Module 19.1: Using joint distributions for classification and sampling

Now that we have some understanding of joint probability distributions and efficient ways of representing them, let us see some more practical examples where we can use these joint distributions

- Consider a movie critic who writes reviews for movies

- **M1:** An unexpected and necessary masterpiece
- **M2:** Delightfully merged information and comedy
- **M3:** Director's first true masterpiece
- **M4:** Sci-fi perfection,truly mesmerizing film.
- **M5:** Waste of time and money
- **M6:** Best Lame Historical Movie Ever

- Consider a movie critic who writes reviews for movies
- For simplicity let us assume that he always writes reviews containing a maximum of 5 words

- **M1:** An unexpected and necessary masterpiece
- **M2:** Delightfully merged information and comedy
- **M3:** Director's first true masterpiece
- **M4:** Sci-fi perfection,truly mesmerizing film.
- **M5:** Waste of time and money
- **M6:** Best Lame Historical Movie Ever

- Consider a movie critic who writes reviews for movies
- For simplicity let us assume that he always writes reviews containing a maximum of 5 words
- Further, let us assume that there are a total of 50 words in his vocabulary

- **M1:** An unexpected and necessary masterpiece
- **M2:** Delightfully merged information and comedy
- **M3:** Director's first true masterpiece
- **M4:** Sci-fi perfection,truly mesmerizing film.
- **M5:** Waste of time and money
- **M6:** Best Lame Historical Movie Ever

- Consider a movie critic who writes reviews for movies
- For simplicity let us assume that he always writes reviews containing a maximum of 5 words
- Further, let us assume that there are a total of 50 words in his vocabulary
- Each of the 5 words in his review can be treated as a random variable which takes one of the 50 values

- **M1:** An unexpected and necessary masterpiece
- **M2:** Delightfully merged information and comedy
- **M3:** Director's first true masterpiece
- **M4:** Sci-fi perfection,truly mesmerizing film.
- **M5:** Waste of time and money
- **M6:** Best Lame Historical Movie Ever

- Consider a movie critic who writes reviews for movies
- For simplicity let us assume that he always writes reviews containing a maximum of 5 words
- Further, let us assume that there are a total of 50 words in his vocabulary
- Each of the 5 words in his review can be treated as a random variable which takes one of the 50 values
- Given many such reviews written by the reviewer we could learn the joint probability distribution

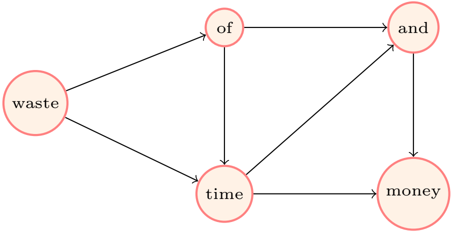
$$P(X_1, X_2, \ldots, X_5)$$

- **M1:** An unexpected and necessary masterpiece
- **M2:** Delightfully merged information and comedy
- **M3:** Director's first true masterpiece
- **M4:** Sci-fi perfection,truly mesmerizing film.
- **M5:** Waste of time and money
- **M6:** Best Lame Historical Movie Ever

- In fact, we can even think of a very simple factorization for this model

$$P(X_1, X_2, \ldots, X_5) = \prod P(X_i|X_{i-1}, X_{i-2})$$

- **M1:** An unexpected and necessary masterpiece
- **M2:** Delightfully merged information and comedy
- **M3:** Director's first true masterpiece
- **M4:** Sci-fi perfection,truly mesmerizing film.
- **M5:** Waste of time and money
- **M6:** Best Lame Historical Movie Ever

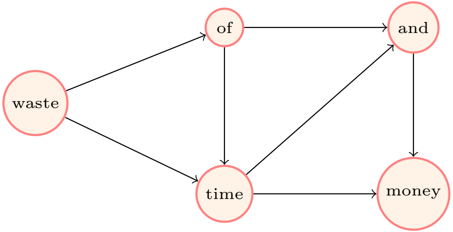


- In fact, we can even think of a very simple factorization for this model

$$P(X_1, X_2, \ldots, X_5) = \prod P(X_i | X_{i-1}, X_{i-2})$$

- In other words, we are assuming that the i-th word only depends on the previous 2 words and not anything before that

- **M1:** An unexpected and necessary masterpiece
- **M2:** Delightfully merged information and comedy
- **M3:** Director's first true masterpiece
- **M4:** Sci-fi perfection,truly mesmerizing film.
- **M5:** Waste of time and money
- **M6:** Best Lame Historical Movie Ever

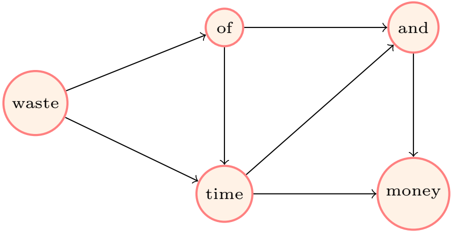


- In fact, we can even think of a very simple factorization for this model

$$P(X_1, X_2, \ldots, X_5) = \prod P(X_i | X_{i-1}, X_{i-2})$$

- In other words, we are assuming that the i-th word only depends on the previous 2 words and not anything before that
- Let us consider one such factor $P(X_i = time | X_{i-2} = waste, X_{i-1} = of)$

- **M1:** An unexpected and necessary masterpiece
- **M2:** Delightfully merged information and comedy
- **M3:** Director's first true masterpiece
- **M4:** Sci-fi perfection,truly mesmerizing film.
- **M5:** Waste of time and money
- **M6:** Best Lame Historical Movie Ever



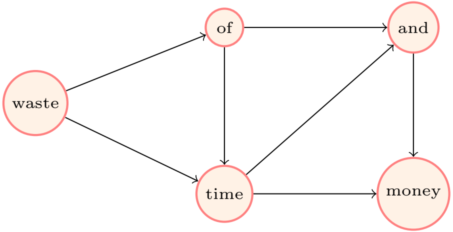
- In fact, we can even think of a very simple factorization for this model

$$P(X_1, X_2, \ldots, X_5) = \prod P(X_i | X_{i-1}, X_{i-2})$$

- In other words, we are assuming that the i-th word only depends on the previous 2 words and not anything before that
- Let us consider one such factor $P(X_i = time | X_{i-2} = waste, X_{i-1} = of)$
- We can estimate this as

$$\frac{count(\text{waste of time})}{count(\text{waste of})}$$

- **M1:** An unexpected and necessary masterpiece
- **M2:** Delightfully merged information and comedy
- **M3:** Director's first true masterpiece
- **M4:** Sci-fi perfection,truly mesmerizing film.
- **M5:** Waste of time and money
- **M6:** Best Lame Historical Movie Ever



- In fact, we can even think of a very simple factorization for this model

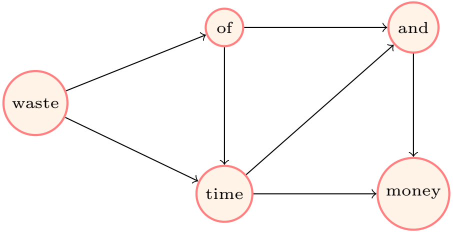
$$P(X_1, X_2, \ldots, X_5) = \prod P(X_i | X_{i-1}, X_{i-2})$$

- In other words, we are assuming that the i-th word only depends on the previous 2 words and not anything before that
- Let us consider one such factor $P(X_i = time | X_{i-2} = waste, X_{i-1} = of)$
- We can estimate this as

$$\frac{count(\text{waste of time})}{count(\text{waste of})}$$

- And the two counts mentioned above can be computed by going over all the reviews

- **M1:** An unexpected and necessary masterpiece
- **M2:** Delightfully merged information and comedy
- **M3:** Director's first true masterpiece
- **M4:** Sci-fi perfection,truly mesmerizing film.
- **M5:** Waste of time and money
- **M6:** Best Lame Historical Movie Ever



- In fact, we can even think of a very simple factorization for this model

$$P(X_1, X_2, \ldots, X_5) = \prod P(X_i | X_{i-1}, X_{i-2})$$

- In other words, we are assuming that the i-th word only depends on the previous 2 words and not anything before that
- Let us consider one such factor $P(X_i = time | X_{i-2} = waste, X_{i-1} = of)$
- We can estimate this as

$$\frac{count(\text{waste of time})}{count(\text{waste of})}$$

- And the two counts mentioned above can be computed by going over all the reviews
- We could similarly compute the probabilities of all such factors

w	$P(X_i = w\|,$ $X_{i-2} = more,$ $X_{i-1} = realistic)$	$P(X_i = w\|,$ $X_{i-2} = realistic,$ $X_{i-1} = than)$	$P(X_i = w\|$ $X_{i-2} = than,$ $X_{i-1} = real)$	$\ldots$
than	0.61	0.01	0.20	...
as	0.12	0.10	0.16	...
for	0.14	0.09	0.05	...
real	0.01	0.50	0.01	...
the	0.02	0.12	0.12	...
life	0.05	0.11	0.33	...

- Okay, so now what can we do with this joint distribution?

M7: More realistic than real life

w	$P(X_i = w\|, X_{i-2} = more, X_{i-1} = realistic)$	$P(X_i = w\|, X_{i-2} = realistic, X_{i-1} = than)$	$P(X_i = w\| X_{i-2} = than, X_{i-1} = real)$	$\ldots$
than	0.61	0.01	0.20	...
as	0.12	0.10	0.16	...
for	0.14	0.09	0.05	...
real	0.01	0.50	0.01	...
the	0.02	0.12	0.12	...
life	0.05	0.11	0.33	...

- Okay, so now what can we do with this joint distribution?
- Given a review, *classify* if this was written by the reviewer

M7: More realistic than real life

w	$P(X_i = w \mid, X_{i-2} = more, X_{i-1} = realistic)$	$P(X_i = w \mid, X_{i-2} = realistic, X_{i-1} = than)$	$P(X_i = w \mid X_{i-2} = than, X_{i-1} = real)$	...
than	0.61	0.01	0.20	...
as	0.12	0.10	0.16	...
for	0.14	0.09	0.05	...
real	0.01	0.50	0.01	...
the	0.02	0.12	0.12	...
life	0.05	0.11	0.33	...

$$P(M7) = P(X_1 = more).P(X_2 = realistic|X_1 = more).$$
$$P(X_3 = than|X_1 = more, X_2 = realistic).$$
$$P(X_4 = real|X_2 = realistic, X_3 = than).$$
$$P(X_5 = life|X_3 = than, X_4 = real)$$
$$= 0.2 \times 0.25 \times 0.61 \times 0.50 \times 0.33 = 0.005$$

- Okay, so now what can we do with this joint distribution?
- Given a review, *classify* if this was written by the reviewer

M7: More realistic than real life

w	$P(X_i = w \mid, X_{i-2} = more, X_{i-1} = realistic)$	$P(X_i = w \mid, X_{i-2} = realistic, X_{i-1} = than)$	$P(X_i = w \mid X_{i-2} = than, X_{i-1} = real)$	...
than	0.61	0.01	0.20	...
as	0.12	0.10	0.16	...
for	0.14	0.09	0.05	...
real	0.01	0.50	0.01	...
the	0.02	0.12	0.12	...
life	0.05	0.11	0.33	...

$$P(M7) = P(X_1 = more).P(X_2 = realistic|X_1 = more).$$
$$P(X_3 = than|X_1 = more, X_2 = realistic).$$
$$P(X_4 = real|X_2 = realistic, X_3 = than).$$
$$P(X_5 = life|X_3 = than, X_4 = real)$$
$$= 0.2 \times 0.25 \times 0.61 \times 0.50 \times 0.33 = 0.005$$

- Okay, so now what can we do with this joint distribution?
- Given a review, *classify* if this was written by the reviewer
- *Generate* new reviews which would look like reviews written by this reviewer

M7: More realistic than real life

w	$P(X_i = w \mid X_{i-2} = more, X_{i-1} = realistic)$	$P(X_i = w \mid X_{i-2} = realistic, X_{i-1} = than)$	$P(X_i = w \mid X_{i-2} = than, X_{i-1} = real)$	...
than	0.61	0.01	0.20	...
as	0.12	0.10	0.16	...
for	0.14	0.09	0.05	...
real	0.01	0.50	0.01	...
the	0.02	0.12	0.12	...
life	0.05	0.11	0.33	...

$$P(M7) = P(X_1 = more).P(X_2 = realistic|X_1 = more).$$
$$P(X_3 = than|X_1 = more, X_2 = realistic).$$
$$P(X_4 = real|X_2 = realistic, X_3 = than).$$
$$P(X_5 = life|X_3 = than, X_4 = real)$$
$$= 0.2 \times 0.25 \times 0.61 \times 0.50 \times 0.33 = 0.005$$

- Okay, so now what can we do with this joint distribution?
- Given a review, *classify* if this was written by the reviewer
- *Generate* new reviews which would look like reviews written by this reviewer
- How would you do this? By sampling from this distribution! What does that mean? Let us see!

w	$P(X_1 = w)$			
the	0.62			
movie	0.10			
amazing	0.01			
useless	0.01			
was	0.01			
⋮	⋮			

- How does the reviewer start his reviews (what is the first word that he chooses)?

w	$P(X_1 = w)$			
the	0.62			
movie	0.10			
amazing	0.01			
useless	0.01			
was	0.01			
⋮	⋮			

The

- How does the reviewer start his reviews (what is the first word that he chooses)?
- We could take the word which has the highest probability and put it as the first word in our review

w	$P(X_1 = w)$	$P(X_2 = w \mid, X_1 = the)$		
the	0.62	0.01		
movie	0.10	0.40		
amazing	0.01	0.22		
useless	0.01	0.20		
was	0.01	0.00		
⋮	⋮	⋮		

The movie

- How does the reviewer start his reviews (what is the first word that he chooses)?
- We could take the word which has the highest probability and put it as the first word in our review
- Having selected this what is the most likely second word that the reviewer uses?

w	$P(X_1 = w)$	$P(X_2 = w \mid, X_1 = the)$	$P(X_i = w \mid, X_{i-2} = the, X_{i-1} = movie)$	
the	0.62	0.01	0.01	
movie	0.10	0.40	0.01	
amazing	0.01	0.22	0.01	
useless	0.01	0.20	0.03	
was	0.01	0.00	0.60	
⋮	⋮	⋮	⋮	

The movie was

- How does the reviewer start his reviews (what is the first word that he chooses)?
- We could take the word which has the highest probability and put it as the first word in our review
- Having selected this what is the most likely second word that the reviewer uses?
- Having selected the first two words what is the most likely third word that the reviewer uses?

w	$P(X_1 = w)$	$P(X_2 = w \mid, X_1 = the)$	$P(X_i = w \mid, X_{i-2} = the, X_{i-1} = movie)$	...
the	0.62	0.01	0.01	...
movie	0.10	0.40	0.01	...
amazing	0.01	0.22	0.01	...
useless	0.01	0.20	0.03	...
was	0.01	0.00	0.60	...
⋮	⋮	⋮	⋮	...

The movie was really amazing

- How does the reviewer start his reviews (what is the first word that he chooses)?
- We could take the word which has the highest probability and put it as the first word in our review
- Having selected this what is the most likely second word that the reviewer uses?
- Having selected the first two words what is the most likely third word that the reviewer uses?
- and so on...

- But there is a catch here!

w	$P(X_1 = w)$	$P(X_2 = w \mid, X_1 = the)$	$P(X_i = w \mid, X_{i-2} = the, X_{i-1} = movie)$	...
the	0.62	0.01	0.01	...
movie	0.10	0.40	0.01	...
amazing	0.01	0.22	0.01	...
useless	0.01	0.20	0.03	...
was	0.01	0.00	0.60	...
⋮	⋮	⋮	⋮	...

The movie was really amazing

w	$P(X_1 = w)$	$P(X_2 = w \vert, X_1 = the)$	$P(X_i = w \vert, X_{i-2} = the, X_{i-1} = movie)$	...
the	0.62	0.01	0.01	...
movie	0.10	0.40	0.01	...
amazing	0.01	0.22	0.01	...
useless	0.01	0.20	0.03	...
was	0.01	0.00	0.60	...
⋮	⋮	⋮	⋮	...

The movie was really amazing

- But there is a catch here!
- Selecting the most likely word at each time step will only give us the same review again and again!

w	$P(X_1 = w)$	$P(X_2 = w \mid, X_1 = the)$	$P(X_i = w \mid, X_{i-2} = the, X_{i-1} = movie)$	...
the	0.62	0.01	0.01	...
movie	0.10	0.40	0.01	...
amazing	0.01	0.22	0.01	...
useless	0.01	0.20	0.03	...
was	0.01	0.00	0.60	...
⋮	⋮	⋮	⋮	...

The movie was really amazing

- But there is a catch here!
- Selecting the most likely word at each time step will only give us the same review again and again!
- But we would like to generate different reviews

w	$P(X_1 = w)$	$P(X_2 = w \mid, X_1 = the)$	$P(X_i = w \mid, X_{i-2} = the, X_{i-1} = movie)$	...
the	0.62	0.01	0.01	...
movie	0.10	0.40	0.01	...
amazing	0.01	0.22	0.01	...
useless	0.01	0.20	0.03	...
was	0.01	0.00	0.60	...
⋮	⋮	⋮	⋮	...

The movie was really amazing

- But there is a catch here!
- Selecting the most likely word at each time step will only give us the same review again and again!
- But we would like to generate different reviews
- So instead of taking the max value we can sample from this distribution

w	$P(X_1 = w)$	$P(X_2 = w \mid, X_1 = the)$	$P(X_i = w \mid, X_{i-2} = the, X_{i-1} = movie)$	...
the	0.62	0.01	0.01	...
movie	0.10	0.40	0.01	...
amazing	0.01	0.22	0.01	...
useless	0.01	0.20	0.03	...
was	0.01	0.00	0.60	...
⋮	⋮	⋮	⋮	...

The movie was really amazing

- But there is a catch here!
- Selecting the most likely word at each time step will only give us the same review again and again!
- But we would like to generate different reviews
- So instead of taking the max value we can sample from this distribution
- How?

w	$P(X_1 = w)$	$P(X_2 = w \mid, X_1 = the)$	$P(X_i = w \mid, X_{i-2} = the, X_{i-1} = movie)$	...
the	0.62	0.01	0.01	...
movie	0.10	0.40	0.01	...
amazing	0.01	0.22	0.01	...
useless	0.01	0.20	0.03	...
was	0.01	0.00	0.60	...
⋮	⋮	⋮	⋮	...

The movie was really amazing

- But there is a catch here!
- Selecting the most likely word at each time step will only give us the same review again and again!
- But we would like to generate different reviews
- So instead of taking the max value we can sample from this distribution
- How? Let us see!

w				
the				
movie				
amazing				
useless				
was				
is				
masterpiece				
I				
liked				
decent				

- Suppose there are 10 words in the vocabulary

w	$P(X_1 = w)$			
the	0.62			
movie	0.10			
amazing	0.01			
useless	0.01			
was	0.01			
is	0.01			
masterpiece	0.01			
I	0.21			
liked	0.01			
decent	0.01			

- Suppose there are 10 words in the vocabulary
- We have computed the probability distribution $P(X_1 = word)$

w	$P(X_1 = w)$			
the	0.62			
movie	0.10			
amazing	0.01			
useless	0.01			
was	0.01			
is	0.01			
masterpiece	0.01			
I	0.21			
liked	0.01			
decent	0.01			

- Suppose there are 10 words in the vocabulary
- We have computed the probability distribution $P(X_1 = word)$
- $P(X_1 = the)$ is the fraction of reviews having *the* as the first word

w	$P(X_1 = w)$	$P(X_2 = w \mid, X_1 = the)$	$P(X_i = w \mid, X_{i-2} = the, X_{i-1} = movie)$	...
the	0.62	0.01	0.01	...
movie	0.10	0.40	0.01	...
amazing	0.01	0.22	0.01	...
useless	0.01	0.20	0.03	...
was	0.01	0.00	0.60	...
is	0.01	0.00	0.30	...
masterpiece	0.01	0.11	0.01	...
I	0.21	0.00	0.01	...
liked	0.01	0.01	0.01	...
decent	0.01	0.02	0.01	...

- Suppose there are 10 words in the vocabulary
- We have computed the probability distribution $P(X_1 = word)$
- $P(X_1 = the)$ is the fraction of reviews having *the* as the first word
- Similarly, we have computed $P(X_2 = word_2 | X_1 = word_1)$ and $P(X_3 = word_3 | X_1 = word_1, X_2 = word_2)$

The movie . . .

Word
the
movie
amazing
useless
was
is
masterpiece
I
liked
decent

- Now consider that we want to generate the 3rd word in the review given the first 2 words of the review

The movie . . .

Index	Word
0	the
1	movie
2	amazing
3	useless
4	was
5	is
6	masterpiece
7	I
8	liked
9	decent

- Now consider that we want to generate the 3rd word in the review given the first 2 words of the review
- We can think of the 10 words as forming a 10 sided dice where each side corresponds to a word

The movie ...

Index	**Word**	$P(X_i = w \mid, X_{i-2} = the, X_{i-1} = movie)$	...
0	the	0.01	...
1	movie	0.01	...
2	amazing	0.01	...
3	useless	0.03	...
4	was	0.60	...
5	is	0.30	...
6	masterpiece	0.01	...
7	I	0.01	...
8	liked	0.01	...
9	decent	0.01	...

- Now consider that we want to generate the 3rd word in the review given the first 2 words of the review
- We can think of the 10 words as forming a 10 sided dice where each side corresponds to a word
- The probability of each side showing up is not uniform but as per the values given in the table

The movie . . .

Index	**Word**	$P(X_i = w\|, X_{i-2} = the, X_{i-1} = movie)$	. . .
0	the	0.01	. . .
1	movie	0.01	. . .
2	amazing	0.01	. . .
3	useless	0.03	. . .
4	was	0.60	. . .
5	is	0.30	. . .
6	masterpiece	0.01	. . .
7	I	0.01	. . .
8	liked	0.01	. . .
9	decent	0.01	. . .

- Now consider that we want to generate the 3rd word in the review given the first 2 words of the review
- We can think of the 10 words as forming a 10 sided dice where each side corresponds to a word
- The probability of each side showing up is not uniform but as per the values given in the table
- We can select the next word by rolling this dice and picking up the word which shows up

The movie ...

Index	Word	$P(X_i = w\|, X_{i-2} = the, X_{i-1} = movie)$	...
0	the	0.01	...
1	movie	0.01	...
2	amazing	0.01	...
3	useless	0.03	...
4	was	0.60	...
5	is	0.30	...
6	masterpiece	0.01	...
7	I	0.01	...
8	liked	0.01	...
9	decent	0.01	...

- Now consider that we want to generate the 3rd word in the review given the first 2 words of the review
- We can think of the 10 words as forming a 10 sided dice where each side corresponds to a word
- The probability of each side showing up is not uniform but as per the values given in the table
- We can select the next word by rolling this dice and picking up the word which shows up
- You can write a python program to roll such a biased dice

```
import numpy
review = [None,None,'the','movie']
words = ["the","movie","amazing","useless","was",
         "is","masterpiece","I","liked","decent"]
probs = dict()
probs[('the','movie')] = ["0.01","0.01","0.01",
    "0.03","0.60","0.30","0.01","0.01","0.01","0.01"]
# Add conditional probabilities for all pairs
outcome = numpy.random.choice(numpy.arange(0,10),
                p=probs[(review[-2],review[-1])])
print words[outcome],
```

```python
import numpy
review = [None,None]
words = ["the","movie","amazing","useless","was",
         "is","masterpiece","I","liked","decent"]
probs = dict()
probs[('the','movie')] = ["0.01","0.01","0.01",
    "0.03","0.60","0.30","0.01","0.01","0.01","0.01"]
# Add conditional probabilities for all pairs
for _ in range(5):
    outcome = numpy.random.choice(numpy.arange(0,10),
                        p=probs[(review[-2],review[-1])])
    review.append(words[outcome])
print ' '.join(review[2:])
```

- Now, at each timestep we do not pick the most likely word but all words are possible depending on their probability (just as rolling a biased dice or tossing a biased coin)

```python
import numpy
review = [None,None]
words = ["the","movie","amazing","useless","was",
         "is","masterpiece","I","liked","decent"]
probs = dict()
probs[('the','movie')] = ["0.01","0.01","0.01",
    "0.03","0.60","0.30","0.01","0.01","0.01","0.01"]
# Add conditional probabilities for all pairs
for _ in range(5):
    outcome = numpy.random.choice(numpy.arange(0,10),
                        p=probs[(review[-2],review[-1])])
    review.append(words[outcome])
print ' '.join(review[2:])
```

- Now, at each timestep we do not pick the most likely word but all words are possible depending on their probability (just as rolling a biased dice or tossing a biased coin)
- Every run will now give us a different review!

```python
import numpy
review = [None,None]
words = ["the","movie","amazing","useless","was",
            "is","masterpiece","I","liked","decent"]
probs = dict()
probs[('the','movie')] = ["0.01","0.01","0.01",
    "0.03","0.60","0.30","0.01","0.01","0.01","0.01"]
# Add conditional probabilities for all pairs
for _ in range(5):
    outcome = numpy.random.choice(numpy.arange(0,10),
                    p=probs[(review[-2],review[-1])])
    review.append(words[outcome])
print ' '.join(review[2:])
```

Generated Reviews

- the movie is liked decent

- Now, at each timestep we do not pick the most likely word but all words are possible depending on their probability (just as rolling a biased dice or tossing a biased coin)
- Every run will now give us a different review!

```python
import numpy
review = [None,None]
words = ["the","movie","amazing","useless","was",
            "is","masterpiece","I","liked","decent"]
probs = dict()
probs[('the','movie')] = ["0.01","0.01","0.01",
    "0.03","0.60","0.30","0.01","0.01","0.01","0.01"]
# Add conditional probabilities for all pairs
for _ in range(5):
    outcome = numpy.random.choice(numpy.arange(0,10),
                    p=probs[(review[-2],review[-1])])
    review.append(words[outcome])
print ' '.join(review[2:])
```

Generated Reviews

- the movie is liked decent
- I liked the amazing movie

- Now, at each timestep we do not pick the most likely word but all words are possible depending on their probability (just as rolling a biased dice or tossing a biased coin)
- Every run will now give us a different review!

```python
import numpy
review = [None,None]
words = ["the","movie","amazing","useless","was",
          "is","masterpiece","I","liked","decent"]
probs = dict()
probs[('the','movie')] = ["0.01","0.01","0.01",
    "0.03","0.60","0.30","0.01","0.01","0.01","0.01"]
# Add conditional probabilities for all pairs
for _ in range(5):
    outcome = numpy.random.choice(numpy.arange(0,10),
                        p=probs[(review[-2],review[-1])])
    review.append(words[outcome])
print ' '.join(review[2:])
```

Generated Reviews

- the movie is liked decent
- I liked the amazing movie
- the movie is masterpiece

- Now, at each timestep we do not pick the most likely word but all words are possible depending on their probability (just as rolling a biased dice or tossing a biased coin)
- Every run will now give us a different review!

```python
import numpy
review = [None,None]
words = ["the","movie","amazing","useless","was",
         "is","masterpiece","I","liked","decent"]
probs = dict()
probs[('the','movie')] = ["0.01","0.01","0.01",
    "0.03","0.60","0.30","0.01","0.01","0.01","0.01"]
# Add conditional probabilities for all pairs
for _ in range(5):
    outcome = numpy.random.choice(numpy.arange(0,10),
                                  p=probs[(review[-2],review[-1])])
    review.append(words[outcome])
print ' '.join(review[2:])
```

Generated Reviews

- the movie is liked decent
- I liked the amazing movie
- the movie is masterpiece
- the movie I liked useless

- Now, at each timestep we do not pick the most likely word but all words are possible depending on their probability (just as rolling a biased dice or tossing a biased coin)
- Every run will now give us a different review!

Returning back to our story....

M7: More realistic than real life

w	$P(X_i = w \mid, X_{i-2} = more, X_{i-1} = realistic)$	$P(X_i = w \mid, X_{i-2} = realistic, X_{i-1} = than)$	$P(X_i = w \mid X_{i-2} = than, X_{i-1} = real)$	...
than	0.61	0.01	0.20	...
as	0.12	0.10	0.16	...
for	0.14	0.09	0.05	...
real	0.01	0.50	0.01	...
the	0.02	0.12	0.12	...
life	0.05	0.11	0.33	...

$$P(M7) = P(X_1 = more).P(X_2 = realistic|X_1 = more).$$
$$P(X_3 = than|X_1 = more, X_2 = realistic).$$
$$P(X_4 = real|X_2 = realistic, X_3 = than).$$
$$P(X_5 = life|X_3 = than, X_4 = real)$$
$$= 0.2 \times 0.25 \times 0.61 \times 0.50 \times 0.33 = 0.005$$

- Okay, so now what can we do with this joint distribution?
- Given a review, *classify* if this was written by the reviewer
- *Generate* new reviews which would look like reviews written by this reviewer

M7: More realistic than real life

w	$P(X_i = w \mid, X_{i-2} = more, X_{i-1} = realistic)$	$P(X_i = w \mid, X_{i-2} = realistic, X_{i-1} = than)$	$P(X_i = w \mid X_{i-2} = than, X_{i-1} = real)$	...
than	0.61	0.01	0.20	...
as	0.12	0.10	0.16	...
for	0.14	0.09	0.05	...
real	0.01	0.50	0.01	...
the	0.02	0.12	0.12	...
life	0.05	0.11	0.33	...

$$P(M7) = P(X_1 = more).P(X_2 = realistic|X_1 = more).$$
$$P(X_3 = than|X_1 = more, X_2 = realistic).$$
$$P(X_4 = real|X_2 = realistic, X_3 = than).$$
$$P(X_5 = life|X_3 = than, X_4 = real)$$
$$= 0.2 \times 0.25 \times 0.61 \times 0.50 \times 0.33 = 0.005$$

- Okay, so now what can we do with this joint distribution?
- Given a review, *classify* if this was written by the reviewer
- *Generate* new reviews which would look like reviews written by this reviewer
- *Correct noisy reviews* or help in completing incomplete reviews

$$\underset{X_5}{argmax}\ P(X_1 = the, X_2 = movie, X_3 = was, X_4 = amazingly, X_5 = ?)$$

Let us take an example from another domain

- Consider images which contain $m \times n$ pixels (say 32×32)

- Consider images which contain $m \times n$ pixels (say 32×32)
- Each pixel here is a random variable which can take values from 0 to 255 (colors)

- Consider images which contain $m \times n$ pixels (say 32×32)
- Each pixel here is a random variable which can take values from 0 to 255 (colors)
- We thus have a total of $32 \times 32 = 1024$ random variables $(X_1, X_2, ..., X_{1024})$

- Consider images which contain $m \times n$ pixels (say 32×32)
- Each pixel here is a random variable which can take values from 0 to 255 (colors)
- We thus have a total of $32 \times 32 = 1024$ random variables $(X_1, X_2, ..., X_{1024})$
- Together these pixels define the image and different combinations of pixel values lead to different images

- Consider images which contain $m \times n$ pixels (say 32×32)
- Each pixel here is a random variable which can take values from 0 to 255 (colors)
- We thus have a total of $32 \times 32 = 1024$ random variables $(X_1, X_2, ..., X_{1024})$
- Together these pixels define the image and different combinations of pixel values lead to different images


- Consider images which contain $m \times n$ pixels (say 32×32)
- Each pixel here is a random variable which can take values from 0 to 255 (colors)
- We thus have a total of $32 \times 32 = 1024$ random variables $(X_1, X_2, ..., X_{1024})$
- Together these pixels define the image and different combinations of pixel values lead to different images

- Consider images which contain $m \times n$ pixels (say 32×32)
- Each pixel here is a random variable which can take values from 0 to 255 (colors)
- We thus have a total of $32 \times 32 = 1024$ random variables $(X_1, X_2, ..., X_{1024})$
- Together these pixels define the image and different combinations of pixel values lead to different images

- Consider images which contain $m \times n$ pixels (say 32×32)
- Each pixel here is a random variable which can take values from 0 to 255 (colors)
- We thus have a total of $32 \times 32 = 1024$ random variables $(X_1, X_2, ..., X_{1024})$
- Together these pixels define the image and different combinations of pixel values lead to different images

- Consider images which contain $m \times n$ pixels (say 32×32)
- Each pixel here is a random variable which can take values from 0 to 255 (colors)
- We thus have a total of $32 \times 32 = 1024$ random variables $(X_1, X_2, ..., X_{1024})$
- Together these pixels define the image and different combinations of pixel values lead to different images

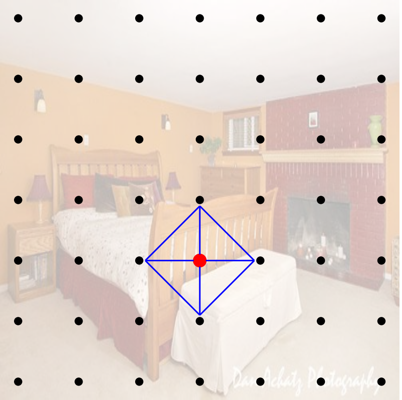
- Consider images which contain $m \times n$ pixels (say 32×32)
- Each pixel here is a random variable which can take values from 0 to 255 (colors)
- We thus have a total of $32 \times 32 = 1024$ random variables $(X_1, X_2, ..., X_{1024})$
- Together these pixels define the image and different combinations of pixel values lead to different images

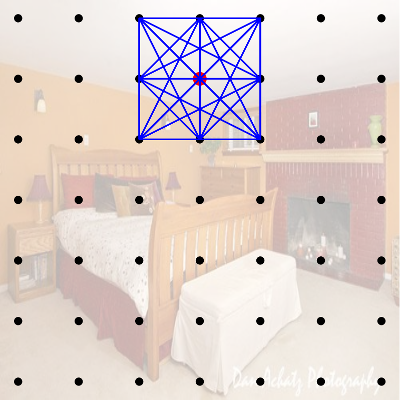
- Consider images which contain $m \times n$ pixels (say 32×32)
- Each pixel here is a random variable which can take values from 0 to 255 (colors)
- We thus have a total of $32 \times 32 = 1024$ random variables $(X_1, X_2, ..., X_{1024})$
- Together these pixels define the image and different combinations of pixel values lead to different images

- Consider images which contain $m \times n$ pixels (say 32×32)
- Each pixel here is a random variable which can take values from 0 to 255 (colors)
- We thus have a total of $32 \times 32 = 1024$ random variables $(X_1, X_2, ..., X_{1024})$
- Together these pixels define the image and different combinations of pixel values lead to different images

- Consider images which contain $m \times n$ pixels (say 32×32)
- Each pixel here is a random variable which can take values from 0 to 255 (colors)
- We thus have a total of $32 \times 32 = 1024$ random variables $(X_1, X_2, ..., X_{1024})$
- Together these pixels define the image and different combinations of pixel values lead to different images
- Given many such images we want to learn the joint distribution $P(X_1, X_2, ..., X_{1024})$

- We can assume each pixel is dependent only on its neighbors





- We can assume each pixel is dependent only on its neighbors
- In this case we could factorize the distribution over a Markov network

$$\prod \phi(D_i)$$

where D_i is a set of variables which form a maximal clique (basically, groups of neighboring pixels)

- Again, what can we do with this joint distribution?

Probability Score = 0.01

- Again, what can we do with this joint distribution?
- Given a new image, *classify* if is indeed a bedroom

- Again, what can we do with this joint distribution?
- Given a new image, *classify* if is indeed a bedroom
- *Generate new images* which would look like bedrooms (say, if you are an interior designer)

- Again, what can we do with this joint distribution?
- Given a new image, *classify* if is indeed a bedroom
- *Generate new images* which would look like bedrooms (say, if you are an interior designer)
- *Correct noisy images* or help in completing incomplete images

- Such models which try to estimate the probability $P(X)$ from a large number of samples are called generative models