Module 19.7: Motivation for Sampling

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

- The trick is to approximate the sum by using a few samples instead of an exponential number of samples

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

- 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

- Suppose you live in a city which has a population of 10M and you want to compute the average weight of this population

- 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

- 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

- 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 $weight(X)$

- 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

- 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

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

- 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)

- 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)

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

- 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)

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

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

- 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)

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

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

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

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

- This looks easy, why can't we do the same for our task ?

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

- 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?

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

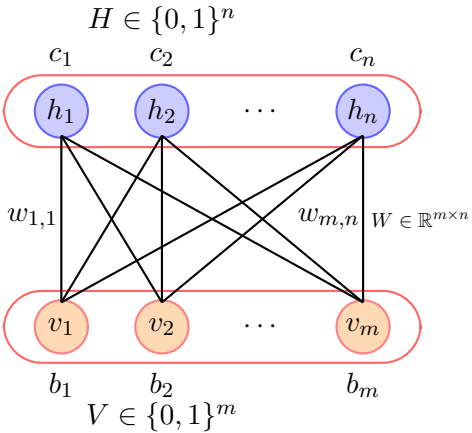
- This looks easy, why can't we do the same for our task ?
- Why can't we simply approximate the sum by using some samples?
- What does that mean? It means that instead of considering all possible values of $\{v, h\} \in 2^{m+n}$ let us just consider some samples from this population

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

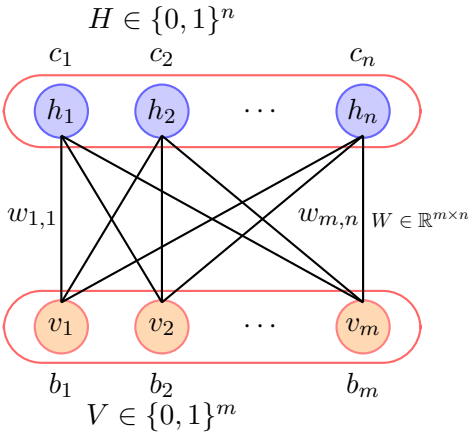
- This looks easy, why can't we do the same for our task ?
- Why can't we simply approximate the sum by using some samples?
- What does that mean? It means that instead of considering all possible values of $\{v, h\} \in 2^{m+n}$ let us just consider some samples from this population
- Analogy: Earlier 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

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

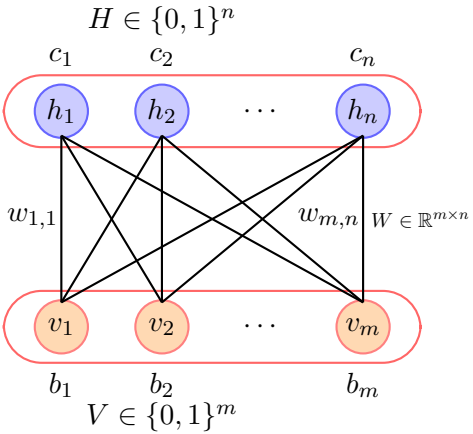
- This looks easy, why can't we do the same for our task ?
- Why can't we simply approximate the sum by using some samples?
- What does that mean? It means that instead of considering all possible values of $\{v, h\} \in 2^{m+n}$ let us just consider some samples from this population
- Analogy: Earlier 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!



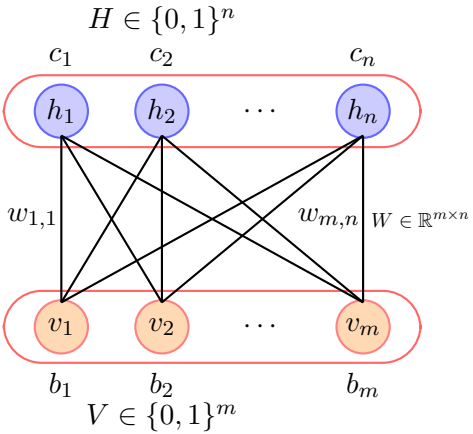
- For simplicity, first let us just focus on the visible variables ($V \in 2^m$) and let us see what it means to draw samples from $P(V)$



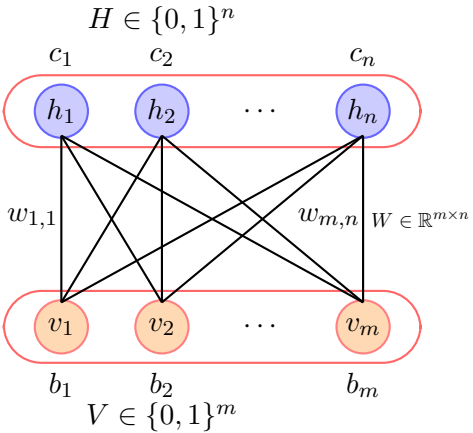
- For simplicity, first let us just focus on the visible variables ($V \in 2^m$) and let us see what it means to draw samples from $P(V)$
- Well, we know that $V = v_1, v_2, \ldots, v_m$ where each $v_i \in \{0,1\}$



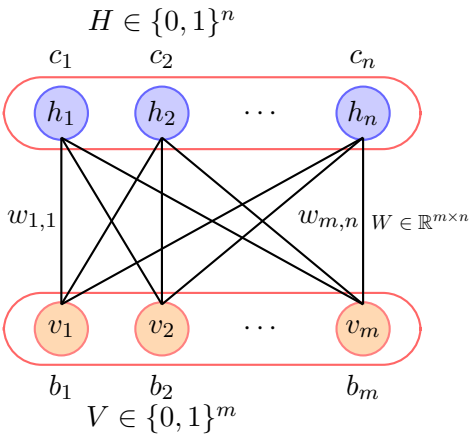
- For simplicity, first let us just focus on the visible variables ($V \in 2^m$) and let us see what it means to draw samples from $P(V)$
- Well, we know that $V = v_1, v_2, \ldots, v_m$ where each $v_i \in \{0, 1\}$
- Suppose we decide to approximate the sum by $10K$ samples instead of the full 2^m samples



- 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 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



- 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 ?

- Well, that's where the catch is!

- Well, that's where the catch is!
- Unlike, our population analogy, here we cannot assume that every sample is equally likely

- Well, that's where the catch is!
- Unlike, our population analogy, here we cannot assume that every sample is equally likely
- Why?

- 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)

Likely

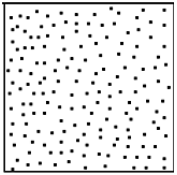
Unlikely

- 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!

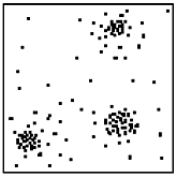
Likely

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 ($V \in 2^m$) are equally likely

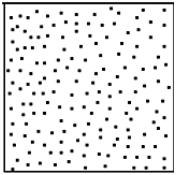


Uniform distribution

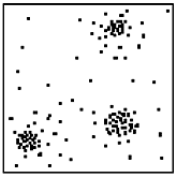


Multimodal distribution

- Let us see this in more detail

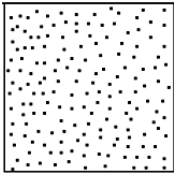


Uniform distribution

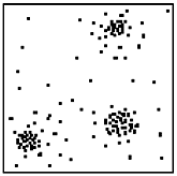


Multimodal distribution

- Let us see this in more detail
- In our analogy, every person was equally likely so we could just sample people uniformly randomly

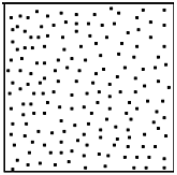


Uniform distribution

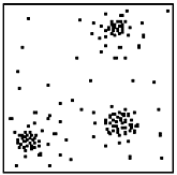


Multimodal distribution

- 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

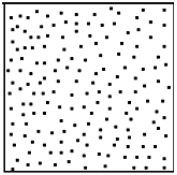


Uniform distribution

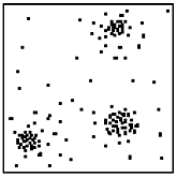


Multimodal distribution

- 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



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 probability and not uniformly

- That is where the problem lies!

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

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

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

- That is where the problem lies!
- To draw a sample (V, H), we need to know its probability $P(V, H)$

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

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

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

- 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

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

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

- 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

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

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

- 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!)

The story so far

- Conclusion: Okay, I get it that drawing samples from this distribution P is hard.

The story so far

- Conclusion: Okay, I get it that drawing samples from this distribution P is hard.
- Question: Is it possible to draw samples from an easier distribution (say, Q) as long as I am sure that if I keep drawing samples from Q eventually my samples will start looking as if they were drawn from P!

The story so far

- Conclusion: Okay, I get it that drawing samples from this distribution P is hard.
- Question: Is it possible to draw samples from an easier distribution (say, Q) as long as I am sure that if I keep drawing samples from Q eventually my samples will start looking as if they were drawn from P!
- Answer: Well if you can actually prove this then why not? (and that's what we do in Gibbs Sampling)