

Physics through Computation Thinking

Dr. Auditya Sharma & Dr. Ambar Jain

Department of Physics

Indian Institute of Science Education and Research, Bhopal

Lecture 49

Random Walks: The central limit theorem.

Hi guys, so we have had quite a few discussions now on random walks. We saw that there are these nice properties of random walks, namely that average of average distance squared, as a function of time will go linearly and so we derived the exact result for for the discrete random walk in 1D, then we saw how there is a connection between these random walks which are discrete in nature and when you make it continuous time and space you get the diffusion equation.

And in that context, I had mentioned that in fact, this result that the average distance covered as a function of time, you know, will go as a square root of the time and and the fact that this holds regardless of whether you make the time and space discrete or continuous or you know you keep one of them discrete and you make the other one continuous, it does not matter.

In all these contexts, you get this basic rule, which is that distance typically goes as square root of the time. And so, this result is very very general and this is called diffusive motion. And so today in this module, I want to motivate where this comes from? What is the cost for this kind of generality? And why do random walks appear in all kinds of context where apparently there is nothing to do with a you know drunkard or a random walker involved?

There is no obvious way to see this and yet these results are so generic in nature and they appear in whenever there is stochastic variables involved, they they appear and so I want to show you how this is connected to a very deep and important theorem called the central limit theorem. Probabilist spent a lot of time going into the rigorous formulation of this theorem and proving it and analyzing various technical aspects of it, but our goal here will be to simply motivate.

And in fact, just state the theorem in a very non-rigorous way and then do a few checks in the true spirit of this course, we will do some numerical checks to explore the generality of this theorem and then set it up in such a way that it will actually, its it makes it amenable to more such exploration, which I will allow you to pot around and play with. OK. So here is my non-rigorous statement of the central limit theorem.

(Refer Slide Time: 2:58)

The Central Limit Theorem

The reason for the ubiquitous appearance of the Gaussian distribution has to do with a deep and important theorem called the Central Limit Theorem. It is difficult to prove, but fairly easy to state in a non-rigorous way. Our goal here is to verify it numerically for a number of cases.

Statement (Non-rigorous)

Let X_1, X_2, \dots, X_N be N independent random variables with means $\mu_1, \mu_2, \dots, \mu_N$ and standard deviations $\sigma_1, \sigma_2, \dots, \sigma_N$ respectively. Then the distribution of the sum

$$S = X_1 + X_2 + \dots + X_N$$

tends to a Gaussian distribution with mean $\mu = \mu_1 + \mu_2 + \dots + \mu_N$, and variance $\sigma^2 = \sigma_1^2 + \sigma_2^2 + \dots + \sigma_N^2$.

Exercise

- Numerically generate a few sample random walks where the lengths of individual steps are drawn from a uniform distribution in the interval $(-1, 1]$ and visualize them.
- Check if the key result $\langle n^2 \rangle = aN$ is still robust by numerically generating many samples of random walks, and averaging.
- Analytically compute a . Check by explicit plotting if your numerical data support this expectation.

Solution

Suppose, you have N random variable and X_1, X_2 so on until X_N and they all have different means, μ_1, μ_2 so on μ_N and standard deviations, $\sigma_1, \sigma_2, \sigma_3$ all the way up to σ_N . Now, it is often of interest to consider the sum of all these random variables, so lot of different context give us sums of a large number of random variables and that is where the central limit theorem comes into play.

Whenever you have a sum of a large number of random variables, regardless of the microscopic information or the precise form of the distribution of each of these random variables, it turns out that the sum of all these random variables tends to go to a definite distribution and that is the Gaussian distribution whose mean is just simply given by the sum of the means of all these different random variables and whose variance sigma squared is simply the sum of the variance of all these random variables.

So, that is where that is why Gaussian, these bell shaped curves are so ubiquitous, you might have encountered these in the context of, you know, your marks distributions, for example, in your among various different students or if you were to study the distribution of heights of various members of some group in your class or somewhere in some other sample, you would find a bell shaped curve or you might have encountered bell shaped curves in the context of your lab experimental results that you you measured of some data set and then you can just make a distribution that is the histogram is likely to be like a bell curve.

So, there is some underlying sum involved that is the implication, which I am trying to get across. Now, so let us do a few exercises to see whether this is a reasonable thing. So, what have we done in the past, we have said that your random walker can either go to the right or left with some probability p and $1 - p = q$ and we have checked this.

So, now I want to consider a somewhat more general distribution. The individual steps, suppose I make those individual steps not precisely $+1$ or -1 , but I allow it to be drawn from some uniform distribution in the spirit of this generalization here, so it tells you that it does not have to be some very restricted sum of variable, which are all the same, nothing like that, it is all these X_1 , X_2 so on can be completely dependent. They can be drawn from different distribution.

So, here what I want to do is, draw these individual steps from a uniform distribution in the interval -1 to $+1$ and let us visualize them. So, if you want to pause the video here, and write your own code to generate a few sample random walks of this kind and visualize them, please feel free to do so. It is easy to do because I have already given you the code for the case where you have $+1$ or -1 type of motion, some small tweaks to it will allow you to write down the code for this.

So, you you are strongly encouraged to pause the video and work this out for yourself. I have also shown you, you know, the next couple of exercise which are coming your way if you want to do all of them and only then continue that is also fine.

(Refer Slide Time: 6:30)

Statement (Non-rigorous)

Let X_1, X_2, \dots, X_N be N independent random variables with means $\mu_1, \mu_2, \dots, \mu_N$ and standard deviations $\sigma_1, \sigma_2, \dots, \sigma_N$ respectively. Then the distribution of the sum

$$S = X_1 + X_2 + \dots + X_N$$


tends to a Gaussian distribution with mean $\mu = \mu_1 + \mu_2 + \dots + \mu_N$, and variance $\sigma^2 = \sigma_1^2 + \sigma_2^2 + \dots + \sigma_N^2$.

Exercise

- Numerically generate a few sample random walks where the lengths of individual steps are drawn from a uniform distribution in the interval $[-1, 1]$ and visualize them.
- Check if the key result $\langle n^2 \rangle = aN$ is still robust by numerically generating many samples of random walks, and averaging over them.
- Analytically compute a . Check by explicit plotting if your numerical data support this expectation.

Solution

```
a = {0};
Do[AppendTo[a, a[[n - 1]] + 1 - 2 RandomReal[]], {n, 1, 500}];
ListPlot[a, Joined -> True];
```



So, here is my solution. So, like before, I start my array with just one element and that element is 0 and then I will use this AppendTo and also the Do command, so AppendTo[a, a[n - 1] + 1 - 2RandomReal(randomreL)]. So, what does this 1 - 2RandomReal do? So, RandomReal will give me a uniform distribution, random variable gives a uniform distribution in the interval 0 to 1, so if I just do this 1 - 2 times this, it will just stretch out this distribution to take take give me numbers between -1 and +1, that is all that is going on.

And then I will allow n to go from to all the way up to 500 and that is it, basically, I keep on adding this and that will generate for me a random walk with this distribution and then I will go ahead and do a ListPlot.

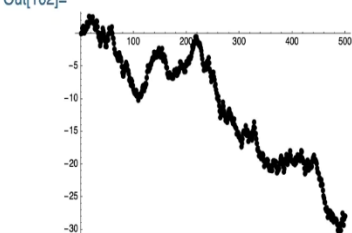
(Refer Slide Time: 7:25)

• Check if the key result $\langle m^2 \rangle = \alpha N$ is still robust by numerically generating many samples of random walks, and averaging over them.
• Analytically compute α . Check by explicit plotting if your numerical data support this expectation.


Solution

```
In[100]:=  
a = {0};  
Do[AppendTo[a, a[[n - 1]] + 1 - 2 RandomReal[]], {n, 2, 500}]  
ListPlot[a, Joined -> True]
```

Out[102]=



Slide 2 of 3

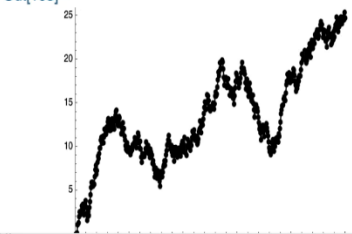


• Check if the key result $\langle m^2 \rangle = \alpha N$ is still robust by numerically generating many samples of random walks, and averaging over them.
• Analytically compute α . Check by explicit plotting if your numerical data support this expectation.


Solution

```
In[103]:=  
a = {0};  
Do[AppendTo[a, a[[n - 1]] + 1 - 2 RandomReal[]], {n, 2, 500}]  
ListPlot[a, Joined -> True]
```

Out[105]=



Slide 2 of 3



So, if I do it once, it looks like this, if I do it a second time it looks like this So, now you see that it is, it is a little more smooth than before. Earlier, we got very very jagged motion, because every step was of the same size, it could only vary in terms of whether you went right or to the left, but now there is also variation in terms of the size of the length at each point.

So, you can play with this, you can generate a large number of these and keep on visualizing them. I mean, these kinds of things are also done in the context of polymers. You can think of

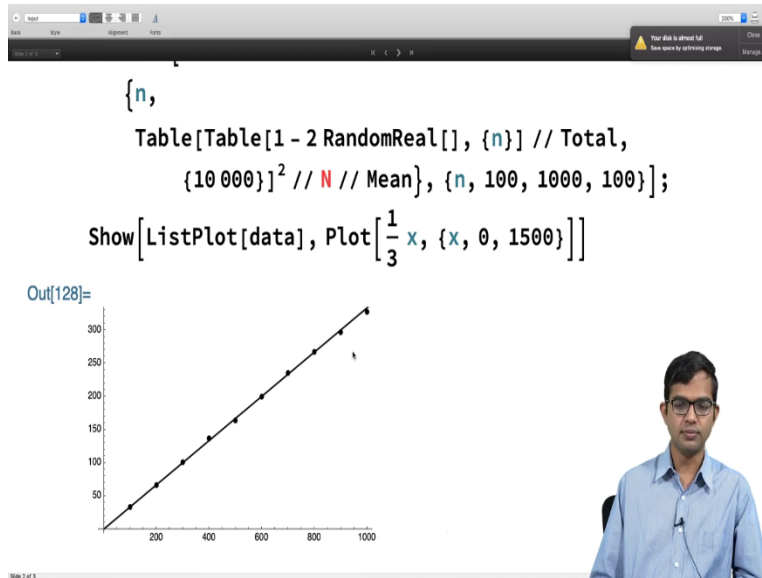
these individual steps as monomers and you are creating a polymer of a certain size, you know, with where the number of monomers is given to you and you can ask what is the typical distance between the between the origin or the starting point of your polymer and endpoint? And typically it will go as \sqrt{n} like we have seen that.

Now, the question is, is that true also for an arbitrary distribution of these individual random variables involved? And so the answer is yes, but let us verify that. And that has to do with the connection to the central limit theorem; we will become clear in a moment. So, in order to do this, what do I do? I will make a table of these random walks, so like I did before, I will generate not just, you know, this AppendTo, I will abandon this approach and instead simply generate a table of these random numbers which are just $1 - 2\text{RandomReal}$ and n of these and then sum them,.

That is completely identical to doing this and this total is really the sum. You see, it is like this S , so there is indeed a sum involved here and when you have the sum of a large number of random variables underlying this is the central limit theorem and that is where the generality of these these results actually come from the central limit theorem.

So, I am going to consider 10000 such samples and then square this stuff and take its mean, mean of m^2 and my claim is that this is going to be linear in the end and I need to find out the α corresponding to this and is there a way to argue for this analytically. So, I am claiming that in this case $\alpha = 1/3$ and I will tell you in a moment why this is the case as it starts doing it, there you go.

(Refer Slide Time: 9:54)



Indeed, it is perfectly linear, so the graph agrees excellently with the numerical data of the random walks and its slope is also one third. Why is the slope one third? Why is $\alpha = 1/3$ and so in order to see this, you have to go back to the central limit theorem. It tells you that, σ^2 for your distribution of your sum is given by the sum of all these σ^2 and if you just compute the variance of the uniform distribution from -1 to 1, how would you do this?

You will have to do $\int x^2 dx$ and it will become $x^3/3$ and then you should do the integral carefully and take the limits from - to +1, you can check that, you will get 1/3, the slope of 1/3 comes simply from doing the integration and using this statement of the central limit theorem.

So, that is what is the third part of this exercise, analytically compute alpha, check by explicit plotting if your numerical data supports the expectation. Indeed, it does that. What would be the distribution of the final position you would expect in this case? So, distribution is also given to us by the central limit theorem. It tells you not only that the mean of this sum is going to be the

sum of these means and variance is going to be the sum of the variances, but in fact it tells you more.

That is where the power of the central limit theorem comes from. It does not matter what the random variables individual random variables, what distributions they come from, if you add a large number of them, together the sum is going to become a Gaussian. That is the beauty and power of this central limit theorem.

So, what I do is, I choose nMax to be a 1000, then I will choose my binsize appropriately, you can play with this, and then I create a histogram of this table of 1 - 2RandomReal nMax and then I have also totaled this here, so it is a sum. I have created a large number of these random variables and created the sum and lots of sums and I am histogramming them all together, this is the code.

(Refer Slide Time: 12:15)

What would be the distribution of the final position you would expect in this case? Test it numerically.

Solution

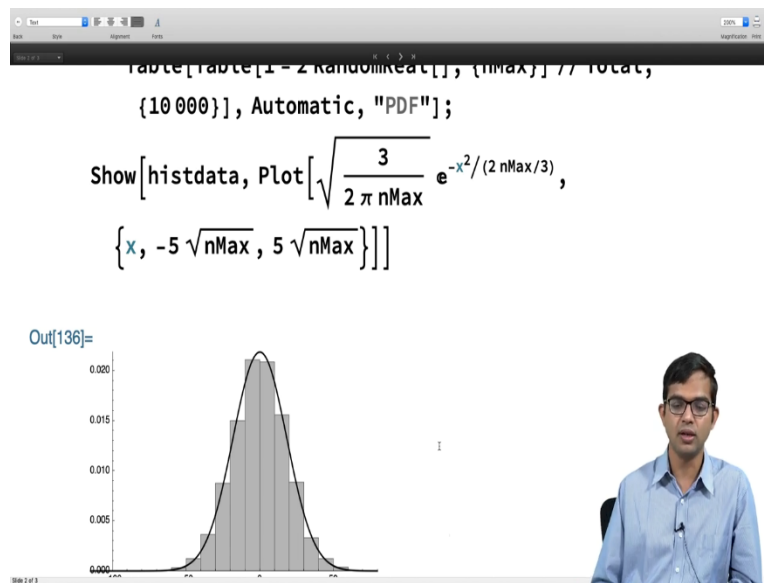
```
nMax = 1000;  
binsize = 10;  
histdata =  
Histogram[  
  Table[Table[1 - 2 RandomReal[], {nMax}] // Total,  
    {10000}], Automatic, "PDF"];  
Show[histdata, Plot[Sqrt[ $\frac{3}{2 \pi nMax}$ ] e-x2/(2 nMax/3),  
  {x, -5 Sqrt[nMax], 5 Sqrt[nMax]}]];
```

You should look at the code for a moment after you have tried it yourself. And then convince yourself. I am operating this histogram in PDF mode, this is something that you have to be careful about, look up the documentation, you can operate it in other modes as well, so there are density peak and so on.

This is a probability distribution function mode, so if I run this, then I will make a comparison between this numerically histogrammed data with the analytical expectation which from the central limit theorem, which tells me that it is a Gaussian whose variance comes in here. It is $1/\sqrt{2\pi\sigma}$ and the σ in this case is just $n\text{Max}/3$. Where is that coming from? It comes from the fact that so $2 n\text{Max}/3$, so $n\text{Max}/3$ is the variance, how do you get that?

You get it from this $1/3$, which comes in exactly from here, so it is $N/3$, so that is what is called $n\text{Max}$ here, so I am just plugging in the the result which you already got from the integration. And now let us check this and I am allowing my x to go from minus $-5\sqrt{n\text{Max}}$ to $5\sqrt{n\text{Max}}$. So, let us see what it looks like.

(Refer Slide Time: 13:42)



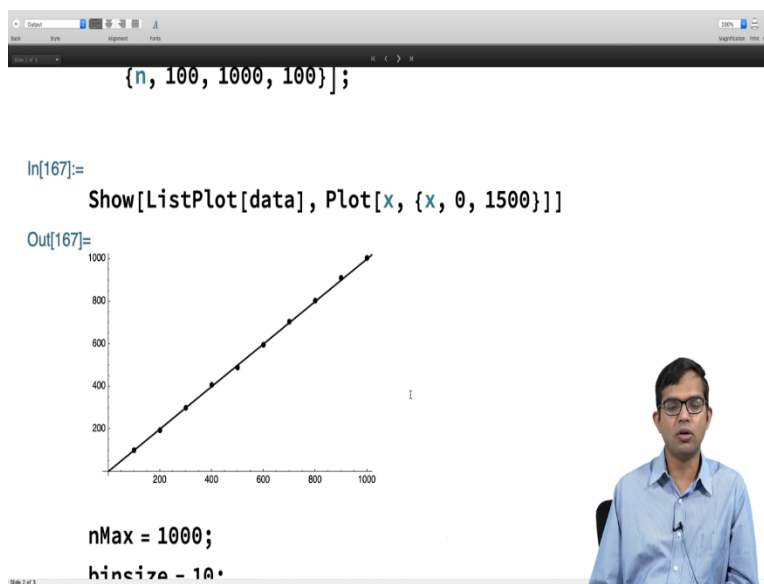
There we go. So we see that indeed there is excellent agreement between the expectation from the central limit theorem and our directly numerically generated data. And then the final thing here is, to play with other kinds of distribution. So, now we are free to check the generality of the central limit theorem, you do not have to restrict yourself to uniform distribution, you can generate other kinds of distribution.

Then see how large number of random variables has to be before it begins to respect the central limit theorem. So try out crazy distributions like exponential freefalling distribution or you know some other kind of distribution and see that the sum is still going to give you a Gaussian.

So, I have here one one another variation of this, which is to choose a normal distribution. I have basically the same kind of thing, but now AppendTo instead of just doing RandomReal, I have RandomVariate of normal distribution and everything else is the same as before. So, this is just to see that you know you get a different kind of plots that looks slightly different, but they are also random walks, so it is going to be every time you try it out, it is going to look different, but by and large, you can see that the distance between the origin and the last point will go as square root of the total number.

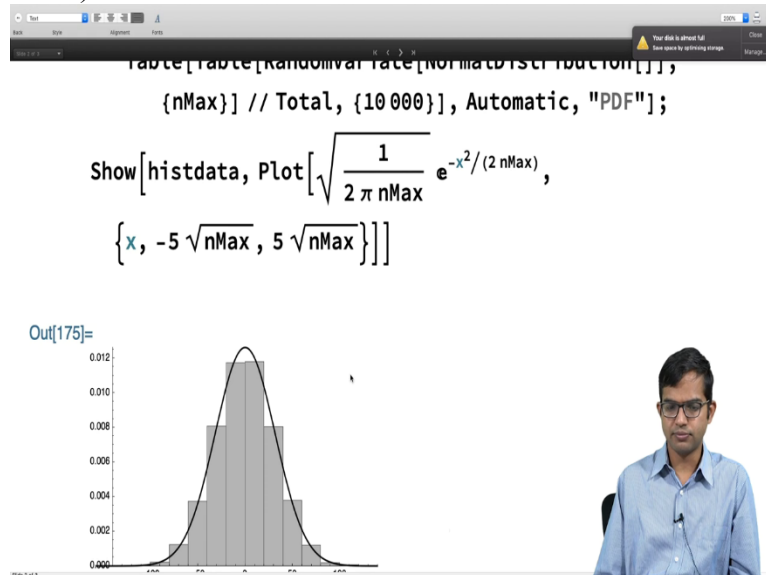
It is also not evident from here, but for that we will do the next test. And then I have basically used the same code as before, except that now I have a RandomVariate of normal distribution where I had is $1 - 2\text{RandomReal}$ and then, I will show both this ListPlot and the plot together on the same graph, there you go and so once again, the data sit on top of each other, so there is excellent agreement of this data as well and there is with this expectation from the central limit theorem.

(Refer Slide Time 15:43)



So, once again, you can go ahead and so I claim that this part is not going to change, because it is still, So, the only thing that has changed is σ^2 , now it is just nMax, it is not, there is no one third here, so that part has changed, but otherwise it is still a Gaussian distribution and you can check this,

(Refer Slide Time 16:10)



Indeed the two of them agree very well. So, this discussion was to give you a quick flavor or to just state the central limit theorem and give you an idea of how to play with random variables and develop a feeling for this theorem. So, there is one more small topic which I want to cover in this context and so the question is why is you know why is diffusive motion so common. One way of thinking about this is, in terms of the central limit theorem and another is the following argument.

(Refer Slide Time: 16:45)

A general argument for the scaling of net displacement.

The law of diffusive scaling (Net displacement proportional to the square root of the time traversed) is very general, and seems to hold in a wide variety of contexts. In particular it does not depend on dimensionality. Here is an argument why that happens. Consider a random walk in which each of the steps is a vector. Let us call them $\vec{r}_1, \vec{r}_2, \vec{r}_3, \dots, \vec{r}_N$. Since every direction is assumed to be equally likely the average of each of the vectors must be zero. That is

$$\langle \vec{r}_j \rangle = 0. \quad (1)$$


We can take the typical length of each vector to be a . This information appears in the mean of the square of each vector. That is

$$\langle \vec{r}_j \cdot \vec{r}_j \rangle = a^2. \quad (2)$$

The net displacement is then given by

$$\vec{r} = \vec{r}_1 + \vec{r}_2 + \vec{r}_3 + \dots + \vec{r}_N. \quad (3)$$

The average of the net displacement is also zero.

$$\langle \vec{r} \rangle = \langle \vec{r}_1 \rangle + \langle \vec{r}_2 \rangle + \langle \vec{r}_3 \rangle + \dots + \langle \vec{r}_N \rangle = 0 + 0 + 0 + \dots + 0 = 0.$$


So, this the law of diffusive scaling is very very general and it comes about not only in the context of the random walker, but in whatever dimension you are in, it holds. So, why does it hold in arbitrary dimension? So, this this argument, I find, is actually quite neat to see why this comes about.

So, consider a random walk in which each of the step as a vector, do not think of it as going to the right or to the left, all directions are completely open, let us say. So, let us call these vectors as r_1, r_2, r_3 , all the way up to n . Since, every direction is assumed to be equally likely, so the average of each of these vectors has to be zero. So, given it is simply a sort of directional average, you can think of this as. It does not matter what a magnitude of this individual vector typically is, but since every direction is equally likely, the average of each of these vectors is zero, it is a vectorial average.

So, we can take the typical length of each vector to be a and so this information can be compactly put into this equation, $\langle \vec{r}_j \cdot \vec{r}_j \rangle$, the dot product of a vector with itself is a measure of its the square of its distance or of its size and so, this the typical length we will take it to be a^2 so now think about this vectorial displacement.

So, if you add n such vectors, $r_1 + r_2 + r_3$, all the way up to n , so the average of this net displacement is, of course, zero because the average of each of these vectors is individually zero,

therefore the sum of all of these is also zero, but if you consider the average of the square of the net displacement and other words, if you take the dot product of this sum with itself, so this is a little bit like the central limit theorem, but here this is much less abstract, this is very simple set of ideas involved and just using the notion of vectors.

So, if you do $r \cdot r$, so you will see that there are all these terms which involve vectors multiplying with themselves, $r_1 \cdot r_1$, $r_2 \cdot r_2$ and so on and then you have all these cross terms. These cross terms is where now now comes the key argument. So, basically you argue that this random walk is completely memory free, so every step simply does not care about what any other step is doing.

At a given point, you go to the next point and then you have a fresh decision to make and you take any direction and within the distribution, the size of the step is completely open to you, and therefore, you can say the $\langle r_i \cdot r_j \rangle$ is actually nothing but the $\langle r_j \rangle \cdot \langle r_j \rangle$, so this is the key point.

If these two are completely independent, that means that this the the average of the product is equal to the product of the average and each of these averages is separately zero because every direction is completely equivalent and therefore, the $\langle r_i \cdot r_j \rangle = 0$. So, they are completely uncorrelated and so that is why, although you have a large number of sums, there is these these n terms, but there is also $n(n-1)/2$, other terms in this in this summation here.

All of that will just simply go to zero and then finally you are left with simply $\langle r \cdot r \rangle$ is just equal to these sums, each of these typically has the same length, they \bar{a}^2 , \bar{a}^2 , \bar{a}^2 all add up and then you get $n\bar{a}^2$. So, these are very nice beautiful arguments which and which immediately gives you the statement of the diffusive motion, basically.

What it tells you is that, if you take n steps, you are typically only going as \sqrt{n} you will have to take the square root of this overall stuff, then you will have $\sqrt{n}a$, you have taken n steps, but you have only moved \sqrt{n} times each individual length stepping one, so that is nothing but the statement of the diffusive motion coming from some very very simple direct arguments. Alright, thank you.