STOCHASTIC APPROXIMATION: THEORY AND APPLICATIONS

Dr. Gugan Thope

Department of Computer Science and Engineering

Indian Institute of Science, Bangalore

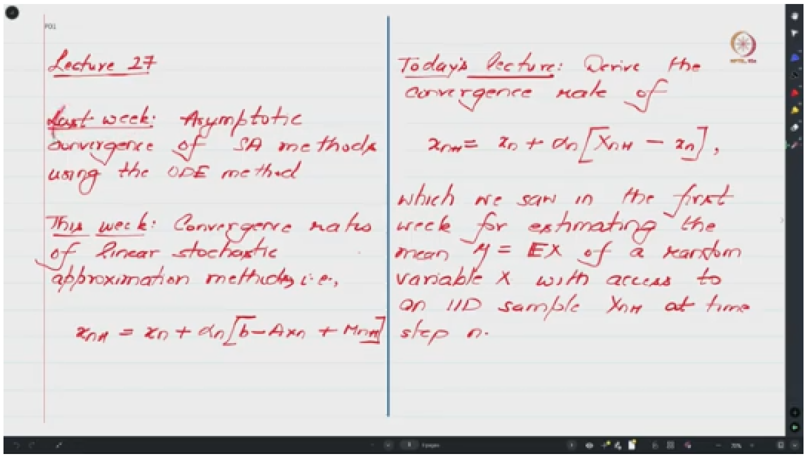
Week 7

Lecture 27

Convergence Rate for Linear Stochastic Approximation - Part 1

Hello and Namaste, everyone. Welcome to week 7, in particular lecture 27 of this NPTEL course on Stochastic Approximation. So, in the previous 2 weeks, we were looking at the asymptotic convergence of stochastic approximation algorithms. So, there was one assumption which we refer to as the stability of the stochastic approximation iterates, which says that almost surely the iterates are bounded. Under this, you know, in some sense a non-trivial assumption, we discussed the asymptotic convergence. In the next week, we will actually discuss ways to remove this assumption. However, in this intermediate class, we will discuss the convergence rates of, you know, simple stochastic approximation algorithms.

So that you get to understand the effect of the step size on how fast you converge, okay? So, let us do a formal summary. So, last week, we looked at asymptotic convergence of stochastic approximation methods using what we called the OD method. This week, we will look at the convergence rates, which means how fast you converge. To keep things simple and to get the key, you know, idea out, we will focus on the simple case of linear stochastic approximation. By that, I mean stochastic approximation algorithms which can be written in the form xn plus 1 equals xn plus alpha n times b minus axn plus mn plus 1.



Lecture 27

Last week: Asymptotic convergence of SA methods using the ODE method

This week: Convergence rates of linear stochastic approximation methods i.e.,

$$x_{n+1} = x_n + \alpha_n[b - Ax_n + M_{n+1}]$$

Today's lecture: Derive the convergence rate of

$$x_{n+1} = x_n + \alpha_n[X_{n+1} - x_n],$$

which we saw in the first week for estimating the mean $\mu = EX$ of a random variable X with access to an IID sample X_{n+1} at time step n.

That is your function h. has the form B minus AX, and because your H function can be written in this fashion, we will refer to this update rule as linear stochastic approximation, right.

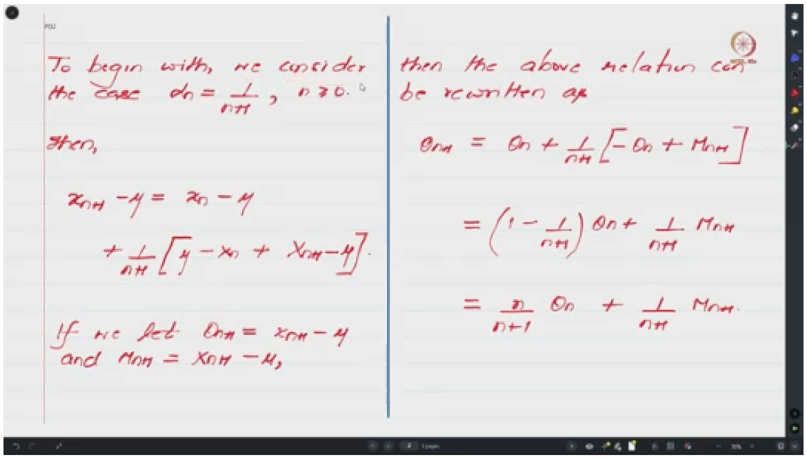
$$x_{n+1} = x_n + \alpha_n\left[b - AX_n + M_{n+1}\right]$$

$$h(x) = b - Ax$$

And in this class, that is in today's class, we will actually keep things even more simple—simpler, that is, we will focus on this very, very basic algorithm used for estimating the mean of a random variable X, which we had discussed during the first week. So, let us recall the problem that we had discussed and let us recall how this algorithm was designed. So, the problem is that we have some random variable X, and its mean is mu equals the expected value of X.

And we have access to IID samples of this random variable X. In particular, at time step n, we have access to the random variable Xn plus 1. Is that okay? That is, at time step 0, we have access to the random variable capital X1. At time step 1, we have access to the random variable X2, and so on. And we had discussed an algorithm of the following form to estimate the mean of the random variable X.

Now, for the simple algorithm, we will discuss how the choice of step size actually affects its convergence rate. To begin with, we will consider the very natural step size choice of alpha n equals 1 over n plus 1. So, I say natural because once you substitute this step size choice over here, you will see that the update rule has the form xn plus 1 equals xn plus 1 over n plus 1 xn plus 1 minus little xn, and by some rearrangement, you can see that this expression actually equals x1 plus dot dot dot all the way up till xn plus 1 over n plus 1. So, you can see that by using a simple inductive argument, you can show that little xn plus 1 is actually the sum of all the random variables divided by n plus 1. So, basically, this expression is the sample mean of all the samples that we have observed so far.



To begin with, we consider the case $\alpha_n = \frac{1}{n+1}$, $n \geq 0$.

Then,

$$x_{n+1} - \mu = x_n - \mu + \frac{1}{n+1}\left[\mu - x_n + X_{n+1} - \mu\right].$$

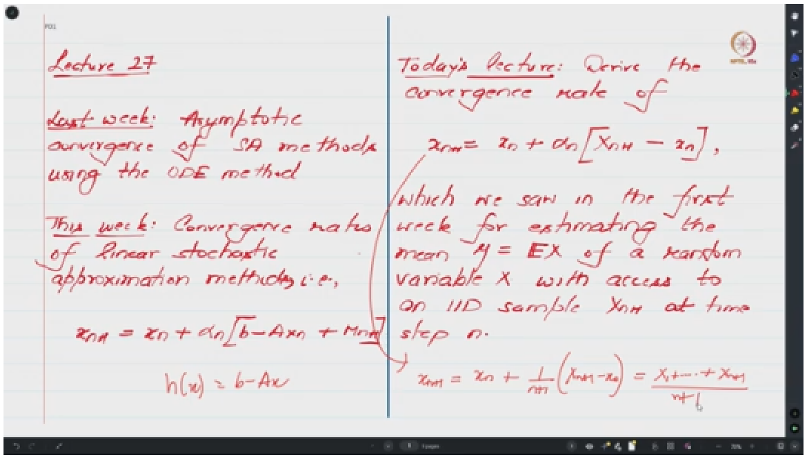
If we let $\theta_{n+1} = x_{n+1} - \mu$ and $M_{n+1} = X_{n+1} - \mu$,

then the above relation can be rewritten as

$$\theta_{n+1} = \theta_n + \frac{1}{n+1}\left[-\theta_n + M_{n+1}\right]$$

$$= \left(1 - \frac{1}{n+1}\right)\theta_n + \frac{1}{n+1} M_{n+1}$$

$$= \frac{n}{n+1}\theta_n + \frac{1}{n+1} M_{n+1}.$$



Lecture 27

Last week: Asymptotic convergence of SA methods using the ODE method

This week: Convergence rates of linear stochastic approximation methods, i.e.,

$$x_{n+1} = x_n + \alpha_n\left[b - Ax_n + M_{n+1}\right]$$

$$h(x) = b - Ax$$

Today's lecture: Derive the convergence rate of

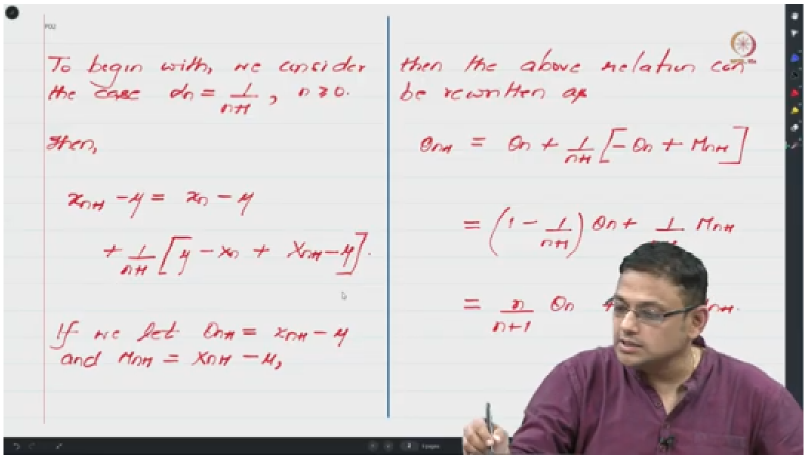
$$x_{n+1} = x_n + \alpha_n\left[X_{n+1} - x_n\right],$$

which we saw in the first week for estimating the mean $\mu = EX$ of a random variable X with access to an IID sample X_{n+1} at time step n.

$$x_{n+1} = x_n + \frac{1}{n+1}\left(X_{n+1} - x_n\right) = \frac{X_1 + \cdots + X_{n+1}}{n+1}$$

Now, given this setup, one can see that since we are taking the sample average, one can view this update rule with alpha n equals 1 over n as the natural thing to do, and one would like to know in this case how fast the algorithm converges. So, what we will do is we will subtract mu from both sides so that we can write the update rule as Xn plus 1 minus mu equals Xn minus mu plus 1 over n plus 1. So, here observe that we have substituted the value of alpha n plus 1, and in the update rule earlier, it was Xn plus 1.

minus little xn. So, what I have done is I have added and subtracted mu from this expression and written mu minus little xn first and capital Xn plus 1 minus mu second, right?



To begin with, we consider the case $\alpha_n = \frac{1}{n+1}$, $n \geq 0$.

Then,

$$x_{n+1} - \mu = x_n - \mu + \frac{1}{n+1}\left[\mu - x_n + X_{n+1} - \mu\right].$$

If we let $\theta_{n+1} = x_{n+1} - \mu$ and $M_{n+1} = X_{n+1} - \mu$,

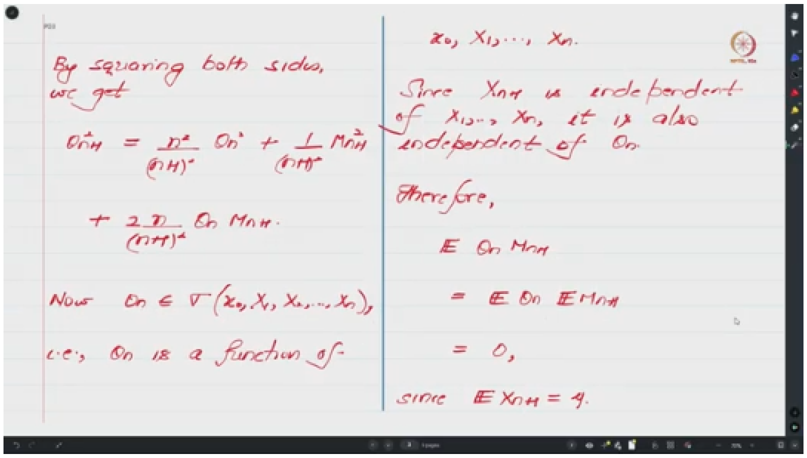
then the above relation can be rewritten as

$$\theta_{n+1} = \theta_n + \frac{1}{n+1}\left[-\theta_n + M_{n+1}\right]$$

$$= \left(1 - \frac{1}{n+1}\right)\theta_n + \frac{1}{n+1} M_{n+1}$$

$$= \frac{n}{n+1}\theta_n + \quad M_{n+1}.$$

So, now if you define θ□$_{+1}$ as X□$_{+1}$ minus μ and M□$_{+1}$ as capital X□$_{+1}$ minus μ, so that we can think of this as the noise, then this update rule can be written as θ□$_{+1}$ equals θ□. So, observe that this expression is θ□, 1 over n plus 1 is as is, and this is minus θ□, and the noise term is M□$_{+1}$. By some rearrangement, one can see that if I combine these θ□ terms together, the two terms can be written as 1 minus 1 over n times θ□ plus 1 over n plus 1 times M□$_{+1}$. And this expression leads to n over n plus 1 times θ□ plus 1 over n plus 1 times M□$_{+1}$. Right now, the goal for us is to find the convergence rate of the squared error, and hence we are going to square both sides. Recall θ□$_{+1}$ is X□$_{+1}$ minus μ, so the error, in some sense, discusses how far our current estimate is from the value we are intending to estimate, which is μ over here, right? And hence, θ□$_{+1}$ squared is the squared error, right?



By squaring both sides, we get

$$\theta_{n+1}^2 = \frac{n^2}{(n+1)^2}\theta_n^2 + \frac{1}{(n+1)^2}M_{n+1}^2 + \frac{2n}{(n+1)^2}\theta_n M_{n+1}.$$

Now $\theta_n \in \sigma(x_0, X_1, X_2, \dots, X_n)$,

i.e., θ_n is a function of $x_0, X_1, \dots, X_n$.

Since X_{n+1} is independent of $X_1, \dots, X_n$, it is also independent of θ_n.

Therefore,

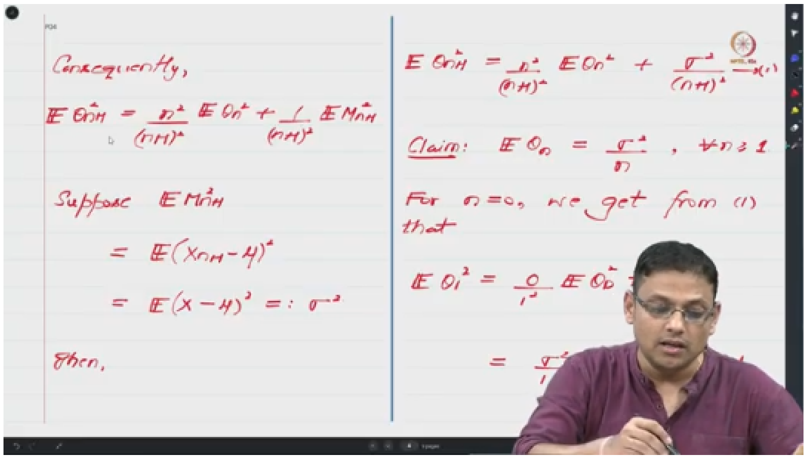
$$\mathbb{E}\,\theta_n M_{n+1} = \mathbb{E}\,\theta_n\,\mathbb{E}M_{n+1} = 0,$$

since $\mathbb{E}X_{n+1} = \mu$.

So, since $\theta_{□+1}$ equals this expression plus this expression, if I square them up, I would end up with n squared over n plus 1 the whole square times $\theta_{□}$ squared plus 1 over n plus 1 the whole square times $M_{□+1}$ squared plus twice the cross terms, which is 2 times n over n plus 1 times $\theta_{□}$ times 1 over n plus 1 of $M_{□+1}$, which will lead to n over n plus 1 the whole square times $\theta_{□}$ and the product of $\theta_{□}$ and $M_{□+1}$, right? Now, it is easy to check that $\theta_{□}$ is actually a function of these quantities, that is, your initial estimate and the samples we have observed, that is, X_1 all the way up to $X_{□}$. More formally, one can show that $\theta_{□}$ is measurable with respect to the sigma field generated by these quantities. And we have presumed that your capital $X_{□+1}$ is actually independent of capital X_1 to capital $X_{□}$.

This is because we have presumed that we have access to this independent sequence of random variables. And since $\theta_{□}$ is a function of these quantities, one can immediately conclude that capital $X_{□+1}$ is also independent of $\theta_{□}$. Hence, if you take the expectation of this cross term, we would end up with the product of the expectations, and if you recall the definition of $M_{□+1}$, right? So, your $X_{□+1}$ is an identically distributed random variable. Hence, if you take its expectation, this will be 0.

Hence, this product of expectations will be 0, right? Which you know follows because your capital Xn plus 1's expectation is actually mu. So, if you take expectation throughout, you can see that this expectation will be there, this expectation will be there, this expectation will be there. However, the expectation of the cross term will vanish, and we will get 0 in its place. Hence, we can see that the expected value of theta n plus 1 squared is n squared over n plus 1 the whole squared times the expected value of theta n squared plus 1 over n plus 1 the whole squared times the expected Mn plus 1 squared.



Consequently,

$\mathbb{E}\,\theta_{n+1}^2 = \frac{n^2}{(n+1)^2}\,\mathbb{E}\,\theta_n^2 + \frac{1}{(n+1)^2}\,\mathbb{E}\,M_{n+1}^2$

Suppose $\mathbb{E}\,M_{n+1}^2$

$= \mathbb{E}(X_{n+1}-\mu)^2$

$= \mathbb{E}(X-\mu)^2 =: \sigma^2$

Then,

$\mathbb{E}\,\theta_{n+1}^2 = \frac{n^2}{(n+1)^2}\,\mathbb{E}\,\theta_n^2 + \frac{\sigma^2}{(n+1)^2} \quad (1)$

Claim: $\mathbb{E}\,\theta_n = \frac{\sigma^2}{n}, \quad \forall n \geq 1$

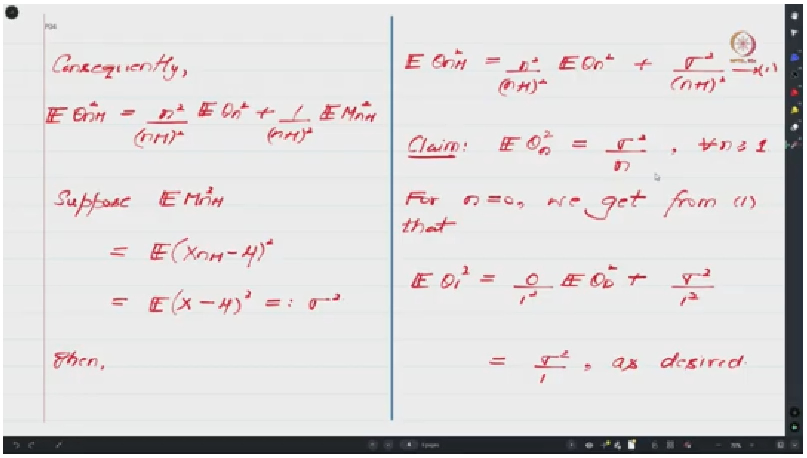
For $n = 0$, we get from (1) that

$\mathbb{E}\,\theta_1^2 = \frac{0}{1^2}\,\mathbb{E}\,\theta_0^2$

$= \frac{\sigma^2}{1}$

So, notice that the cross term is no longer there. And for simplicity, what we will again do is we will presume that the expected value of Mn plus 1 squared, which is the expected value of Xn plus 1 minus mu the whole squared. And since Xn plus 1 is identically distributed as X, this expression equals this. And we will presume that this quantity, which is the variance of X, is denoted by sigma squared. So, since your Xn's are identically distributed, your expected Mn plus 1 squared will actually be expected X minus mu the whole squared, which is the variance of capital X, and we will denote that by sigma squared.

Then this expression will change to expected theta n plus 1 squared equals n squared over n plus 1 the whole squared of expected theta n squared plus sigma squared over n plus 1 the whole squared. Now, we claim that this update rule shows that—so there is a typo here—that the expected theta n squared is actually sigma squared over n. So, let us verify this expression. So, if I substitute n equals 0 in this relation. One can see that the expected theta 1 squared, right? Because n is 0, this term is 0 over 1 squared plus sigma squared over 1.



Consequently,

$$\mathbb{E}\,\theta_{n+1}^2 = \frac{n^2}{(n+1)^2}\,\mathbb{E}\,\theta_n^2 + \frac{1}{(n+1)^2}\,\mathbb{E}\,M_{n+1}^2$$

Suppose $\mathbb{E}\,M_{n+1}^2$

$$= \mathbb{E}(X_{n+1}-\mu)^2$$

$$= \mathbb{E}(X-\mu)^2 =: \sigma^2$$

Then,

$$\mathbb{E}\,\theta_{n+1}^2 = \frac{n^2}{(n+1)^2}\,\mathbb{E}\,\theta_n^2 + \frac{\sigma^2}{(n+1)^2} \quad \text{---}(1)$$

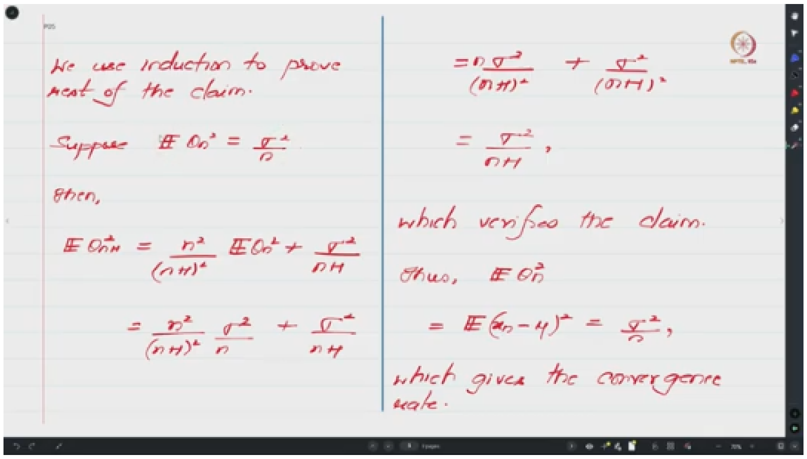
Claim: $\mathbb{E}\,\theta_n^2 = \frac{\sigma^2}{n}$, $\forall n \geq 1$

For $n=0$, we get from (1) that

$$\mathbb{E}\,\theta_1^2 = \frac{0}{1^2}\,\mathbb{E}\,\theta_0^2 + \frac{\sigma^2}{1^2}$$

$$= \frac{\sigma^2}{1}, \text{ as desired.}$$

So, you can see that indeed the expected value of theta n squared equals sigma squared over 1 satisfies this claim over here, right? That is, this claim is true for the case of n equals 1. Now, for proving the general claim, we plan to use induction, and hence as an induction hypothesis, let us presume that the expected theta n squared is sigma squared over n. Then, from equation 1, we know that the expected theta n plus 1 squared is n squared over n plus 1 the whole squared of the expected theta n squared over sigma squared of n plus 1, and this expression by our induction hypothesis is sigma squared over n. Hence, we substitute it over here, and this n and one of the n's will cancel out here, leaving us with n over n plus 1 the whole squared. And here we have 1 over n plus 1 the whole squared.



We use induction to prove rest of the claim.

Suppose $\mathbb{E}\,\theta_n^2 = \frac{\sigma^2}{n}$

Then,

$$\mathbb{E}\,\theta_{n+1}^2 = \frac{n^2}{(n+1)^2}\,\mathbb{E}\,\theta_n^2 + \frac{\sigma^2}{n+1}$$

$$= \frac{n^2}{(n+1)^2}\,\frac{\sigma^2}{n} + \frac{\sigma^2}{n+1}$$

$$= \frac{n\sigma^2}{(n+1)^2} + \frac{\sigma^2}{(n+1)^2}$$

$$= \frac{\sigma^2}{n+1},$$

which verifies the claim.

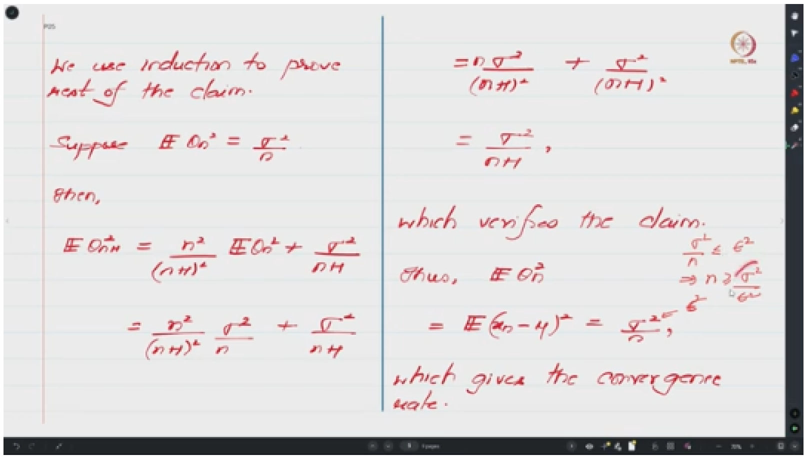
Thus, $\mathbb{E}\,\theta_n^2$

$$= \mathbb{E}(x_n-\mu)^2 = \frac{\sigma^2}{n},$$

which gives the convergence rate.

So, if you add them up, this n and this quantity over here will give us an n plus 1, which will cancel off with one of the factors in the denominator, leaving us with sigma squared over n plus 1. This indeed verifies the claim because the claim was that the expected theta n plus 1 squared is sigma squared over n plus 1. So, now if you try to summarize, one can see that the expected theta n squared is the squared error, and this is sigma squared over n, which tells us the rate at which the squared error converges. So, one can see that the squared error converges at the rate of 1 over n. Which means when you have n samples, the error, at least for this basic algorithm, will at most be sigma squared over n.

If you want this error to be less than epsilon squared, one can see that sigma squared over n less than or equal to epsilon squared implies n should be greater than sigma squared over epsilon squared. Which means that to get an error in your estimate to be less than epsilon squared, one will need roughly 1 over epsilon squared many samples. To get to that error, for example, if your epsilon was 10 to the power of, let us say, 10 to the power minus 1, so that epsilon squared is 10 raised to minus 2, right. So, roughly around 100 samples, or orders of 100 samples, you would need to get the error to be less than 10 raised to minus 2. So, in that sense, one can see or get an estimate of how many samples you would need to get the error to be of order epsilon squared in terms of squared error.



We use induction to prove rest of the claim.

Suppose $\mathbb{E}\,\theta_n^2 = \frac{\sigma^2}{n}$

then,

$$\mathbb{E}\,\theta_{n+1}^2 = \frac{n^2}{(n+1)^2}\,\mathbb{E}\,\theta_n^2 + \frac{\sigma^2}{n+1}$$

$$= \frac{n^2}{(n+1)^2}\,\frac{\sigma^2}{n} + \frac{\sigma^2}{n+1}$$

$$= \frac{n\,\sigma^2}{(n+1)^2} + \frac{\sigma^2}{(n+1)^2}$$

$$= \frac{\sigma^2}{n+1},$$

which verifies the claim.

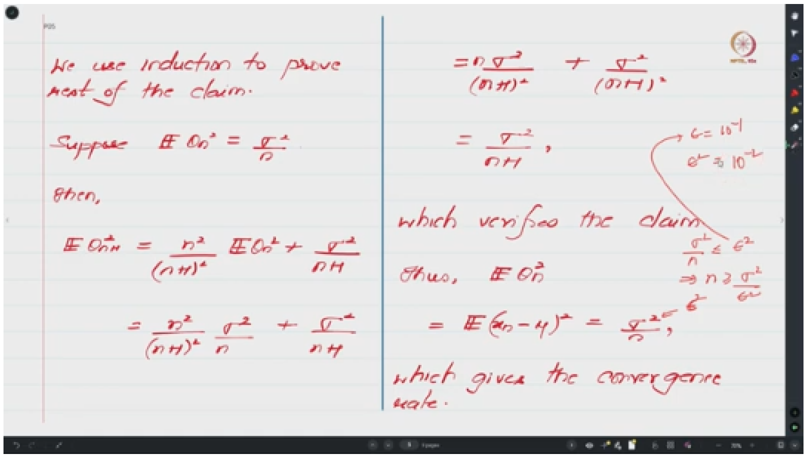
Thus, $\mathbb{E}\,\theta_n^2$

$$= \mathbb{E}(\hat{x}_n - \mu)^2 = \frac{\sigma^2}{n},$$

$\frac{\sigma^2}{n} \le \epsilon^2$

$\Rightarrow n \ge \frac{\sigma^2}{\epsilon^2}$

which gives the convergence rate.



We use induction to prove rest of the claim.

Suppose $\mathbb{E}\,\theta_n^2 = \frac{\sigma^2}{n}$.

Then,

$$\mathbb{E}\,\theta_{n+1}^2 = \frac{n^2}{(n+1)^2}\,\mathbb{E}\,\theta_n^2 + \frac{\sigma^2}{n+1}$$

$$= \frac{n^2}{(n+1)^2}\,\frac{\sigma^2}{n} + \frac{\sigma^2}{n+1}$$

$$= \frac{n\sigma^2}{(n+1)^2} + \frac{\sigma^2}{(n+1)^2}$$

$$= \frac{\sigma^2}{n+1},$$

which verifies the claim

Thus, $\mathbb{E}\,\theta_n^2$

$$= \mathbb{E}(x_n-\mu)^2 = \frac{\sigma^2}{n} \leq \epsilon^2,$$

$\epsilon = 10^{-1}$

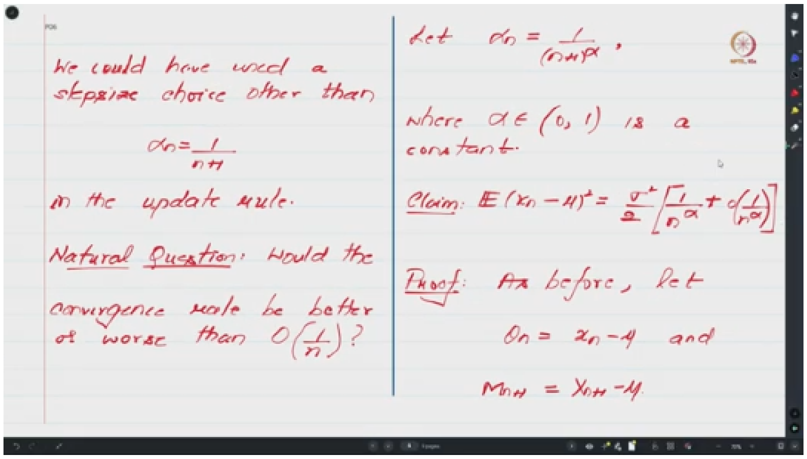
$\epsilon^2 = 10^{-2}$

$\frac{\sigma^2}{n} \leq \epsilon^2$

$\Rightarrow n \geq \frac{\sigma^2}{\epsilon^2}$

which gives the convergence rate.

So, now one can ask, you know, in this update rule, we had used this step size sequence 1 over n plus 1. Why did we use that step size sequence? Maybe if you had used a different step size sequence, let us say alpha between 0 and 1, perhaps we would have gotten a better convergence rate. Now observe that. You know, this quantity over here, right, compared to 1 over n plus 1, right? So if you compare this step size sequence where your alpha is between 0 and 1, that is strictly less than 1, right, then this step size sequence decays lower. This step size sequence decays faster, which means, you know, your step sizes are going to 0 much faster, right? So if you go back to this update rule that we had over here.



We could have used a stepsize choice other than

$$\alpha_n = \frac{1}{n+1}$$

in the update rule.

Natural Question: Would the convergence rate be better or worse than $O\left(\frac{1}{n}\right)$?

Let $\alpha_n = \frac{1}{(n+1)^\alpha}$,

where $\alpha \in (0,1)$ is a constant.

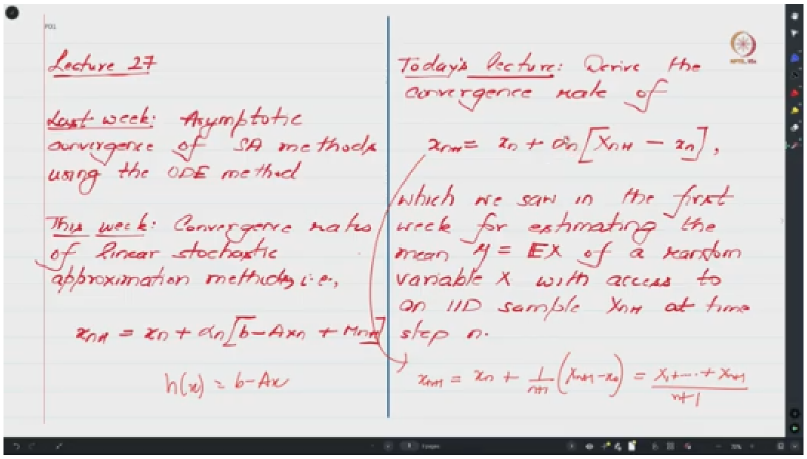
Claim: $\mathbb{E}(x_n-\mu)^2 = \frac{\sigma^2}{2}\left[\frac{1}{n^\alpha} + o\left(\frac{1}{n^\alpha}\right)\right]$

Proof: As before, let

$\theta_n = x_n - \mu$ and

$M_{n+1} = X_{n+1} - \mu$.

So, here, if your step size decays to 0 much faster, you know, we can conclude that the value over here is going to be very, very small. And since this value is going to be very, very small quickly, right, the difference between Xn plus 1 and Xn is going to be, you know, not very different. Like, Xn plus 1 and Xn are not going to be very different if alpha n becomes very, very small. And hence, a natural thought would be, what if we choose alpha n's to be maybe not 1 over n plus 1 but maybe something that decays slower, in which case your Xn plus 1's will actually be much different compared to Xn relative to if you had chosen the step size sequence to be 1 over n plus 1. Right, and often in practice, you know, we may have, you know, computed such estimates by choosing alpha n to be, you know, a constant which never decays, right. So, one may ask, okay, if we choose step sizes which do not decay that fast, do we get a better convergence rate or not? So, that is the question that we are going to start answering.



Lecture 27

Last week: Asymptotic convergence of SA methods using the ODE method

This week: Convergence rates of linear stochastic approximation methods i.e.,

$$x_{n+1} = x_n + \alpha_n[b - Ax_n + M_{n+1}]$$

$$h(x) = b - Ax$$

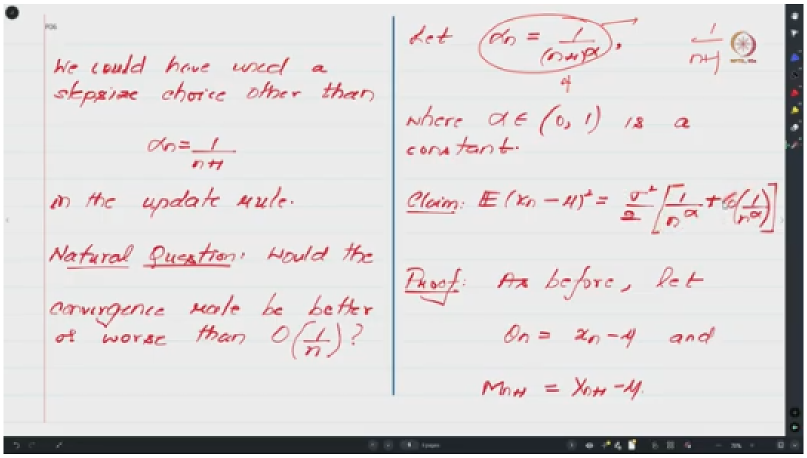
Today's lecture: Derive the convergence rate of

$$x_{n+1} = x_n + \alpha_n[X_{n+1} - x_n],$$

which we saw in the first week for estimating the mean $\mu = EX$ of a random variable X with access to an IID sample X_{n+1} at time step n.

$$x_{n+1} = x_n + \frac{1}{n+1}(X_{n+1} - x_n) = \frac{X_1 + \cdots + X_{n+1}}{n+1}$$

We are partially answering that question in this lecture, and in the next lecture, we will completely answer that question. Okay, and so that there is no suspense, let me already tell you what we are trying to prove. We are going to show that when the step size sequence is of the form 1 over n plus 1 to the power alpha, then your squared error is of the form. Sigma square over 2 times 1 over n to the power alpha plus some term which is little o of 1 over n to the power alpha, okay. So, this means that this expression decays faster than 1 over n to the power alpha. In other words, the term that decays slowly in this expression or the dominant expression is actually 1 over n to the power alpha.



We could have used a stepsize choice other than

$$\alpha_n = \frac{1}{n+1}$$

in the update rule.

Natural Question: Would the convergence rate be better or worse than $O\left(\frac{1}{n}\right)$?

Let $\boxed{\alpha_n = \frac{1}{(n+1)^\alpha}}$, $\frac{1}{n+1}$

where $\alpha \in (0, 1)$ is a constant.

Claim: $\mathbb{E}(x_n - \mu)^2 = \frac{\sigma^2}{2}\left[\frac{1}{n^\alpha} + o\left(\frac{1}{n^\alpha}\right)\right]$

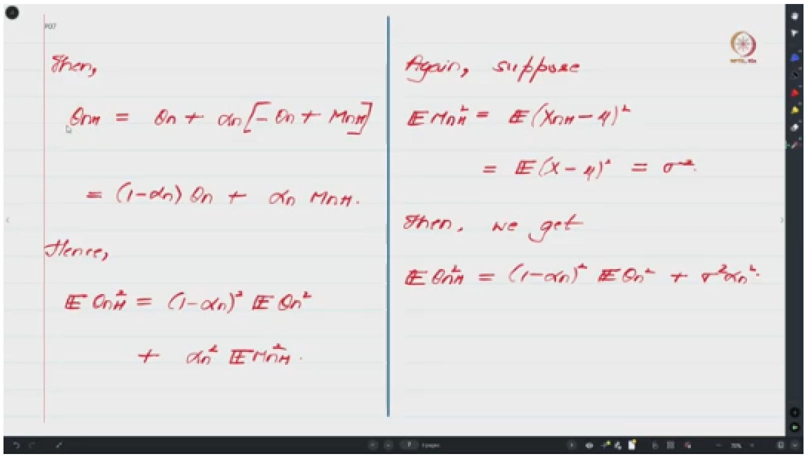
Proof: As before, let

$\theta_n = x_n - \mu$ and

$M_{n+1} = X_{n+1} - \mu$.

So, what this tells us is that if you choose such a step size, then the squared error is actually decaying at a slower rate compared to what had happened when we had chosen the step size sequence 1 over n plus 1. Recall that then the error actually decreased at the rate 1 over n, but here we are saying that the error will actually decay only at the rate of 1 over n to the power alpha, which suggests that choosing this step size is actually bad from the perspective of this convergence rate. So, from the perspective of convergence rate, one needs to actually choose a step size of this form. So, now we are going to derive this result. As I said a few minutes back, we are going to build an intermediate expression in today's class, and we will obtain a bound on this intermediate expression in the next class.

So, let us go back to this derivation. Right, we have Xn plus 1 is Xn plus alpha n times, you know, capital Xn plus 1 minus little xn, and keeping that in mind, let us again define theta n to be Xn minus mu and capital Mn plus 1 to be capital Xn plus 1 minus mu. Then again, the update rule that we will have would be theta n plus 1 equals theta n plus alpha n minus theta n plus mn plus 1. Recall that our update rule was xn plus 1 equals xn plus alpha n. Here we had capital Xn plus 1 minus little xn. So, again by adding and subtracting mu to this expression, one can see that we would have this update rule, and again what we will do is we will collect the common terms over here. So, that we end up with 1 minus alpha n times theta n plus alpha n times mn plus 1.



Then,

$$\theta_{n+1} = \theta_n + \alpha_n[-\theta_n + M_{n+1}]$$

$$= (1-\alpha_n)\theta_n + \alpha_n M_{n+1}.$$

Hence,

$$\mathbb{E}\,\theta_{n+1}^2 = (1-\alpha_n)^2\,\mathbb{E}\,\theta_n^2 + \alpha_n^2\,\mathbb{E}\,M_{n+1}^2.$$

Again, suppose

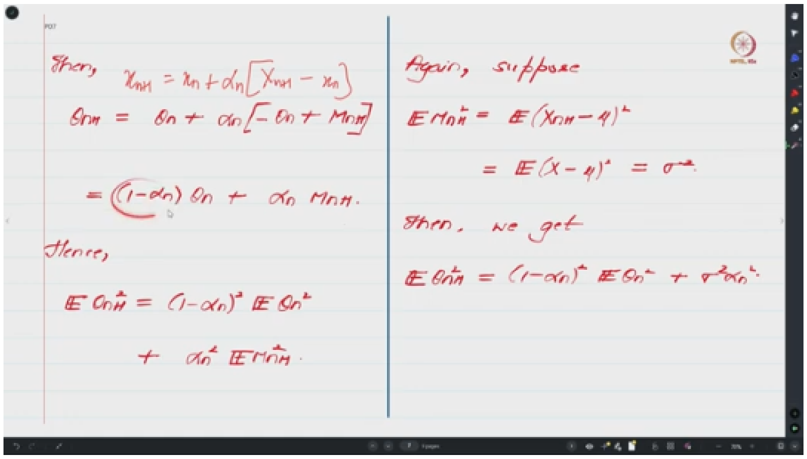
$$\mathbb{E}\,M_{n+1}^2 = \mathbb{E}(X_{n+1}-\mu)^2$$

$$= \mathbb{E}(X-\mu)^2 = \sigma^2.$$

Then, we get

$$\mathbb{E}\,\theta_{n+1}^2 = (1-\alpha_n)^2\,\mathbb{E}\,\theta_n^2 + \sigma^2\alpha_n^2.$$

So, when we had chosen alpha n equals 1 over n plus 1, this was the expression that led little n over little n plus 1. However, in our current derivation, this alpha n does not have that nice form, and hence we will stick to this expression. So, let us again square this expression and take expectation. So, expected value of theta n plus 1 square is what we have over here.



Then, $x_{n+1} = x_n + \alpha_n[X_{n+1} - x_n]$

$$\theta_{n+1} = \theta_n + \alpha_n[-\theta_n + M_{n+1}]$$

$$= (1-\alpha_n)\theta_n + \alpha_n M_{n+1}.$$

Hence,

$$\mathbb{E}\,\theta_{n+1}^2 = (1-\alpha_n)^2\,\mathbb{E}\,\theta_n^2 + \alpha_n^2\,\mathbb{E}\,M_{n+1}^2.$$

Again, suppose

$$\mathbb{E}\,M_{n+1}^2 = \mathbb{E}(X_{n+1}-\mu)^2$$

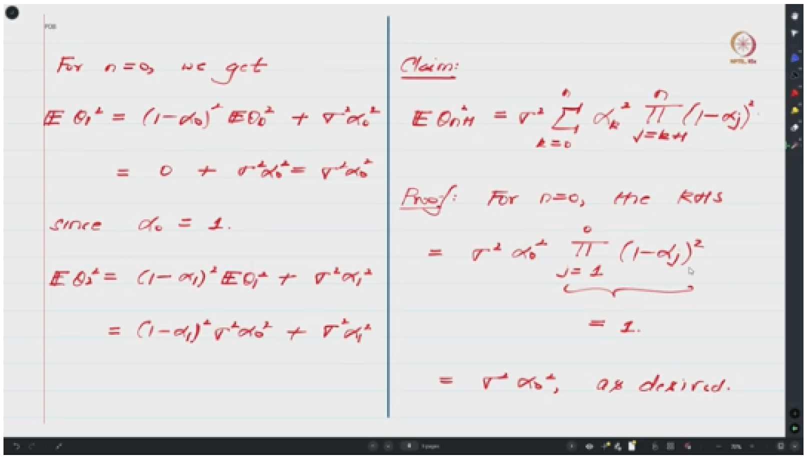
$$= \mathbb{E}(X-\mu)^2 = \sigma^2.$$

Then, we get

$$\mathbb{E}\,\theta_{n+1}^2 = (1-\alpha_n)^2\,\mathbb{E}\,\theta_n^2 + \sigma^2\alpha_n^2.$$

Then the square of this expression is 1 minus alpha n squared times the expected value of theta n squared plus alpha n squared times the expected value of m n plus 1 squared. So, we would again you know, have ended up with a cross term. But since the expected value of Mn plus 1 is 0, that cross term vanishes, right? And hence, we only have, you know, the square of the first term and the square of the second term.

And again, let us assume that the MNs have a common second moment; in other words, let us presume that the XN plus 1 are identically distributed so that this expectation is equal to the expected value of capital X minus mu the whole square, and let us say we denote this by sigma squared, right? And this expression then can be expressed in the following way, which is that the expected value of theta n plus 1 squared is 1 minus alpha n squared times the expected theta n squared plus sigma squared alpha n squared, right? Now, again, let us do some, you know, derivation of some expressions for small values of n and, based on those derivations, let us try to come up with a conjecture which we will prove using induction, right. So, for n equals 0, right, if you go back over here, if you substitute n equals 0, one can see that the expected theta 1 squared is 1 minus alpha 0 the whole squared times the expected theta 0 squared plus sigma squared alpha 0 squared.



For $n=0$ we get

$$\mathbb{E}\,\theta_1^2 = (1-\alpha_0)^2\,\mathbb{E}\theta_0^2 + \sigma^2\alpha_0^2$$

$$= 0 + \sigma^2\alpha_0^2 = \sigma^2\alpha_0^2$$

since $\alpha_0 = 1$.

$$\mathbb{E}\,\theta_2^2 = (1-\alpha_1)^2\,\mathbb{E}\theta_1^2 + \sigma^2\alpha_1^2$$

$$= (1-\alpha_1)^2\sigma^2\alpha_0^2 + \sigma^2\alpha_1^2$$

Claim:

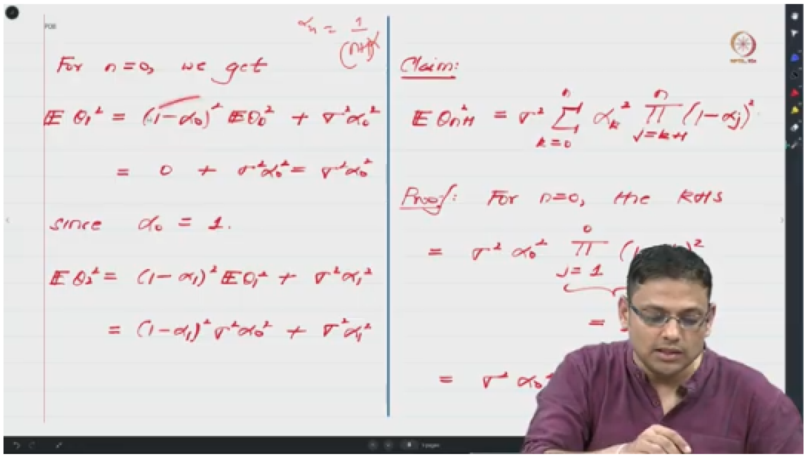
$$\mathbb{E}\,\theta_{n+1}^2 = \sigma^2\sum_{k=0}^{n}\alpha_k^2\prod_{j=k+1}^{n}(1-\alpha_j)^2$$

Proof: For $n=0$, the RHS

$$= \sigma^2\alpha_0^2\underbrace{\prod_{j=1}^{0}(1-\alpha_j)^2}_{=1.}$$

$$= \sigma^2\alpha_0^2,$$ as desired.

Right, and since your alpha, you know, your step size sequences are of the form 1 over n plus 1 to the power alpha, where alpha is some number between 0 and 1, if I substitute n equals 0 here, I hope you agree that alpha 0 is 1, and because alpha 0 is 1, this expression over here becomes 0, and hence I have written 0 over here, and the second expression is actually sigma squared, but you know, I want to retain this form so that I can use it in my analysis. So, this is sigma squared alpha 0 squared, and hence this expression is sigma squared alpha 0 squared. Similarly, if I substitute n equals 1 over here, I would end up with 1 minus alpha 1 squared times the expected value of theta 1 squared and sigma squared alpha 1 squared. So, that is precisely what I have written here.



$\alpha_n = \frac{1}{(n+1)}$

For $n=0$, we get

$\mathbb{E}\theta_1^2 = (1-\alpha_0)^2\,\mathbb{E}\theta_0^2 + \sigma^2\alpha_0^2$

$= 0 + \sigma^2\alpha_0^2 = \sigma^2\alpha_0^2$

since $\alpha_0 = 1$.

$\mathbb{E}\theta_2^2 = (1-\alpha_1)^2\,\mathbb{E}\theta_1^2 + \sigma^2\alpha_1^2$

$= (1-\alpha_1)^2\sigma^2\alpha_0^2 + \sigma^2\alpha_1^2$

Claim:

$\mathbb{E}\theta_{n+1}^2 = \sigma^2\sum_{k=0}^{n}\alpha_k^2\prod_{j=k+1}^{n}(1-\alpha_j)^2$

Proof: For $n=0$, the RHS

$= \sigma^2\alpha_0^2\prod_{j=1}^{0}(\;\;)^2$

$=$

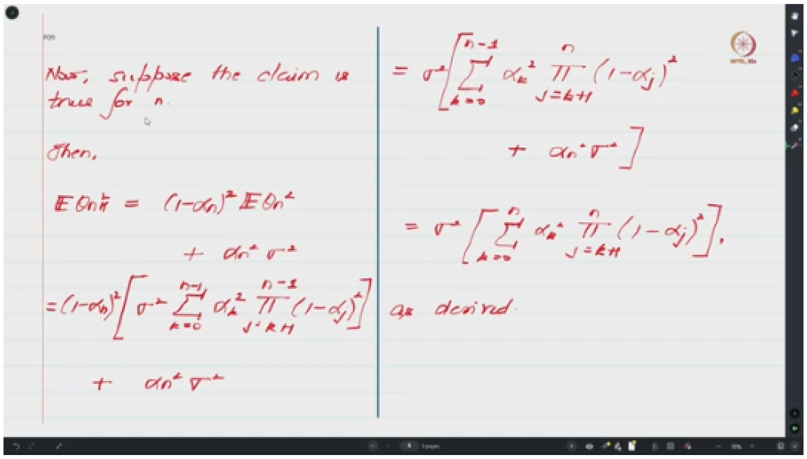
$= \sigma^2\alpha_0$

This is 1 minus alpha on the whole square expected theta 1 square plus sigma square alpha 1 square. Now, here I am going to substitute this expression over here, so that we end up with 1 minus alpha 1 the whole square times sigma square alpha 0 square plus sigma square alpha 1 square, right. So, keeping this expression and this expression in mind, we conjecture or claim that expected theta n plus 1 square is actually 1. So, you can observe that the sigma square terms are common. So, we can pull it out, right?

And one can see that for n equals 2, you had two terms, right? And for n equals 1, there was a single term accordingly for n plus 1. We start with 0 to n so that there are n plus 1 many terms, and the conjecture is that the summation is of terms of the form alpha k square times the product of j equals k plus 1 to n 1 minus alpha j the whole square. So, this is the conjecture, and we are going to verify this conjecture using induction. So, if I substitute n equals 0, this right-hand side will be sigma square, and this n will be 0 here.

So, there will be only one term to add. So, I will now substitute k equals 0. So, that I end up with 0. alpha 0 square, and here I have product from j equals k plus 1 to n, so n is 0, so this is what I have, and j will become 1 to 0 1 minus alpha j square. So, whenever you have a product where the lower index is larger than the upper index, we interpret such products as being taking the value 1. Is that okay? Because such products are, you know, in some sense, technically not possible, hence we will interpret this product as 1, right? And once we interpret this product as 1, we see that this expression is actually sigma 0 alpha square, which is exactly what we had verified here, okay.

So now what we are going to do is we are going to presume that the claim is actually true for some arbitrary value of n and then check, you know, if the expression for n plus 1 follows the pattern that we have indicated. Now, the expected value of theta n plus 1 squared from our earlier expression equals 1 minus alpha n squared times the expected value of theta n squared plus alpha n squared sigma squared. So now what we are going to do is we are going to take this expression and, you know, substitute our induction hypothesis. So, wherever we had this expected theta n squared, we replace it by sigma squared plus the sum that goes from 0 to n minus 1. This is n minus 1 because we have n here, and similarly here also we have n minus 1 because we have n here and we have this expression.



Now, suppose the claim is true for n.

Then,

$$\mathbb{E}\theta_{n+1}^2 = (1-\alpha_n)^2\mathbb{E}\theta_n^2 + \alpha_n^2\sigma^2$$

$$= (1-\alpha_n)^2\left[\sigma^2 \sum_{k=0}^{n-1} \alpha_k^2 \prod_{j=k+1}^{n-1}(1-\alpha_j)^2\right] + \alpha_n^2\sigma^2$$

$$= \sigma^2\left[\sum_{k=0}^{n-1} \alpha_k^2 \prod_{j=k+1}^{n}(1-\alpha_j)^2 + \alpha_n^2\sigma^2\right]$$

$$= \sigma^2\left[\sum_{k=0}^{n} \alpha_k^2 \prod_{j=k+1}^{n}(1-\alpha_j)^2\right],$$

as desired.

And this alpha n squared sigma squared we retain as before. Now, because this 1 minus alpha n squared is outside, we can take it inside, and every expression here can now be pre-multiplied by 1 minus alpha n squared so that I can increase the upper index of this product from n minus 1 to n. And hence this whole expression will become sigma squared, okay, which is over here times this summation, which is k equals 0 to n minus 1 times alpha k squared this product, where observe that the upper index has now changed. This and alpha n squared times sigma squared, okay. So there is a typo here since I have pulled this sigma squared out common; there will not be any sigma squared right. And this whole expression again by interpreting the product of, you know, starting from a higher index and moving to a lower index as 1, one can see that this 1 over here can be written as product J equals n plus 1, okay. j equals n plus 1 to n of 1 minus alpha j squared.

So, if you interpret this as 1, I can, you know, replace it back over here, and one can see that this whole summation can be jointly written as summation k equals 0 to n alpha k squared product going from j equals k plus 1 to n 1 minus alpha j squared.

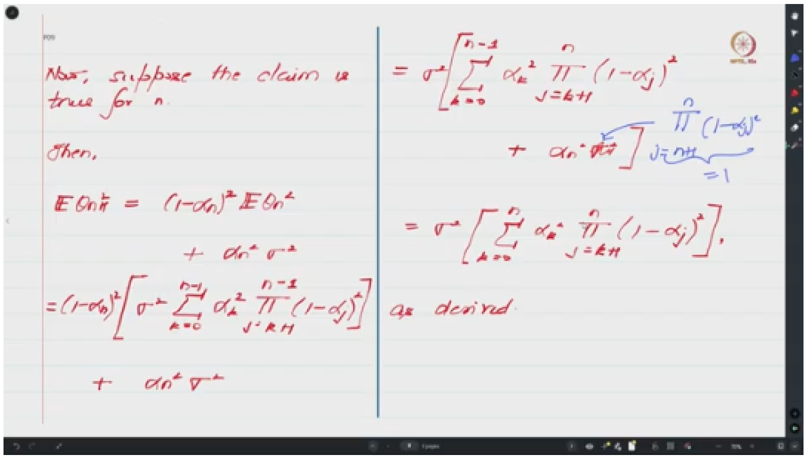
$$E\theta_{n+1}^2 = \left(1-\alpha_n\right)^2 E\theta_n^2 + \alpha_n^2\sigma^2$$

$$= \left(1-\alpha_n\right)^2\left[\sigma^2\sum_{k=0}^{n-1}\alpha_k^2\prod_{j=k+1}^{n-1}(1-\alpha)^2\right] + \alpha_n^2\sigma^2$$

$$= \sigma^2\left[\sum_{k=0}^{n-1}\alpha_k^2\prod_{j=k+1}^{n}\left(1-\alpha_j\right)^2 + \alpha_n^2\sigma^2\right]$$

$$= \sigma^2\left[\sum_{k=0}^{n}\alpha_k^2\prod_{j=k+1}^{n}\left(1-\alpha_j\right)^2\right]$$

So, this finishes the verification of this claim, and this brings us to the end of today's class. In the next class, what we are going to do is we are going to obtain a bound on this expression so that we can conclude this result, in particular, show that the convergence rate actually is lower when you choose a slowly decaying step size. We actually require a step size of the form 1 over n plus 1 to ensure that the convergence rate is indeed the squared error decreases at the rate of 1 over n. This is something that we will prove in the next class. Until then, goodbye and thank you.



Now, suppose the claim is true for n.

Then,

$$\mathbb{E}\theta_{n+1}^2 = (1-\alpha_n)^2\mathbb{E}\theta_n^2 + \alpha_n^2\sigma^2$$

$$= (1-\alpha_n)^2\left[\sigma^2\sum_{k=0}^{n-1}\alpha_k^2\prod_{j=k+1}^{n-1}(1-\alpha_j)^2\right] + \alpha_n^2\sigma^2$$

$$= \sigma^2\left[\sum_{k=0}^{n-1}\alpha_k^2\prod_{j=k+1}^{n}(1-\alpha_j)^2 + \alpha_n^2 \underbrace{\prod_{j=n+1}^{n}(1-\alpha_j)^2}_{=1}\right]$$

$$= \sigma^2\left[\sum_{k=0}^{n}\alpha_k^2\prod_{j=k+1}^{n}(1-\alpha_j)^2\right].$$

as desired

Bye.