Lec 18.7
Motivation for
Sampling

Refer Slide Time :(0: 14)

$$\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}} = \mathbb{E}_{p(H|V)}[v_i h_j] - \mathbb{E}_{p(V,H)}[v_i h_j]$$

- The trick is to approximate the sum by using a few samples instead of an exponential number of samples
- We will try to understand this with the help of an analogy

So, the trick is basically to approximate the summation, by a few samples, instead of an exponential number of samples. Right? And we'll, try to understand with the help of a simple analogy, I'm sure most of you know this, but I'll still, just go over this analogy.

Refer Slide Time :(0: 27)

- Suppose you live in a city which has a population of 10M and you want to compute the average weight of this population
- You can think of X as a random variable which denotes a person
- The value assigned to this random variable can be any person from your population
- For each person you have an associated value denoted by $weight(X)$
- You are then interested in computing the expected value of $weight(X)$ as shown on the RHS

$$\mathbb{E}[weight(X)] = \sum_{(x \in P)} p(x) weight(x)$$

- Of course, it is going to be hard to get the weights of every person in the population and hence in practice we approximate the above sum by sampling only few subjects from the population (say 10000)

$$\mathbb{E}[weight(X)] \approx \sum_{x \in P[:10000]} [p(x) weight(x)]$$

- Further, you assume that $P(X) = \frac{1}{N} = \frac{1}{10K}$, i.e., every person in your population is equally likely

$$\mathbb{E}[weight(X)] \approx \frac{\sum_{x \in Persons[:10000]} [weight(x)]}{10^4}$$

So, suppose you live in a city which has, a population of10 million and you want to compute the average, weight of this population. What will you do? You will go and ask everyone, in the population to give their weights to you, most of them. So, you can think of X, as a random variable, which denotes a person. Right? So, of all the10 million values that, X can take, I mean X and take any one of these 10 million values that you have, 10 million people that you have in your population. Now, with every random variable, you also, have the weight associated with that. So, technically what are you interested in computing? Have X and you have weight of X, I want to compute the average weight of the population. So, what's the technical term for that? Expectation, what expectation I need to compute? Expectation of the weight and this is how you'll compute it? Now, what's the number of terms in this summation? 10million. Right? This is all the people in your population. Okay? But, this is going to be hard. So, in

practice what do we do? We just take some random 10,000 people, compute their weights and from that we compute the expectation. Okay? But, we do while doing that. Right? We actually, do a lot of stuff which we don't actually realize. So, let's see, what is it that we actually do? So, this is exactly, what I'm going to do? I'm going to approximate that expectation, by just Some 10,000 people, from all the people that, I have in my population. Okay? Now, what are the assumptions that I'm making while doing. So, Okay? That's fair enough, in fact, what is this formula going to simplify further to? It's going to simplify to this, everyone agrees with that. So, what's the assumption that I've made here? How many forget that? We make an implicit assumption without actually realizing it, what is the assumption that we have made? All samples from our population, are equally likely, in some cases this makes sense. Right? So, if actually if you look at it as a population point of view, unless like the population is really divided that, there are very few people staying in certain areas and most people stay in one area, then there is going to be, less likelihood of, eye sampling from this, smaller population region. Right? But, if I just assume that, along the city people are equally distributed, then I can just go and pickup, random samples, I can just keep, walking along the city and just take10,000 random samples and they are a representative of the population. So, in that sense, I can assume: that all people in my population are equally, likely is that fair. Okay? You see a problem with this in many other cases, can you give me one example, where this is a problem, where you have a, space, a universal set, wait what's the universal set here? All people in our population and everything in this universal set, is not uniform, it's not equally likely to get any event from this universe aside, what's the event here? Any person that we want to pick, yeah! Can you give an example, from I am sure most of you did not hear that. So, I am repeating the question, from something that you have learnt recently, or from the discussion that we have been having over the past, couple of four three four lectures. Give me an example, of a space, in which you do not, expect every, outcome from that space to be equally likely images. Right? Okay? So, we will just go to that. Right? So, remember that, when you're computing these expectations you make this assumption that all samples are equally, likely which may not be the case. Okay? And we will see that, if that is not the case, what's the right way of doing it? Yeah! But, you are able to solve draw these samples, because you are able you're making this assumption that every sample is equally, likely. Right? I'll just continue with this discussion and then I will give you an example, where this is, what I want to say? Maybe I'm not saying it correctly, but, it should become clear. Right?

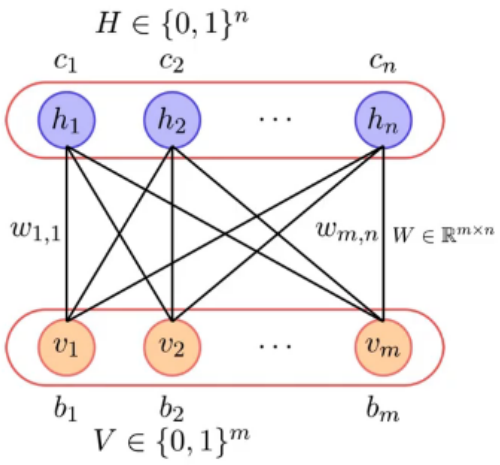
Refer Slide Time :(4: 22)

$$\mathbb{E}[X] = \sum_{(x \in P)} xp(x)$$

- This looks easy, why can't we do the same for our task ?
- Why can't we simply approximate the sum by using some samples?
- What does that mean? It means that instead of considering all possible values of $\{v, h\} \in 2^{m+n}$ let us just consider some samples from this population
- Analogy: Earlier we had 10M samples in the population from which we drew $10K$ samples, now we have 2^{m+n} samples in the population from which we need to draw a reasonable number of samples
- Why is this not straightforward? Let us see!

So, this looks easy. Right? Why can't we do it for our task? Our task was also that, we had these all possible values of V comma H, just as we had all 10 million possibilities for the people, this does V sample any 10 K from there, why can't we approximate this all possible values of V comma H, by any k values of V comma H or any reasonable values. Right? Let's make it 1 million for that matter, why can't we do that? It's, just that's exactly the analogy, you had10 million samples in your space, is approximated by H, small number of samples? You have to raise to M plus and possible values, just draw a reasonable number of values from there, say 1million or 2 million and just approximate the summation by that. Okay? That looks easy. Right? That is what we should do? This is what I'm, trying to tell you and instead of all these two days to M plus n values, let's consider some samples, then let's be vague about the number of samples, let's be generous. So, it will assume 1 million samples, because two days to M plus n is a really large space. So, even if you draw 1million, is going to be a very small number of samples, from that space. Okay? Okay? Why is this not straightforward?

Refer Slide Time :(5: 28)

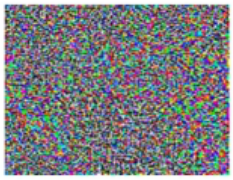


- For simplicity, first let us just focus on the visible variables ($V \in 2^m$) and let us see what it means to draw samples from $P(V)$
- Well, we know that $V = v_1, v_2, \ldots, v_m$ where each $v_i \in \{0,1\}$
- Suppose we decide to approximate the sum by $10K$ samples instead of the full 2^m samples
- It is easy to create these samples by assigning values to each v_i
- For example, $V = 11111 \ldots 11111, V = 00000 \ldots 0000, V = 00110011 \ldots 00110011, \ldots V = 0101 \ldots 0101$ are all samples from this population
- So which samples do we consider ?

And for simplicity, what we will do is we'll just focus on the visible variables. Right? So, let's assume: that we are only interested in summation over all V's, we don't have a term which has a summation over V comma H. So, we just have to raise to M terms and not to raise to M cross n terms, this is just for the sake of simplicity and even with that, we can really see that, it is slightly hard to do that. So, what does it mean, to draw samples from P of V, we know that, that VI is belong to 0 comma 1 and every V vector belongs to 2 raised to M. Now, suppose I decide: that I want to approximate this sum by 10 K values. Right? Can you construct these 10 K values, just as in the case of the population, you just took 10 K people. Right? Because there, was just there and you just took them, in the case of this, in the case of visible variables, if I asked you to give me 10 K samples, can you give that 10 K samples to me? How will you come up with these 10 K samples? How will you come up, if I ask you to give me one valid configuration of V, give me an answer, all zeroes, all ones, some 0, some ones. So, it's not very easy, very difficult to construct samples from this space. Right? You just need to decide, I'll set some values to 1, some values to 0 and all these things which I've written here, are valid samples from this space. I can just construct any of these samples, it's not very hard, to construct these 10 K samples, I'll just take these 10k samples and then approximate the sum using this 10 K samples, what's wrong with this? What are we again assuming here? All samples are equally, likely that's why I'm saying that I just have this large page, I'll just pick one from here, one from there everything. Right? Everything is uniformly possible. Actually if you are considering images. Okay? And let's assume there are no blank images that you ever see, should I've actually drawn this sample? No this was a sample which's zero probability and I'm actually drawn it, that's the problem. Right? As compared to the population example, that's the problem that we have that, all samples. In this space are not equally likely. So, when we are constructing these10 K samples, which 10 K from thepossible to raise to M plus n should we consider, that's the question that we need to answer, is that here, is everyone fine with that. Okay?

Refer Slide Time :(7: 45)

Likely

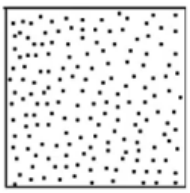


Unlikely

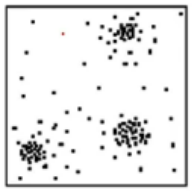
- Well, that's where the catch is!
- Unlike, our population analogy, here we cannot assume that every sample is equally likely
- Why? (Hint: consider the case that visible variables correspond to pixels from natural images)
- Clearly some images are more likely than the others!
- Hence, we cannot assume that all samples from the population ($V \in 2^{m+n}$ are equally likely

Now, that's where the catch is! Unlike, our population analogy, all these are not equally, likely why? Because if you consider the case of images, this the top image, is more likely to be seen, the bottom image is not likely to be seen. Right? If you are looking at images of natural scenes or yeah! ,Or even animals or whatever, the bottom image is definitely not likely and MySpace is, so high dimensional that off that entire space, actually there are very, very few samples which are really legitimate, all the other samples are noise. Now, if I am going to draw everything uniformly, I'm going to get a lot of these, noisy samples, which do not, actually belong to my distribution, I'll just not be able to get only those images, which are skies or Birds or whatever, all of you get this that to raise to M plus n, is a very, very, very high dimensional space, it has many, many, many points, of which only a very, very, very few points, are the points that we are interested in and while computing this expectation, we need to focus on those points, because all the other points have a zero probability, does that make sense. Okay?

Refer Slide Time :(8: 58)



Uniform distribution



Multimodal distribution

- Let us see this in more detail
- In our analogy, every person was equally likely so we could just sample people uniformly randomly
- However, now if we sample people uniformly randomly then we will not get the true picture of the expected value
- We need to draw more samples from the high probability region and fewer samples from the low probability region
- In other words each sample needs to be drawn in proportion to its p not uniformly

So, this is the difference Right? So, earlier you had this uniform distribution. But, now you have some kind of a multi-modal distribution or just a skewed distribution. Where only some regions have, high density samples and everything else is not likely. Right? There is nothing likely from this region which I have just painted, in our distribution. So, if I draw samples from there, I can actually tell you what this point is that, this is some in this, if I think of this as a square grid, then this is some three comma one or some point it, I can tell you what this point is, I can always sample this point. But, when I sample this point, I am doing something which is wrong, because this is not a good sample. Okay? Okay? So, we need to draw samples, from the high, priority region and fewer samples, from the lower probability region. So, whatever sample we draw, should be proportional to the probability, this was not a problem earlier, because everything was equally, likely. Okay?

Refer Slide Time :(9: 51)

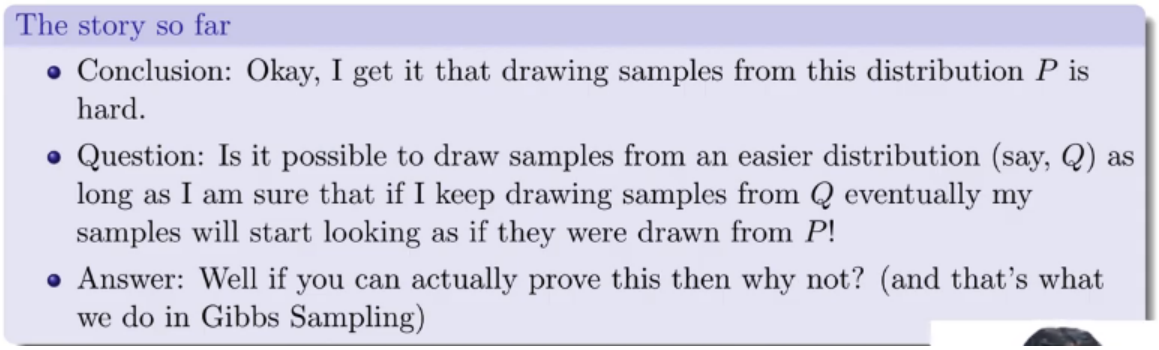
$$\frac{\partial \mathscr{L}(\theta|V)}{\partial w_{ij}} = \mathbb{E}_{p(H|V)}[v_i h_j] - \mathbb{E}_{p(V,H)}[v_i h_j]$$

$$Z = \sum_V \sum_H \Big(\prod_i \prod_j \phi_{ij}(v_i, h_j) \prod_i \psi_i(v_i) \prod_j \xi_j(h_j) \Big)$$

- That is where the problem lies!
- To draw a sample (V, H), we need to know its probability $P(V, H)$
- And of course, we also need this $P(V, H)$to compute the expectation
- But, unfortunately computing $P(V, H)$ is intractable because of the partition function Z
- Hence, approximating the summation by using a few samples is not straightforward! (or rather drawing a few samples from the distribution is hard!)

So, that is where the problem lies. Now, if I want to pick up a sample, according to its probability, what do I need to compute first? The probability, can I compute the probability, to compute the probability, what do I need to do, what's the term that will, I will divide with, Z and Z by itself is, intractable because again I have an exponential term number of terms in it. Right? So, I have this dilemma. Right? I want to draw samples, according to a probability distribution. But, I cannot actually, compute the probability distribution, because computing the probability of any sample itself, is exponentially complex, is that fine, you see the problem here, some to draw these samples, we need to compute P of e comma H and also, to compute the expectation which is saying the same thing we need to know, the probability of V comma H. Right? Because you have this in the expectation and this is hard to do, because you have the partition function and the partition, function has an exponential number of terms. Okay? So, hence unlike our population cases, approximating this sum, by a fewer samples, is not straightforward, because we don't know, how to draw these fewer samples, to draw these samples, we again go back in circles and we realize that we again need to compute the probability. Okay? So, so just go back and think about this.

Refer Slide Time :(11: 18)



And this is, what the conclusion is. Right? I get it samples, from the distribution P is hard, what's the distribution P? P of e comma H, I cannot draw samples, from this distribution. So, the question that we are going to ask and try to answer in the next lecture or what a few lectures, is that, is it possible to draw samples, from another distribution, Q which is and quote, unquote easy distribution, as long as I am sure that, if I am drawing samples, from this distribution, they are almost as if they came from my original distribution, a very convoluted question. But, do you get the motivation for that, P is a hard distribution, I cannot compute, P because, I need Z which is interact able, what my proposal is? What if I give you a distribution Q, this does not have this, hardness in computational terms, I can easily compute that and I can give you some guarantee that P and Q are very, similar that means, if I am drawing samples from Q, if I'm drawing visible variables from Q, if I am drawingthe vector V from Q, if I can give you an argument that those these are actually very similar, to what I could have gotten from P. Okay? Then it's, Okay? To use Q, does that make sense, because Q is easier to use, Q is easier to deal. So, that's what we are going to do in the next lecture, when we talk about, keep sampling that's the overall idea: that if you have a P and if you have a Q, such that P is very hard to draw from, but, Q is very easy to draw from, can you set up a Q, which has two requirements, one is it's easy and the other is that samples from Q, after a while, start

looking very similar to samples from P, that's the motivation behind Gibbs sampling and that’s what we are going to do in the next lecture. Okay? So, go back and revise today’s, discussion in particular be very comfortable, with the idea of why we need to do sampling and why it is not straightforward in the case of RBMS. Okay? And not just RBMS actually, what you have seen is applicable to all graphical models, which have these exponential of number of terms, because all the graphical models will have this partition function, which you have to normalize over or marginalize over and whenever that is hard to do, you have to resort to some kind of sampling metro’s. Right? Now that's what we will do in the next lecture. Thank you.