λ drawn from some distribution Q over \mathbb{R}^d of $M_t(\lambda)$ ok.

So, think of so, this is essentially you know what this is doing is basically integrating or mixing across martingales M_t and the convenient fact here is that this is actually a martingale itself because it is the average of a bunch of a martingales ok. So, it retains the martingale property for any probability distribution Q on \mathbb{R}^d ok.

So, no matter how many martingales even an uncountable family of martingales you average you will still get a martingale ok. And since this is a martingale its fluctuations are very easy to control using standard martingale results and indirectly we can also get a handle on controlling its supremum which is what we seek to control finally, ok.

(Refer Slide Time: 44:58)



So, this is the high level idea and the details will be spelt out shortly of how this helps. So, let me just put down this claim that the mixture of all these M_t where λ is independently chosen as a vector from some convenient distribution Q over \mathbb{R}^d forms a martingale with respect to the same filtration as we had earlier with respect to X_1, Y_1, X_2, Y_2 and so on in the same way ok. So, this is something that you can easily check offline ok.

And in fact, the beauty here is that for the convenient choice of that the distribution Q the measure Q there is actually an explicit closed form solution for closed form expression for M_t . M_t is expressible in closed form thanks to properties of the Gaussian measure ok. So, this is remarkable. So, to show this let us define for any positive definite matrix P c of P as $\sqrt{2\pi}^d / \det(P)$. So, P is a d by d positive definite matrix divided by the determinant of P .

So, this is also equivalent to integral of e raised to $-\frac{1}{2} X^T P X$ over \mathbb{R}^d for any positive definite matrix P d by d ok.

(Refer Slide Time: 47:14)

Let's take $Q \equiv \mathcal{N}(0, \varepsilon I_{d_d}), \varepsilon > 0$.

$$M_t = \int_{\mathbb{R}^d} \exp\left(\lambda^T S_t - \frac{1}{2} \|\lambda\|_{V_t}^2\right) \cdot \overset{\text{density of } Q}{f(\lambda)} d\lambda$$

So, recall so, let us chose. In fact, so, let us take the distribution of the distribution Q again as a standard normal. Let us say with mean vector 0 and covariance matrix some epsilon times identity ok, where epsilon is some is this epsilon greater than 0 number. So, think of it epsilon as basically a small number, ok.

So, with this what have is that we can explicitly start writing M_t as the integral of M_t lambda. So, $e^{\lambda^T S_t - \frac{1}{2} \lambda^T V_t \lambda}$. So, with respect to the density of the Q distributions, so, $f(\lambda)$ is a notation for the density of the Q measure which is the multivariate normal with covariance epsilon I .

(Refer Slide Time: 48:28)

$$\begin{aligned} & \mathbb{R}^d \\ & \text{(Check!)} = \frac{c(\varepsilon I + V_t)}{c(\varepsilon I)} \cdot \exp\left(\frac{1}{2} \|S_t\|^2_{(\varepsilon I + V_t)^{-1}}\right) \\ & = \left(\frac{\varepsilon^d}{\det(\varepsilon I + V_t)}\right)^{1/2} \cdot \exp\left(\frac{1}{2} \|S_t\|^2_{(\varepsilon I + V_t)^{-1}}\right). \end{aligned}$$

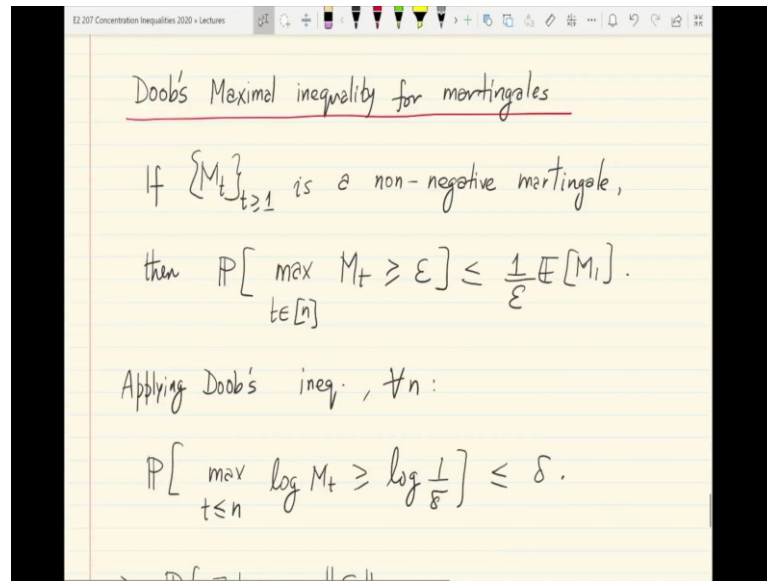
Doob's Maximal inequality for martingales

And then after some straightforward linear algebra which you can verify you can show that this is actually exactly equal to the c of $\varepsilon I + V_t$ divided by C of εI into remarkably e raised to half S_t square $\varepsilon I + V_t$ inverse ok. So, this is exactly so, this exactly turns out to be an explicit function of what we actually wanted to control in the first place which is this S_t inverse weighted by a norm V_t inverse with some extra term in the front.

And this extra term is precisely the determinant of a εI by definition which is ε to the d divided by the determinant of $\varepsilon I + V_t$ whole raised to half into e raised to half the weighted norm of S_t , the self normalized term here ok. And this is basically a martingale ok.

To finish the proof of the result you will appeal to standard result in martingale theory which we will not prove during this course. You can go and look the look up the proof separately where this is called Doob's inequality for Doob's maximal inequality for martingales. So, it is a very well known and its elementary, but very powerful inequality.

(Refer Slide Time: 50:26)



Doob's Maximal inequality for martingales

If $\{M_t\}_{t \geq 1}$ is a non-negative martingale,

then $P\left[\max_{t \in [n]} M_t \geq \epsilon\right] \leq \frac{1}{\epsilon} E[M_1]$.

Applying Doob's ineq., $\forall n$:

$$P\left[\max_{t \leq n} \log M_t \geq \log \frac{1}{\delta}\right] \leq \delta.$$

So, it says basically that if you have a non negative martingale, so, M_t as above actually satisfies this property. If M_t over all time is a non negative martingale then the probability that it is maximum over t going from 1 to n . So, let me just write this precisely as t belonging to n M_t exceeding any level ϵ is at most 1 over ϵ into expected value of M_1 .

This is very similar to Markov's inequality except that it strengthens it by allowing you to replace \max the supremum of this martingale over an interval of time 1 through n by actually replace it with it is only its first element M_1 ok. So, that is what Doob's maximal inequality allows you to do.

So, applying Doob's inequality we get that for any natural number n the probability that the \max over t less than equal to n of $\log M_t$ exceeding $\log 1$ by δ ok. So, you can just exponentiate both sides and then you will get 1 over 1 over δ into expected value of M_1 which is 1. So, you finally, get δ .

(Refer Slide Time: 52:12)

Applying Doob's inequality, for n :

$$P\left[\max_{t \leq n} \log M_t \geq \log \frac{1}{\delta}\right] \leq \delta.$$

$$\Rightarrow P\left[\exists t \leq n : \left\| S_t \right\|_{(\epsilon I + V_t)^{-1}}^2 \geq 2 \log \frac{1}{\delta} + \log \left(\frac{\det(\epsilon I + V_t)}{\epsilon^d} \right) \right] \leq \delta.$$

Can be replaced by ∞ .

And so, this means that you can just turn this around to say that the probability that there exists t less than equal to n for which; so, you can sort of replace you can substitute for $\log M_t$. And you will get the probability that the inverse weighted norm of S_t exceeds twice $\log 1$ by $\delta + \log$ the determinant of $\epsilon I + V_t$ by ϵ to the d .

So, this is so, you basically get the same thing as you get this is the same expression here is at most δ ok. And so, there is only a one last thing remaining which is a technical improvement. So, there is nothing here in this part which depends on the n here. So, this n is just any convenient n . So, you can let actually n tend to infinity. So, you can replace infinity which is a technical procedure that we will not concern ourselves with in this class.

But suffice it suffice to say that this n is really a placeholder n , it can be as large as you want because the rest of the property here a really have nothing to do with this finite number n . So, you can actually take this n all the way to infinity ok by a limiting argument.

And that is what gives you the result that you get this sort of order one term or order $\log 1$ by δ term and we get sort of this order $d \log t$ term as uniform fluctuation for the self normalized quantity S_t matrix waited norm with $\epsilon I + V_t$ the whole inverse which is roughly a V_t inverse normalization of S_t and that concludes this lecture.

Thank you.