The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

PROFESSOR: Today what we want to do is discuss various approaches that you might want to take towards trying to understand stochastic systems. In particular, how is it that we might model or simulate a stochastic system?

Now, we will kind of continue our discussion of the master equation from last time. Hopefully now you've kind of thought about it a bit more in the context of the reading. And we'll discuss kind of what it means to be using the master equation and how to formulate the master equation for more complicated situations, for example, when you have more than one chemical species.

And then we'll talk about the idea of this Gillespie method, which is an exact way to simulate stochastic systems, and it's both exact and computationally tractable as compared to what you might call various naive methods. And the Gillespie method is really sort of qualitatively different from the master equation because in the master equation, you're looking at the evolution of probability distributions across the system, whereas the Gillespie method is really a way to generate individual stochastic trajectories.

So if you start with somehow similar initial conditions, then you can actually get-- you can get, for example, the probability distributions from the Gillespie method by running many individual trajectories. But it's kind of conceptually rather different because of this notion of whether you think about probabilities or you're thinking about individual instantiations of some stochastic trajectory. So we'll try to make sense of when you might want to use one or the other.

And then finally we'll talk about this Fokker-Planck approximation, which, as the reading indicated, for intermediate ends, it's useful to make this kind of continuous

approximation, and then you can get a lot of intuition from your knowledge about diffusion on effective [INAUDIBLE] landscapes.

Are there any questions about this or administrative things before we get going? I just want to remind you that the midterm is indeed next Thursday evening, 7-9 PM. If you have a problem with that time, then you should have emailed [? Sarab. ?] And if you haven't emailed him yet, you should do it right now. And-- yes.

All right. So let's think about the master equation a little bit more. Now before what we did is we thought about the simplest possible case of the master equation, which is, if you just have something being created at a constant rate and then being degraded at a rate that's proportional to the number of that chemical species. And I'm going to be using the nomenclature that's a little bit closer to what was in your reading, just for, hopefully, clarity. And I think that some of my choices from last lecture were maybe unfortunate.

So here, this is, for example, m would be the number of mRNA, for example, in the cell. This is the rate of creation of the mRNA, and then the rate of degradation of the mRNA. So m is the number of mRNA. And if we want understand gene expression, we might include an equation for the protein, so we might have some p dot, where some Kp.

Now-- oh, sorry. Again, I always do this. All right. So we're going to have this be an n dot. So now n is going to be the number of the protein.

Now this really is kind of the simplest possible model that you might write down for gene expression that includes the mRNA and the protein. So there's no autoregulation of any sort. It's just that the mRNA is involved in increasing the protein, but then we have degradation of the protein as well.

So what we want to do is kind of try to understand how to formulate the master equation here. But then also, we want to make sure that we understand what the master equation is actually telling us and how it might be used.

So first of all, in this model, I want to know is there, in principle, protein bursts? So

before we talked about the fact that in-- at least in [? Sunny's ?] paper that we read-- they could observe protein bursts, at least in those experiments in e Coli. Question is, should this model somehow exhibit protein bursts, and why or why not? I just want to see where we are on this.

I think this is something that, depending on how you interpret the question, you might decide the answer is yes or no. But I'm curious-- I think it's worth discussing what the implications are here. And the relevant part of this is going to be the discussion afterwards, so I'd say don't worry too much about what you think right now. But I'm just curious. This model, does it include, somehow, protein bursts? Ready? Three, two, one.

OK. So we got-- I'd say at least a majority of people are saying no. But then some people are saying yes. So can somebody volunteer why or why not? Yes?

AUDIENCE: I think the difference is if we're-- are we using this in a continuous fashion or are we using this in a discrete fashion [INAUDIBLE].

PROFESSOR: Yeah. OK. All right. All right. So he's answered both possible sides of the argument. And the point here is that if you just simulate this from the standpoint-- certainly, for example, this continuous, this discrete-- so if you just simulate this as a deterministic pair of differential equations, then will there be bursts? No. Because everything is well-behaved here.

On the other hand, if we go and we do a full Gillespie simulation of this pair of equations, then in the proper parameter regime, we actually will get protein bursts, which is, in some ways, weird, that depending upon the framework that you're going to be analyzing this in, you can get qualitatively different behaviors for things.

But there's a sense here that the deterministic, continuous evolution of these quantities would be the average over many of these stochastic trajectories, and the stochastic ones do have bursts, but if you average over many, many of them, then you end up getting some well-behaved pair of equations.

So we'll kind of try to make sense of this more later on. But I think this just highlights

that you can get really qualitatively different behaviors for the same set of equations depending upon what you're looking at.

And these protein bursts can be dramatic events, right, where the protein number pops up by a lot. So this really, then, if you look at the individual trajectories here, they would look very different whether you were doing kind of a stochastic treatment or the deterministic one.

Can somebody remind us the situation in which we get protein bursts in the stochastic model? In particular, will we always get these discrete protein bursts? Or what determines the size of a protein burst? Yes.

AUDIENCE: Does it have to do with the lag time between when the mRNA is created [INAUDIBLE]?

PROFESSOR: OK. Right. So there's a lag time between the time that mRNA is created, and then the next thing would be--

AUDIENCE: When the protein [INAUDIBLE].

PROFESSOR: When the protein is [? totaled-- ?] OK. So there are multiple time scales, right? So after an mRNA is created, and that's through this process here-- so now out pops an mRNA-- now there are multiple time scales. There's the time scale for mRNA degradation. That goes as 1 over gamma m. There's a time scale for protein degradation after a protein is made. That goes as 1 over gamma p. But then there's also a time scale associated with kind of the rate of protein production from each of those mRNAs, and that's determined by Kp. So we get big protein bursts if what? What determines the size of these protein bursts? Yes.

AUDIENCE: [INAUDIBLE]

PROFESSOR: Right. It's always confusing. We talk about times. But in particular, we have protein bursts in the stochastic situation if we do a stochastic simulation. And that's in the regime if Kp, the rate of protein synthesis from the mRNA is somehow much larger than this gamma m. Have I screwed up? Yes.

AUDIENCE: So this is also-- in the sense of being different from the deterministic equations, we probably also want the total number of mRNAs [INAUDIBLE]. Is that sort of--

PROFESSOR: Right. Yeah, I think that it-- and the question of what mRNA number you need. I mean, it depends on what you mean by protein bursts. I would say that so long as this is true, what that means is that each mRNA will, indeed, kind of lead to a burst of proteins being made, where the burst is, again, geometrically distributed with some-- now there's another question, which is, are those protein bursts kind of large compared to the steady state protein concentration? And that's going to depend upon Km and gamma p as well. Is that--

AUDIENCE: Yeah. So I guess [INAUDIBLE] which is, I guess it would also depend on how big [INAUDIBLE].

PROFESSOR: All right, well-- and you're saying time resolution in terms of just measuring--

AUDIENCE: Yeah. [INAUDIBLE]

PROFESSOR: Yeah. Well, OK, but right now we're kind of imagining that we live in this perfect world where we know at every moment of time exactly how many of everything there is. So in some ways we haven't really said anything yet about time resolution. We're assuming that our time resolution and our number resolution is actually perfect.

But still, depending upon the regime that you're in, the protein numbers could look something like-- so if you look at the protein number, which is defined as this n as a function of time, then in one regime, you're going to see where it's kind of low. You get a big burst and then it kind of comes down, and then a big burst, and then it kind of comes down, and burst, and it kind of comes down, right? So this is in the regime where you have infrequent mRNAs being produced, and then large size bursts from each mRNA. And then you kind of get this effective degradation or dilution of the protein numbers over time. And this distribution, if you take a histogram of it, is what?

AUDIENCE: [INAUDIBLE]

PROFESSOR: Right. So I'm imagining that we look at this for a long period of time. And then we come over here and we histogram it. So now we come over here, we turn to the left, we say number has a function of-- this is number n. The frequency that we observe, some number of proteins. So frequency. And this is going to do something.

So what about-- it may not be a beautiful drawing, but you're supposed to know the answer. I'm trying to review things for you because I hear that you have a big exam coming up, and I want to make sure that--

Gamma. It's a gamma, right? So this is what we learned earlier. So this is a gamma distribution. And you should know what this gamma distribution looks like. In particular, there are these two parameters that describe this gamma distribution as a function of underlying parameters in the model.

AUDIENCE: [INAUDIBLE]

PROFESSOR: Maybe. I don't want to get too much into this because, well, on Thursday we spent a long time talking about it. Once we get going, we'll spend another long time taling about it again. But you should review your notes from Thursday before the exam.

So this thing is gamma distributed. And if we looked at the mRNA number as a function of time and we did a histogram of that, the mRNA distribution would be what? It's poisson. So it's important to remember that just because I tell you that a protein number is gamma distributed, that doesn't immediately tell you exactly what you should be expecting for the distribution of, say, the number of protein as a function of time.

I mean, there are many different things I could plot over here that would all kind of come down to a gamma distribution over here. So it's important to kind of keep in mind the different representations that you might want to think about the data.

So what we want to do now is we want to think a little bit more about this master equation in the context of if we're going to divide it up into these states. Now I would

say that any time that you are asked to write down the master equation for something-- so now how many equations will the master equation-- I say master equation, but there is really more than one, maybe. So how many equations will be involved in the master equation kind of description of this model?

Infinitely many. But there were infinitely many already when we had just one, when we just had the mRNA distribution. Well, you know, infinite times infinite is still infinite. So long as it's a countably infinite number. But yeah, but it's still infinite, always. All right.

So what we want to do is divide up the states. So when somebody asks you for-- the equation's describing how those probabilities are going to vary, really what we're interested in is some derivative with respect to time of some probabilities described by m,n. We want to know the derivative with respect to time for all m,n's. So that's why there are infinite number, because m goes in one direction, n goes in another. Lots of them, OK?

Now it's always tempting to just write down this derivative and then just write down the equation. If you do that, that's fine, but I would recommend that in general what you do is you try to write a little chart out to keep track of what directions things can go. So for example, here we have the probability of being the m,n state. Now there's going to be ways of going here. And this is going to be going probability of being an m plus 1,n.

What I'm going to do is I'm going to give you just a couple minutes. And in two minutes, I want you to try to write down as many of the rates, the f's and n's that correspond to all these transitions. You may not be able to get through all of them, but if you don't try to figure out some of them, then you're going to have trouble doing it at a later date.

Do you understand what I'm asking you to do? So next to each one of these arrows, you should write something. So I'll give you two minutes to kind of do your best of writing these things down.

All right. Why don't we reconvene, and we'll see how we are? So this is very similar to what we did on Thursday. We have to remember that m's are the mRNAs, and this is what we solved before, where it's just a long row.

Now first of all, the mRNA distributions and the rates, do they depend on the protein numbers? No. So what that mean about, say, this arrow as compared to the arrow that would be down here? It's going to be the same, because n does not appear in that equation describing mRNA. If we had autoregulation of some sort, then it would. So let's go through.

All right. What we're going to do is we're going to do a verbal yelling out. OK, ready. This arrow.

AUDIENCE: Km.

PROFESSOR: This one here is, 3,2,1--

AUDIENCE: Km.

PROFESSOR: Km. All right. All right. Ready, 3, 2, 1.

AUDIENCE: Gamma m times m.

PROFESSOR: Gamma m times m. 3, 2, 1.

AUDIENCE: Gamma n times m plus 1.

PROFESSOR: Gamma m times m plus 1. Now remember that there are more mRNA over here then there are here, which means that the rate of degradation will increase. Now coming here, now this is talking about the creation and destruction of the proteins, changes in n. All right, this arrow here. Ready, 3, 2, 1.

AUDIENCE: Kp times m.

PROFESSOR: It's Kp times m. So this is the rate of creation, going from n minus 1 to n. That's fine. You know, I was looking at my notes from last year, and I got one of these things incorrect, so-- and then, OK, ready. This one here, 3, 2, 1. Kp times m. So here the

same rate, and should we be surprised by that?

So the number of proteins are changing, but here it's the number of mRNA that matters, because we're talking about the rate of translation, right? Now this one here, 3, 2, 1. Gamma p times n. And here, 3, 2, 1.

AUDIENCE: Gamma p times n plus 1.

PROFESSOR: Gamma p times n plus 1. All right. Perfect. Now this is, of course, as you can imagine, the simplest possible kind of set of equations that we could have written down. If you have other crazy things, you get different distributions, if you have autoregulation or if you have interactions of something with something else, or the same thing, so forth.

But I think it's really very useful to kind of write this thing down to clarify your thinking in these problems. And then you can fill out-- for change of probability, you have mn. You come here and you just go around and you count, take all the arrows coming in, and those are ways of increasing your probability. And ways going out are ways of decreasing your probability.

Now in all those cases you have to multiply these raw rates by the probabilities of being in all these other states.

So can you use the master equation to get these probabilities if you're out of equilibrium, out of steady state? So that's a question. So the master equation useful out of steady state? Yes. Ready. 3, 2, 1. All right. So we got a fair number of-- there is some disagreement, but yeah.

So it actually-- the answer is yes. And that's because you can start with any distribution of probabilities across all the states that you'd like. It could be that all of the probabilities at one state. It could be however you like. And the master equation tells you about how that probability distribution will change over time.

Now if you let that run forever, then you come to some equilibrium steady state. And that's a very interesting quantity, is the steady state distribution of these

probabilities. But you can actually calculate from any initial distribution of probabilities evolving to any later time t what the probability would be later.

This comes to another question here. All right. So let's imagine that at time t equal to 0, I tell you that there are m not mRNA and P not-- I always do this. I don't know, somehow my brain does not like this. Because the P's we want to be probabilities. We start with m not mRNA, n not protein.

And maybe it's a complicated situation. We can't calculate this analytically. So what we do is we go to our computer, and we have it solve how this probability distribution will evolve so that time T equal to some time-- if we'd like we can say this is T1. I'll tell you, oh, the probability of having m and n mRNA and protein is going to be equal to something P1.

Now the question is, let's say I then go and I do this simulation again. Now I calculate some other at time T1 again, the probability that you're in the m,n state. The question is, will you again get P1? So this is a question mark. And A is yes, B is no. All right. I'm going to give you 15 seconds. I think this is very important that you understand what the master equation is doing and what it is not doing.

AUDIENCE: [INAUDIBLE]

PROFESSOR: I'm sorry, what's that? Right. OK. So I mean, this is just-- you know, you program in your computer to use the master equation to solve how the probabilities are going to evolve. I'm just telling you, start with some initial distribution. And if you do it once, it says, oh, the probability that you're going to have m-- this time you're going to have mRNA proteins is going to be P1, so it's 10%. Great.

Now I'm asking just, if you go back and do it again, will you again get 10%, or is this output stochastic? It's OK that if you're confused by this distinction. I think that it's easy to get confused by, which is why I'm doing this. But let's just see where we are. Ready? 3, 2, 1.

All right. So I'd say a majority again. We're kind of at the 80-20, 75-25. A majority

here are saying that, yes, you will get the same probability. And this is very important that we understand kind of where this where the stochasticity is somehow embedded in these different representations of these modelings.

The master equation is a set of differential equations telling you about how the probabilities change over time given some initial conditions. Now we're using these things to calculate the evolution of some random process, but the probabilities themselves evolve deterministically. So what that means is that although these things are probabilities, if you start somewhere and you use the master equation to solve, you get the same thing every time you do it.

Now this is not true for the Gillespie simulation, because that, you're looking at an individual trajectory. An individual trajectory, then the stochasticity is embedded in that trajectory itself, whereas in the master equation, the stochasticity arises because these are probabilities that are calculating, so any individual instantiation will be probabilistic because you are sampling from those different probability distributions.

Now this is, I think, a sufficiently important point that if there are questions about it, we should talk about it. Yeah.

AUDIENCE: How do you make the simulations? Would you essentially-- can you take a sum over different Gillespie?

PROFESSOR: So it's true that you can do a sum over different Gillespie. But we haven't yet told you about, what the Gillespie algorithm is, so I can't use that. But indeed, you can just use a standard solver of differential equations. So whatever program you use is going to have some way of doing this.

And once you've written down these equations, the fact that these are actually probabilities doesn't matter. So those could have been something else. So this could be the number of eggs, whatever, right? So once you've gotten the equations, then equations just tell you how the problems are going to change over time. Yeah.

AUDIENCE: Maybe this is a silly question, but in practice, do you have to assume all the

probabilities are 0 above some number?

PROFESSOR: Oh no, that's not at all a silly question, because--

AUDIENCE: [INAUDIBLE]

PROFESSOR: Exactly. Right. And yes, it's a very good question. So I told you this is an infinite set of differential equations. But at the same time I told you this master equation's supposed to be useful for something, and kind of at the face of it, these are incompatible ideas.

And the basic answer is that you have to include all the states where there is a sort of non-negligible probability. We could be concrete, though. So let's imagine that I tell you we want to look at the mRNA number here. And I tell you that OK, Km is equal to-- well, let me make sure. Gamma m. What are typical lifetimes of mRNAs in bacteria again?

AUDIENCE: [INAUDIBLE]

PROFESSOR: Right. Order a minute. So that means that-- let's say this is 0.5 minutes minus 1. To get a lifetime of around 2 minutes. And then let's imagine that this is then 50 per minute. So an mRNA is kind of made once a minute. There's 50 of them. That's a lot, but whatever. There are a few genes. Minute. I wanted the number to be something.

So there's a fair rate of mRNA production. Now how many equations do you think you might need to simulate? So we'll think about this. First of all, does it depend upon the initial conditions or not?

AUDIENCE: Maybe.

PROFESSOR: Yeah. It does. So be careful. But let's say that I tell you that we start with 50 mRNA. The question is, how many equations do you think you might have to write down? And let's say we want to understand this once it gets to, say, the equilibrium.

All right. Number of equations. Give me a moment to come up with some

reasonable options. Well, these are-- let's say that this could show up on your homework. So the question is, how many equations are you going to program into your intersimulation? And it may be-- doesn't have to be exactly any of these number, but order. Do you guys understand the question?

So we need a different equation for each of these probabilities. So in principle we have-- the master equation gives us an infinite number of equations. So we have d the probability of having 0 mRNA with respect to time. That's going to be-- any idea what this is going to be?

AUDIENCE: [INAUDIBLE]

PROFESSOR: Right. So we have a minus Km times what? Times p0, right. So this is because if we start out down here at P0. Now we have Km. So I was just about to violate my rule and just write down an equation without drawing this thing. So it's Km times p0. That's a way you lose probability, but you can also gain probability at a rate that goes as gamma m times P1.

So that's how this probability is going to change over time. But we have a different equation for you for p1, for p2, for p3, for p4, all the way, in principle, to p1,683,000, bla bla bla, right? So that's problematic, because if we have to actually in our program code up 100,000,000 equations, or it could be worse. Then we're going have trouble with our computers. So you always have to have some notion of what you should be doing.

And this also highlights that it's really important to have some intuitive notion of what's going on in your system before you go and you start programming, because in that case-- well, you're likely to write down something that's wrong. You won't know if you have the right answer, and you might do something that doesn't make any sense. So you have to have some notion of what the system should look like before you even start coding it. My question is, how many of these equations should we simulate?

OK. Let's just see where we are. Ready. 3, 2, 1. OK. So I'd say that we have, it's

basically between C and D. Yeah. I would say some people are maybe more careful than I am. Can one of the Ds maybe defend why they're saying D?

AUDIENCE: [INAUDIBLE]

PROFESSOR: The mean is 100, and when you say-- I mean, I think that whatever you're thinking is correct, but I think that the words are a little dangerous. And why am I concerned about-- you said-- is the mean 100 for all time?

AUDIENCE: [INAUDIBLE] and steady state.

PROFESSOR: And steady state. Right. I think that was the-- for long times, the mean number of mRNA will, indeed, be 100. So the mean number of m, in this case, will be Km divided by gamma m, which is going to be equal to 50 divided by that. That gets us 100. Now will it be exactly 100? No. It's going to be 100 plus or minus what? Plus or minus 10. Right. Because this distribution at steady state is what?

AUDIENCE: It's Poisson.

PROFESSOR: It's Poisson. What's the variance of a Poisson distribution? It's equal to the mean. So for Poisson, the variance is equal to the mean. Variance is the square of the standard deviation, which means that this is going to be plus or minus 10. That's kind of the typical width of the distribution. So what it means is that at equilibrium, we're going to be at 100 and it's going to kind of look like this. So this might be 2 sigma, so this could be 20. But each of these is 10.

So if you want to capture this, you might want to go out to a few sigma. So let's say you want to go out to 3 sigma, then you might want to get out to 130 maybe. So then, if you want to be more careful you go out to 140 or 150. But this thing is going to decay exponentially, so you don't need to go up TO 1,000, because the probability's going to be 0 0 0. Certain once you're at 200 you don't have to worry about it.

But of course, you have to remember the initial condition we started at 50. So we started at this point, which means we definitely have to include that equation.

Otherwise we're in trouble. Now how much do we have to go to below 50 Any--

AUDIENCE: My guess would be that it would be not much more than the [? few ?] times 5, because if it were already at equilibrium that would be the mean. But it's not, and so the driving force is still going to push it back to [INAUDIBLE].

PROFESSOR: That's right. So it's going to be a bias random walk here, where it's going to be sort of maybe twice as likely at each step to be moving right as to be moving left. That means it could very well go to 49, 48. But it's not really going to go below 40, say. Of course you have to quantify these things if you want to be careful. But certainly I would say going from, I don't know, 35 to 135 would be fine with me. You would get full credit on your problem set.

So we'll say-- I'm going to make this up-- from 35 to 135, 134 just so it could be 100 equations. So I'd say I'd be fine with 100 equations. So you would simulate the change in the probabilities of P35 to P134, for example. So although in principle, the master equation specifies how the probabilities for an infinite number of equations are going to change, you only need to simulate a finite number of them depending upon the dynamics of your system. Yes. Thank you for the question, because it's a very important practical thing.

AUDIENCE: So in practice, you don't know what the solution is, which is sort of why you would [INAUDIBLE]. Do you explain your range and see if the solution changes?

PROFESSOR: So the question is, in this case, it's a little bit cheating because we already kind of knew the answer. We didn't know exactly how the time dependence was going to go. How is it that the mean is going to change over time on average? Exponentially, right? So on average you will start at 50. You exponentially relax to 100. But in many cases, we don't know so much about the system. And I'd say that what, in general, you can do is, you have to always specify a finite number of equations. But then what you can do is, you can put in, like, reflecting boundary conditions or so on the end, so you don't allow probability to escape.

But then what you can do is you can run the simulation, and if you have some

reasonable probability to any of your boundaries, then you know you're in trouble and you have to extend it from there. So you can look to say, oh, is it above 10 to the minus 3 or 4, whatever. And then if it is, then you know you have to go further. Any other questions about how-- you're actually going to be doing simulations of this, so these are relevant questions for you. All right.

So that's the master equation. But I'd say the key, key thing to remember is that it tells you how to calculate the deterministic evolution of the probability of these states given some potentially complicated set of interactions.

Now, a rather orthogonal view to the master equation is to use the Gillespie algorithm, or in general, to do direct stochastic simulations of individual trajectories. Yeah. Question before we go.

AUDIENCE: So if we just set it to 0, the probabilities outside the range we think we need, would we be losing probability?

PROFESSOR: So the question is whether we're somehow losing probability. So what I was proposing before is that you always want probabilities to sum to 1. Otherwise it's not our probability and the mathematicians get upset. And the key thing there is that you want to start with-- you have to include all the states that have probability at the beginning.

So in that sense, you're given an initial distribution, and you have to include all those states. Otherwise you're definitely going to do something funny. You start out with a normalized probability distribution. And then I guess what I was proposing is that you have a finite number of equations, but you don't let the probability leave or come in from those ends.

And if you do that, then you will always have a normalized probability distribution. Of course, at the ends you've kind of violated the actual equations, and that's why you have to make sure that you don't have significant probability at any of your boundaries. Does that answer? Not quite?

AUDIENCE: Because I'm wondering if [INAUDIBLE].

PROFESSOR: So I was not suggesting that you set the probabilities equal to 0. I was suggesting that you do what's kind of like what the equations actually here, which is that you don't allow any probability to leave. There's no probably flux on this edge.

So for example, out at P134, I would just say, OK, well, here's the probability that you have 134 mRNA. And in principle there are these two arrows, but you can just get rid of them. So now any probability that enters here can only come back. And I've somehow violated my equations. But if P134 is essentially 0, then it doesn't matter.

So instead of looking at these probabilities evolve kind of as a whole, we can instead look at individual trajectories, right? So the idea here is that if we start with the situation-- actually, we can take this thing here. So we know that at steady state it's going to be 100. Starts out at 50. And in this case, with the master equation you say, OK, well, you start out with all the probability right here.

So you have kind of a delta function at 50. But then what happens is this thing kind of evolves, and over time this thing kind of spreads until you have something that looks like this, where you have a Poisson distribution centered around 100. And this Poisson distribution's going to be very close to a Gaussian, because you have a significant number.

So the master equation tells you how this probability distribution evolves. Now this is the number m and this is kind of the frequency that you observe it. So we can also kind of flip things so we instead plot the number m on the y-axis. And we already said the deterministic equations will look like this. And the characteristic time scale for this is what?

1 over mm, right? So this thing relaxes to the equilibrium, time scale determined by the degradation time of the mRNA. So these are things that should be really-- you want to be kind of drilled into your head, and I'm trying to drill, so you'll hear them again and again.

Now the master equation, indeed, since everything's linear here, the expectation

value over the probability distributions actually does behave like this. So the mean of the distributions as a function of time look like that. And in some ways, if we were to plot this, we would say, OK, well, first of all it's all here. Then it kind of looks like this. So this is somehow how those probability distributions are kind of expanding over time.

Now for an individual trajectories, if we run a bunch of stochastic simulations, we'll get something that on average looks like this, but it might look like this. A different one might look like this, and so on, although they shouldn't converge there because that's not consistent.

And if you did a histogram at all those different times of the individual stochastic trajectories, you should recover the probability distribution that you got for the master equation.

So this is a powerful way just to make sure that, for example, your simulations are working, that you can check to make sure that everything behaves in a consistent way.

Now there's a major question, though, of how is it that you should generate these stochastic trajectories? And the sort of most straightforward thing to do is to just divide up time into a bunch of little delta t's, and just ask whether anything happened. So let me--

So what we want to do is we want to imagine we have maybe m chemical species. So now these are different m's and n's. Be careful. m chemical species, they could be anything, could be proteins, they could be small molecules, something. And there are n possible reactions.

And indeed, in some cases people want to study the stochastic dynamics of large networks. So you could have 50 chemical species and 300 different reactions. So this could be rather complicated. And these m chemical species have, we'll say, numbers or if you'd like, in some cases it could be concentrations, Xi, so then the whole thing can be described as some vector X.

And the question is, how should we assimilate this? The so-called, what we often call the naive protocol-- and this is indeed what I did in graduate school because nobody told me that I wasn't supposed to do it-- is that you divide time into little time segments delta t.

Small delta t. And you just do this over and over. And for each delta t you ask, did anything happen? If it did, then you update. If not, you keep on going. Now the problem with this approach-- well, what is the problem with this approach?

AUDIENCE: [INAUDIBLE]

PROFESSOR: Yeah. Time is continuous. So one problem is that, well, you don't like discrete time. That's understandable. But I'm going to say, well, you know, the details-- a delta t may be small, so you won't notice. I'm saying, if I said delta t being small, then I'm going to claim that you're not going to notice that I've--

AUDIENCE: [INAUDIBLE]

PROFESSOR: But then the simulation is slow, right? So there's a fundamental trade-off here. And in particular, the problem with this protocol is that for it to behave reasonably, delta t has to be very small. And what do I mean by very small, though?

AUDIENCE: [INAUDIBLE]

PROFESSOR: That's right. For this to work, delta t has to be such that unlikely for anything to happen. But this is already a problem, because that means that we're doing a lot of simulations, and then just nothing's happening. How do we figure out what that probability is?

So in particular, we can ask about-- well, given possible reactions, we'll say with rates rs of i. So the probability that the i'th reaction occurs is equal to r i times delta t for small delta t, because each of these reactions will occur kind of at a rate-- they're going to be exponential distributions of the times for them to occur. This is a Poisson process because it's random.

Now what we want to know is the probability that nothing is going to happen because that's how we're going to have set delta t. Well, what we can imagine is, then we say, well, what's the probability that is, say, not reaction 1 and not 2 and dot dot dot. OK. Well, and this is in some time delta t.

Well, actually, we know that if the fundamental process just looks like this, then we're going to get exponential distributions for each of those. So we end up with e to the r1, and indeed, once we write an exponential, we don't have to write delta t. This is just some time t. For this to be true requires a delta t is very small. But if we want to just ask, what's the probability that reaction 1 has not happened in some time t, this actually is, indeed, precisely equal to e to the r1t. Yeah, details.

And this is e to the minus r2t dot dot dot minus. And we go up to n, r to the nt, because each of those chemical reactions are going to be exponentially distributed in terms of how long you have to wait for them to happen.

And what's neat about this is that this means that if you just ask about the probability distribution for all of them combined by saying that none of them have happened, this is actually just equal to the exponent of minus-- now we might pull the t out and we just sum over ri.

So this is actually, somehow, a little bit surprising, which is that each of those chemical reactions occur, and they're occurring at different rates. Some of them might be fast, some of them might be slow. The ri's can be different by orders of magnitude. But still, over these hundreds of chemical reactions, if the only thing you want to know is, oh, what's the probability that none of them have happened, that is also going to end up-- that's going to decay exponentially.

And this actually tells us something very interesting, which is that if we want to know the distribution of times for the first thing to happen, that's also going to be exponentially distributed. And it's just exponentially distributed with a rate that is given by the sum of these rates. Now that's the basic insight behind this GIllespie algorithm, where instead of dividing things up into a bunch of little times delta t, instead what you do is you ask, how long am I going to have to wait before the first

thing happens? And you just sample from an exponential with this rate r that is the sum of the rates.

Maybe it's even worth saying that, OK, so there's the naive algorithm where you just divide a bunch of delta t's, you just take a little steps, you say, OK, nothing, nothing, nothing, nothing, and then eventually something happens, and then you update, you keep on going.

There's the somewhat less naive algorithm, which is exact, so it's not the same concerns, the j hat which is that you could just sample from n different exponentials, each with their own rates, and then just take the minimum of them and say, OK, that's the that happened first, and then update from that. And that's an exact algorithm.

But the problem is that you have to sample from possibly many different exponentials. And that's not a disaster, but again, it's computationally slow. So the Gillespie algorithm removes the requirement to from those n exponentials, because instead what you do is you just say, the numbers, or the concentrations, give all of the ri, give you all the rates.

And then what you do is you sample from an exponential with rate r, which is the sum over all the ri. That tells you, when is the first reaction going to occur. And then what you do is you ask, well, which reaction did occur? Because you actually don't know that yet. And there, it's just the probabilities of each of them. So the probabilities Pi is just going to be the ri divided by the sum over the ri, so this big R.

So it may be that you had 300 possible chemical reactions, but you only have to do two things here. And they're both kind of simple, right? You sample from one exponential, gives you how long you had to wait for something to happen. And then you just sample from another simple probability thing here that just tells you which of the n possible chemical reactions was it that actually occurred. And of course, the chemical reactions that were occurring at a faster rate have a higher probability of being chosen.

So this actually is an exact procedure in the sense that there's no digitization of time or anything of the sort. So this actually is computationally efficient and is exact, assuming that your description of the chemical reactions was accurate to begin with.

So then what we do is we update time. This is in some ways-- when you do computations, when you actually do simulations-- this is maybe the annoying part about the Gillespie algorithm, which is that now your times are not equally spaced, and so then you just have to make sure you remember that, you don't plot something that's incorrect. Because your times are going to hop at different time intervals. But that's doable. You have to update your time and you have to update your abundances. And then what you do is repeat.

I think the notes kind of allude to this Gillespie algorithm but are not quite explicit about what you actually do to go through this process. For the simulations that you're going to do in this class, I would say that you don't get the full benefits of the Gillespie in the sense that you're not going to be simulating hundreds of differential equations with hundreds of different things. But it's in those complicated models that you really have to do this kind of Gillespie approach, as compared to even this somewhat better model, which is you sample from the different exponentials.

Are there any questions about why this might work, why you might want to do it? Yes.

AUDIENCE: What do you mean by sample the exponentials?

PROFESSOR: Right. What I mean is that you go to Matlab and you say, random-- I'm sort of serious, but-- sorry, I'm trying to get a new-- All right. So you the exponential. So it's a probability distribution. So this is the probability is a function of time and then t. And it's going to look something like this. This thing is going to be some-- given that, in general, it's going to be the probability t is going to be e to the minus rt. And then do I put r here or do I put 1 over r?

AUDIENCE: [INAUDIBLE]

PROFESSOR: Is it 1 over r? Well, what should be the units of a probability distribution? 1 over

time, in this case. It's 1 over whatever's on this x-axis, because if you want to get the actual, honest to goodness probability-- so if you want the probability that t is, say, between t1 and t1 plus delta t. If you want an actual probability, then this thing is equal to the probability density at t1, in this case, times delta t. So that means thing has to have a 1 over time, and that gives us r here.

So this is probability density, and what I'm saying is that when I say sample from this probability distribution, what it means is that it's like rolling a die, but that it's a biased die because it's continuous thing over the time. But just like when you have a six-sided die and I say, OK, sample from the die, you're playing Monopoly, you throw the die and you get 1, 2, 3, 4, 5, 6. And you do that over and over again.

Same thing here. You kind of roll the die and see what happens. And indeed, you're going to get some practice with probability distributions on the homework that you're doing right now because you're asked to demonstrate that you can sample from a uniform distribution, which something that's just equally probable across the unit line, and do a transformation and get an exponential distribution.

And it used to be that everybody knew all these tricks because you had to kind of know them in order to do computation. But now, Matlab, or whatever program you use, they know all the tricks, so you just ask it to sample from an exponential with this property and it does it for you. But you still need to know what it's doing.

So just to be clear, what is the most likely time that you're going to get out from the exponential? 0. It has a peak here but the mean is over here. Any other questions about how the Gillespie algorithm works?

Can somebody tell me how a protein burst arises? So we had this original question about whether there were protein bursts in that model that I wrote down, where we just had m dot is equal to--

Now what we said was that the master equation would not-- the protein burst would somehow be there are but you would never see them, or somehow the protein burst would influence how the mean and everything have evolved, but you wouldn't

actually see any big jumps. But then we said, oh, but if you did a stochastic simulation, you would. So the claim here is that the Gillespie algorithm, what I've just told you here, will lead to protein bursts. When I make that statement, what is it that I actually mean?

If we do a Gillespie of this, will the-- OK, let's just hold on. Let me do a quick vote. Will we have cases where delta n is greater than 1? If I go through this process, if I'm using the Gillespie and I'm tracking how mRNA and protein number are changing over time, will I get these things, protein bursts, where delta n is larger than 1 in one of these time cycles?

Ready? 3, 2, 1. So most of the group is saying that it's going to be no. But again, it's mixed. So can somebody say why we don't get--

AUDIENCE: [INAUDIBLE] It seems like the structure of the simulation is to make sure [INAUDIBLE].

PROFESSOR: That's right. Yeah. So the simulation as written-- you could imagine some sort of phenomenological version of this where you allowed, actually, for protein bursts. But as kind of specified is that we ask, what's the time for one thing to happen? But the claim somehow is, OK, well, we can still get protein bursts from this. And how does that happen?

AUDIENCE: You can have the rate for something happening increase suddenly, and that would happen if we go from m equals 0 to m equals 1--

PROFESSOR: Yeah, for example, if we didn't have an mRNA before and we got an mRNA. What it means that if you look at n as a function of time during one of these protein bursts-- before, I was drawing it just hopping up, but really, in the context of the Gillespie, it would be that it would hop, hop. So there would be little time jumps. So this is a protein burst, but it's really before this mRNA is degraded, you get 1, 1, 1, 1.

So each of these as is delta n of 1. So this is whatever, 6, 7. And then what can happen is that we get the mRNA degraded. And so then we're going to get a slower thing where it-- looks like that. So the Gillespie, everything is being created and

destroyed in units of 1. But it could be that the time interval over this burst is just very short, so then it goes up very quickly, but then it's slower to go away.

So what I want to do in just the last 15 minutes is talk a bit about the Fokker-Planck approximation. I would say that all these different approaches are useful to varying degrees in terms of actually doing simulations, doing analytic calculations, getting intuition. And the Fokker-Planck approach, I'd say it's more or less useful for different people depending on what you're doing.

So the basic idea, as kind of you answered in the pre-class reading, is that in cases where n is large enough that you don't feel like you need to take into account the discrete nature of the molecules, yet at the same time it's not so large that you can totally ignore the fluctuations, then the Fokker-Planck approach is nice because it allows you to get some sense of what's going on without all of the crazy details of, for example, the master equation. And then it also, because of this idea of an effective potential, it allows you to bring all the intuition from that into your study of these gene circuits.

Now I'm not going to go through the whole derivation, but if you have questions about that, please come up after class and I'm happy to go through it with you, because it's sort of fun. But the notes do go over it. I think that's what's perhaps useful to just remind ourselves of is how it maybe leads to a Gaussian with some width depending upon the shapes of the production degradation curves.

So the basic notion here is that, depending on the f's and g's, the production degradation terms, we get different shaped effective potentials. So in general we have something that looks like-- we have some n dot, there's some fn, and then there's a minus gn.

So for example, for something that is just simple expression, in the case of-- let's just imagine now that there is-- if you want we can say it's a protein where it's just some k minus gamma n. Or if you'd like, we could say, oh, this is mRNA number. But something that's just simple production, and then first order degradation.

The question is, how do we go about understanding this in the context of the Fokker-Planck approximation? And it turns out that you can write it in what is essentially a diffusion equation where you have some probability flux that's moving around. And within that realm, you can write that the probability distribution of the number is going to be something that-- so there's going to be some constant. There's f plus g. And these are both functions of n. And then you have e to the minus [INAUDIBLE]

So the idea here is that this behaves as some effective potential. Of course, it's not quite true because f and g also are functions of n, they're are not in here. But this is the dominant term because it's in the exponential. And here phi n is defined as the following. So it's minus this integral over n of the f minus g and f plus g dn that we integrate over n prime.

And we're going to kind of go through what some of these different f's and g's might look like to try to get a sense of why this happened. It is worth mentioning that you can do this for any f and g when it's just in one dimension, so you just have n. Once you have it in two dimensions, so once you actually have mRNA and protein, for example, you're not guaranteed to to be able to write it as an effective potential. Although I guess if you're willing to invoke a vector potential, then maybe you can.

But in terms of just a simple potential, then you can do it one dimension, but not necessarily in more. And I think that, in general, our intuition is not as useful when you have the equivalent of magnetic fields and so forth here anyway.

What I want to do is just try to understand why this thing looks the way it does for this simple regulation case. And then we're going to ask if we change one thing or another, how does it affect the resulting variance.

So for unregulated expression, such as here, if we look at the production and degradation as a function of n, fn is just some constant k, whereas gn is a line that goes up as gamma n. Now in this situation, if you do this integral-- and really, what you can imagine is what this integral looks like right around that steady state, because that's kind of what we want to know, if we want to something about, for

example, the width of a distribution.

Well, there's going to b e two terms. In the numerator there's an f minus g. In the denominator there's an f plus g. Now f minus g is actually equal to 0 right at that steady state, and that's why it's a steady state, because production and degradation are equal. Now as you go away from that location, what you're doing is you're integrating the difference between the f and the g.

And you can see that around here these things are separating kind of-- well, everything's a line here. And indeed, even if f and g were not linear, close to that steady state they would be linear. What we can see is that as you're integrating, you're integrating across something that is growing linearly. That's what gives you a quadratic. And that's why this effect of potential ends up behaving as if you're in a quadratic trap.

Now I encourage you to go ahead and do that integral at some point. I was planning on doing it for you today, but we are running out of time. Once again, I'm happy to do it, just after class. And indeed, what you can see is that because you're integrating across here, you end up getting a quadratic increase in the effective potential. And if you look at what the variance of that thing is, you indeed find that the variance is equal to the mean here.

So what I want to ask in terms of trying to get intuition is, what happens if we pull these curves down? So in particular, let's imagine that we have a situation where-- I'm going to re-parameterize things, so again, we're kind of keeping the number of the equilibrium constant. But now what I'm going to do is I'm going to have an fn that looks like this, and gn looks like-- now gn is going to be some 1/2 of lambda, and this fn is equal to k minus 1/2 of gamma n.

Now the question is, in this situation, what will be the variance over the mean? Well, first of all, the variance over the mean here was equal to what? Although should we do vote? Here are going to be some options.

Question is variance over the mean in this situation. I'm worried that this is not going

to work, but let's just see where are. Ready, 3, 2, 1. All right. So I'd say that at least broadly, people are agreeing that the variance over the mean here is equal to 1.

And again, this is the situation that we've analyzed many times, which is that in this situation we get a poisson, where the poisson only has one free parameter, and that parameter specifies both the mean and the variance. So for a poisson, the variance of the mean is indeed equal to 1. So the Fokker-Planck approximation actually accurately recapitulates that.

Now the question is, what will the variance over the mean be in the situation that I've just drawn here? So I'm going to give you a minute to try to think about what this means. And there are multiple ways of figuring it out. You can look at, maybe, the integral. You can think about the biological intuition to make at least a guess of of what it should do.

The question is, if the production rate and the degradation rate look like this, what does that mean for the variance over the mean? So I'll give you a minute to kind of play with it.

Why don't we go ahead and vote, just so I can get a sense of where we are? And also, it's OK if you can't actually figure this out or you're confused. But go ahead and make your best guess anyways, because it's also useful if you can guess kind of the direction it'll go, even if you can't figure out its magnitude.

So let's vote. Ready, 3, 2, 1. OK. So it's a mixture now, I'd say, of A, B, C, Ds. Yeah, I think this is, I think, hard and confusing. I maybe won't have-- all right. I'll maybe say something. It may be that talking to each other won't help that much.

OK, so in this case, what's relevant is both the f minus g and the f plus g. And it turns out that f minus g actually behaves the same way, because at the fixed point, or at the equilibrium, it starts at 0 and then it actually grows in the same way as you go away from it. The difference is the f plus g, where that's very much not equal to 0. And f plus g at the equilibrium, this f plus g here is around 2k, whereas f plus g over here is around 1k.

What that means is that in both cases you have a quadratic potential. But here the quadratic potential actually ends up being steeper. So if this were unregulated, then over here we still get a quadratic, but it's with steeper walls. So actually here, this, the variance over the mean, ends up being 1/2.

It's useful to go ahead and just play with these equations to see why that happens. And I think that's a nice way to think about this is, in this limit, where we pull this crossing point all the way down to 0, now we have something that looks kind of like this. So very, very low rate of degradation.

But then also the production rate essentially goes to 0 when we're at this point. So we could still parameterize as k over gamma if we want, with some-- but we could just think about this as being at 100 of these mRNAs, say. But then we're changing the production degradation rate.

And the variance over the mean here-- does anybody have a guess of where that goes? In this case it actually goes to 0. And this is an interesting situation, because really, in the limit where there's no degradation, and it's all at the production side, what it's saying is that you produce, you produce, you produce, until you get to this number, which might be 100, and then you simply stop doing anything. You're not degrading, you're not producing. In that case all the cells will have exactly 100, maybe, mRNA.

And what the Fokker-Planck kind of formalism tells you is that just because production and degradation rates are equal, f minus g is equal to 0, doesn't mean that-- that tells you that that's the equilibrium, but it doesn't tell you how much spread there's going to be around the equilibrium. If f and g are each larger, that leads to a larger spread because there's more randomness, whereas here, f and g are both essentially 0 at that point. What that means is that you kind of just pile up right at that precise value.

We are out of time, so I think we should quit. But I am available for the next half hour if anybody has any questions. Thanks.