Distributed Optimization and Machine Learning

Prof. Mayank Baranwal

Systems and Control Engineering

Indian Institute of Technology Bombay

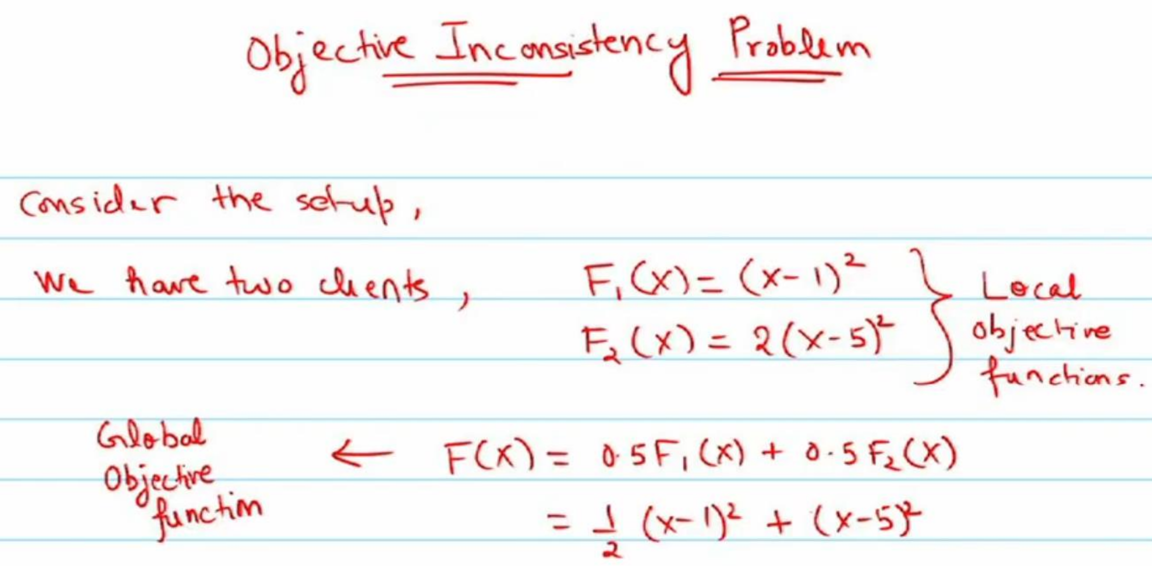
Week-12

Lecture 46: Objective Inconsistency Problem

So, besides this computational heterogeneity there is also this objective inconsistency problem that we had a glimpse of in the last lecture right where we had 2 different 2 different functions or 2 different clients optimizing their own functions. So, there is something called objective in consistency problem. So, as I said let us say there I mean we have just two clients. So, consider the setup. So, we have two clients. So, it is the same example that we looked at in the last class.

and the local objective function are for the first client f 1 x is x minus 1 whole square and f 2 x is 2 times x minus 5 whole square. So, these are local objective functions ok and because I mean these are deterministic functions. So, the global objective function I mean it is not data driven thing. So, it would simply be 0.

5 of f 1 x plus 0.5 of f 2 x you can just consider pi's to be 0.5 0.5 right ok. So, f of x turns out to be half.



Objective Inconsistency Problem

Consider the setup,

We have two clients,

$$\left.\begin{aligned} F_1(x) &= (x-1)^2 \\ F_2(x) &= 2(x-5)^2 \end{aligned}\right\} \text{Local objective functions.}$$

Global Objective function ←

$$F(x) = 0.5F_1(x) + 0.5F_2(x)$$

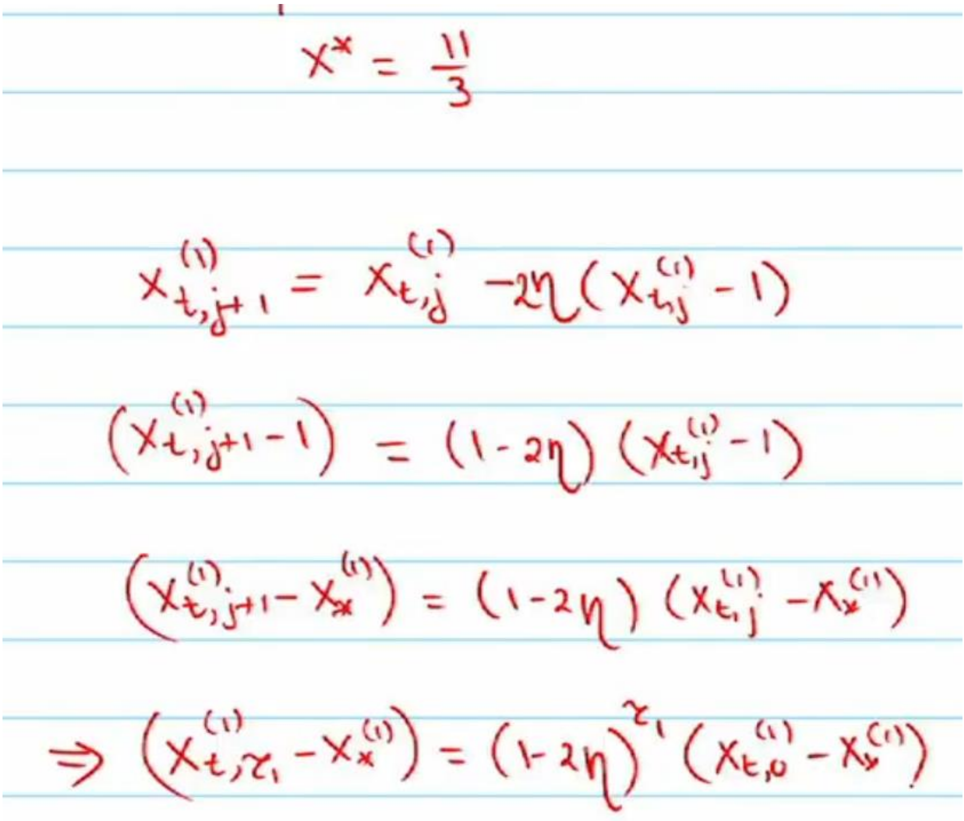
$$= \frac{1}{2}(x-1)^2 + (x-5)^2$$

So, this is your global objective function. f of x turns out to be half x minus 1 whole square plus x minus 5 whole square and what is x star that minimize f of x 11 by 3 right.

So, x star is going to be 11 by 3 x star is the optimal for this capital f like the global objective function. So, one thing that we looked at was suppose there are too many local updates. at both like at both the client ends.

So, then f 1 x the first client would send a value 1, the second client would send the value 5 because these are like optimal solutions or the optimizers for the local objective functions right. And if I take the average of those that is going to be 3 which is going to be different from 11 over 3. So, there is this objective inconsistency problem right. So, how do the step sizes or let us say the number of local updates they affect this particular like they have this I mean this they essentially result in this kind of inconsistency and we are going to quantify that for this particular example ok. So, let us see how each step of this particular thing looks like in the context of this particular example.

So, we have the two agents. So, what would be the? So, we are going to be assuming that each agent runs tau i number of like tau 1 and tau 2 number of local updates. So, for agent 1 what is what does this look like? So, what would be the output? What would be the equation like update rule or update equation? it could be the gradient x t j minus step size times the gradient of f 1 of f 1 x and the gradient is 2 times x minus 1. So, minus y so ok, is this clear. So, I can also write this as x t j plus 1 1.



$$x^* = \frac{11}{3}$$

$$x^{(1)}_{t,j+1} = x^{(1)}_{t,j} - 2\eta\left(x^{(1)}_{t,j} - 1\right)$$

$$\left(x^{(1)}_{t,j+1} - 1\right) = (1-2\eta)\left(x^{(1)}_{t,j} - 1\right)$$

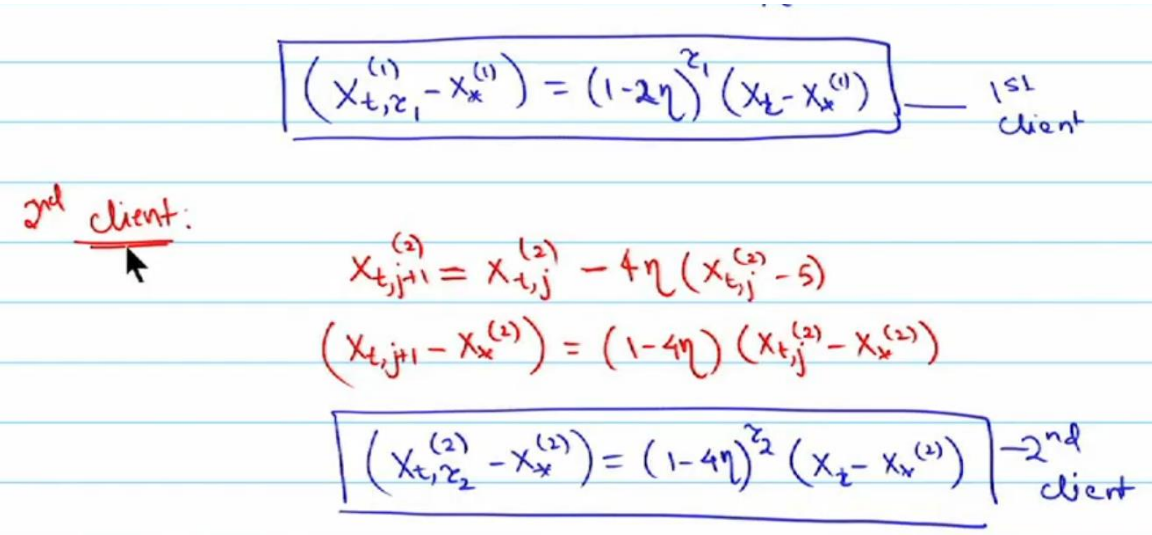
$$\left(x^{(1)}_{t,j+1} - x^{(1)}_{*}\right) = (1-2\eta)\left(x^{(1)}_{t,j} - x^{(1)}_{*}\right)$$

$$\Rightarrow \left(x^{(1)}_{t,\tau_1} - x^{(1)}_{*}\right) = (1-2\eta)^{\tau_1}\left(x^{(1)}_{t,0} - x^{(1)}_{*}\right)$$

So, I am subtracting the optimal which is x, in this case x 1 star is essentially 1 right, is going to be 1 minus 2 root of 1. subtracting minus 1 from it this is what it it would look like. Because the reason I wanted to do this was because this is the x star the optimal for the first agent right is 1 minus 2 eta of x tj 1 minus x star of 1. and this implies if I run this for tau i number of updates. So, x t tau 1 minus x star of 1 this would be 1 minus 2 eta raise to the tau 1 x t 0 1 minus x star.

So, this by the way this xt is the same as the one that is communicated by the central server at the beginning of the tth round right, tth communication round. So, we can also write this as the same as writing simply xt because every agent is going to get at the 0th round or basically 0th update every agent will have this thing. So, this is nothing but xt. So, effectively what we have here is. x t tau 1 1 minus x star 1 is 1 minus 2 times this is for the first agent first client.

So, let us see what if we repeat the same step. So, this is for the first client. Let us see what happens when we do the similar process for this or similar analysis for the second client. For the second client, t j plus 1 2 is equal to x t j 2 minus Now what is the derivative of 4 times this thing? The gradient is essentially 4 times x minus 5. So, it will be minus 4 e to x t into j 2 minus 5 ok.



$$\boxed{\left(x^{(1)}_{t,\tau_1} - x^{(1)}_*\right) = (1-2\eta)^{\tau_1}\left(x_t - x^{(1)}_*\right)} \text{ — 1st client}$$

2nd client:

$$x^{(2)}_{t,j+1} = x^{(2)}_{t,j} - 4\eta\left(x^{(2)}_{t,j} - 5\right)$$

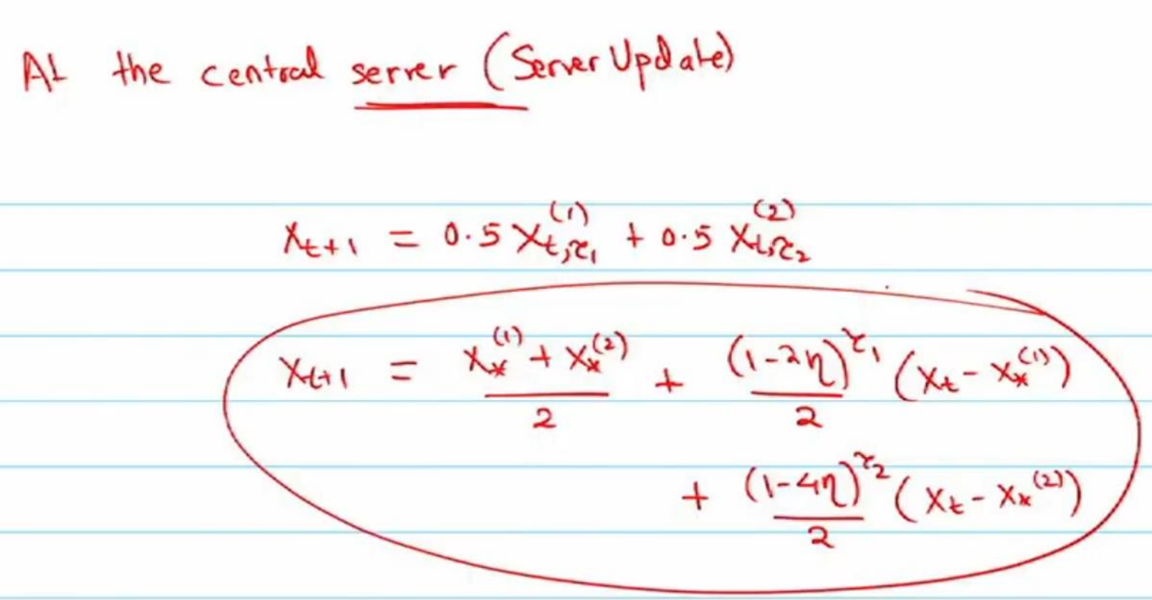
$$\left(x_{t,j+1} - x^{(2)}_*\right) = (1-4\eta)\left(x^{(2)}_{t,j} - x^{(2)}_*\right)$$

$$\boxed{\left(x^{(2)}_{t,\tau_2} - x^{(2)}_*\right) = (1-4\eta)^{\tau_2}\left(x_t - x^{(2)}_*\right)} \text{ — 2nd client}$$

And again if I subtract, so here x star 2 would be 5 right. So, if I subtract x star this is 1 minus 4 times theta x t into 2 minus x star 2 and using a similar analysis you can show that x t tau 2 if you have tau 2 number of updates minus x star 2 this is 1 minus 4 times theta raised to the power tau 2. times xt minus extra 2, this is your second event ok. Everyone with me on this so far? So, what what is now xt plus 1 which is basically the update from the central server. So, it is going to be receiving xt tau 1 and xt tau 2 from the two clients and then we will basically take 0.5, 0.5 of both and then essentially form the xt plus 1 right. So, what happens at the central server? This is server update essentially. So, we are looking at what the server update looks like. and this would be x t plus 1 is essentially 0.5 of x t tau 1 1 plus 0.

5 of x t tau 2 because it is as I said it is not data driven thing. So, you allocate equal sort of I mean it is a deterministic function. So, basically you have 50 50 percent weightage on both the clients and this is what x t plus 1 would look like for the at the server side which is going to be. So, we if I use this particular thing. So, x t tau 2 essentially is x star t x star 2 plus this term.

Likewise x t tau 1 is x star 1 plus this particular term, right. So, if I take half and 0.5, 0.5 of this, so it is essentially x star 1 1 plus x star 2 by 2, ok, plus then you have additional terms, right. And the additional terms are, just want to make sure I do not make, so these 1 minus 2 times eta.



At the central server (Server Update)

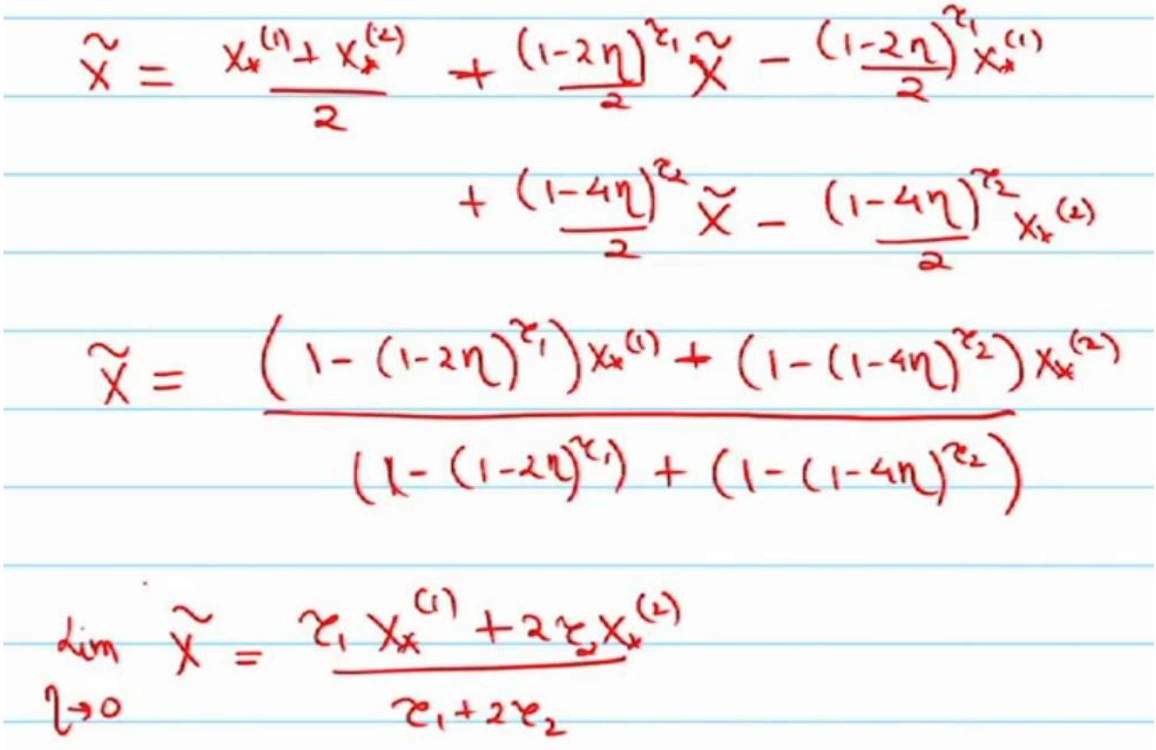
$$x_{t+1} = 0.5\, x^{(1)}_{t,\tau_1} + 0.5\, x^{(2)}_{t,\tau_2}$$

$$\boxed{x_{t+1} = \frac{x_*^{(1)} + x_*^{(2)}}{2} + \frac{(1-2\eta)^{\tau_1}}{2}\left(x_t - x_*^{(1)}\right) + \frac{(1-4\eta)^{\tau_2}}{2}\left(x_t - x_*^{(2)}\right)}$$

Now, we are trying to characterize the solution. So, this is x t plus 1 in terms of x t right. So, this is your x t plus 1 is specified in terms of x t and we are now trying to characterize the solutions of this particular equation ok. So, this is your update at the server side and for the server side to converge to some x let us say x bar or something x tilde. So, what would happen let us say.

So, let us analyze solutions with this. So, if the x let us say x at the server side this has converges some x tilde that means x t plus 1 is same as x t as x tilde right. So, this is saying that here some x tilde is equal to 1 plus x star 2 by 2 plus 1 minus 2 eta tau 1 by 2 x tilde minus 1 minus 2 eta tau 1 by 2 x star 1 plus 1 minus 4 eta tau 2 by 2 x tilde. let us start ok. And this essentially if you recollect basically collect all the terms on the left hand side this turns out to be 1 minus this whole thing divided by 1 minus ok.

This is what the solution looks like x tilde. I mean ideally you would want the solution to converge to. So, you would want to choose your tau 1 and tau 2 and your learning rates. So, that this solution converges to your 11 by 3 right, but let us see what happens. So, if you look at this particular solution and if you apply L'Hopital's rule and in consider the limit when eta goes to 0.



$$\tilde{x} = \frac{x_*^{(1)} + x_*^{(2)}}{2} + \frac{(1-2\eta)^{\tau_1}}{2}\tilde{x} - \frac{(1-2\eta)^{\tau_1}}{2}x_*^{(1)}$$

$$+ \frac{(1-4\eta)^{\tau_2}}{2}\tilde{x} - \frac{(1-4\eta)^{\tau_2}}{2}x_*^{(2)}$$

$$\tilde{x} = \frac{\left(1-(1-2\eta)^{\tau_1}\right)x_*^{(1)} + \left(1-(1-4\eta)^{\tau_2}\right)x_*^{(2)}}{\left(1-(1-2\eta)^{\tau_1}\right) + \left(1-(1-4\eta)^{\tau_2}\right)}$$

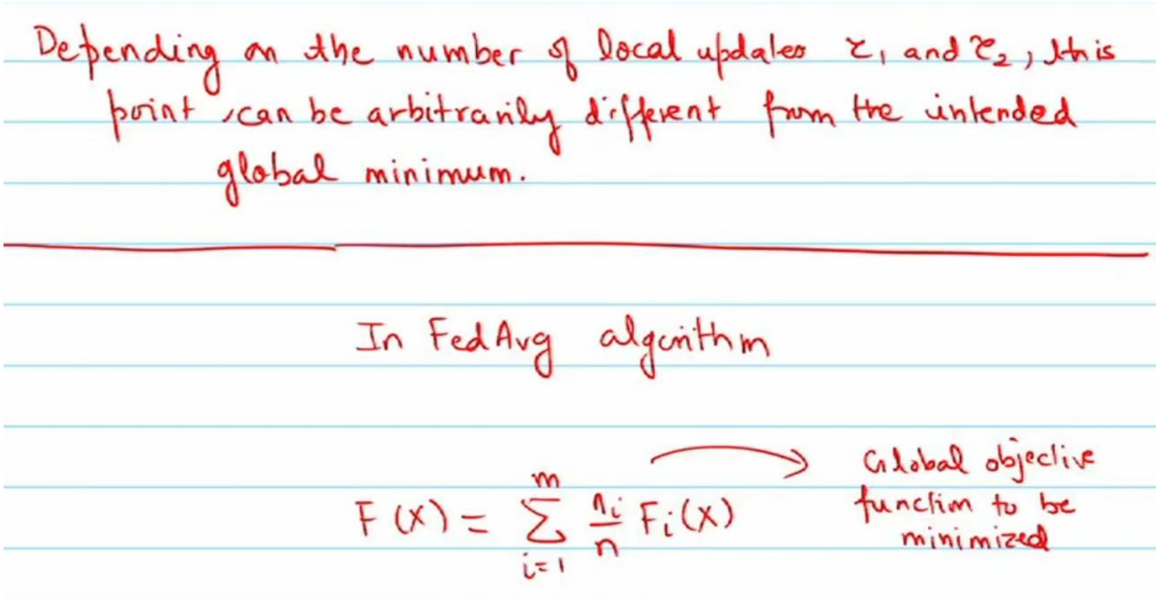
$$\lim_{\eta \to 0} \tilde{x} = \frac{\tau_1 x_*^{(1)} + 2\tau_2 x_*^{(2)}}{\tau_1 + 2\tau_2}$$

So, limit eta goes to 0 your x tilde basically turns you can show that this turns out to and tau 1 x star 1 plus 2 tau 2 star 2 divided by tau 1 plus 2 tau 2. So, you would have to apply the L'Hopital rule ah when once you make once you take eta goes ah eta very small apply the L'Hopital rule and this is what you would get in terms of the solution. And you can see if tau 1 is equal to tau 2 the solution that you get is 11 by 3 right and which is what you want. But if you let us say fix this total number of local updates or if you choose tau 1 and tau 2 slightly differently you would get an entirely different solution ok. So, this is this is the this is called the objective inconsistency problem essentially you are trying to optimize a function global objective function which is 0.

5 of f 1 x and plus 0.5 of f 2 x, but because of the number of local updates that you are making at each client you may be getting come like a different answer right then what you expect. So, depending on the number of local updates tau 1 and tau 2 . This can be arbitrarily different from the intended problem intended So, again you can see that the number of local updates plays a significant role in terms of what solution you converge to right. And so, how do you account for this in typical thread averaging? So, how do you

make sure that there is no objective inconsistency? So, maybe you would have to choose a different weighting scheme.

instead of making 0.5 0.5 some like depending on something depending on the number of local updates or something else you would have to choose a different weighting scheme. So, that you do not basically you end up converting to the intended global minimum ok. So, we were looking at the global or the objective and consistency problem right. So, in typical fair So, what is the intended objective function f of x? It is summation. So, there are m clients i 1 through m.



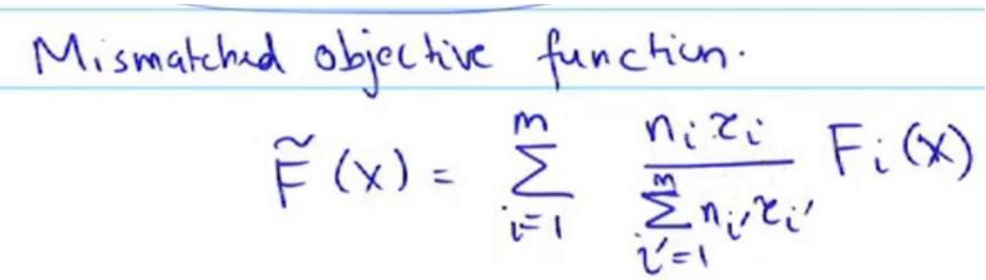
Depending on the number of local updates τ_1 and τ_2, this point can be arbitrarily different from the intended global minimum.

In FedAvg algorithm

$$F(x) = \sum_{i=1}^{m} \frac{n_i}{n} F_i(x) \longrightarrow$$

Global objective function to be minimized

So, this is what we want to minimize right. So, this is your global objective function to be minimized But as you as you saw in the previous example depending on the values of local updates and also I mean that example was about deterministic optimization, but if you have let us say stochastic optimization with n 1, n 2 and so on and data point like these the actual the real objective function that the fed averaging algorithm optimizes based on these parameter values. So, in fact it turns out that the mismatched objective. objective function that the algorithm ends up minimizing.



Mismatched objective function.

$$\tilde{F}(x) = \sum_{i=1}^{m} \frac{n_i \tau_i}{\sum_{i'=1}^{m} n_{i'} \tau_{i'}} F_i(x)$$

Let me call this f tilde of x. So, this turns out to be 1 through m ni tau i i prime 1 through m ni prime tau i prime So, instead of minimizing this intended global objective function, the fed like if you run the fed averaging algorithm with tau i number of local updates, each client having n i number of data points. So, this is what it ends up minimizing. And you can see that if I choose tau i's to be the same, the number of local updates to be the same, then this is the same I mean essentially the mismatch object objective matches your intended objective. But the moment you start having the same number of local updates, slow clients would take more time and so on right. So, there are other issues with that, but this is this I mean you can really see that the mismatch objective, I mean this is the objective that is getting optimized here. Thank you very much.