1 Word-Level Neural Bi-gram Language Model

1.a

$$J = CE(y, \hat{y}) = -\sum_i y_i log(\hat{y}_i) = -log(\frac{e^{\theta_i}}{\sum_j e^{\theta_j}}) = log(\sum_j e^{\theta_j}) - \theta_i$$

$$\hat{y} = softmax(\theta)$$

$$\hat{y}_i = softmax(\theta_i) = \frac{e^{\theta_i}}{\sum_j e^{\theta_j}}$$

$\nabla_{\theta_i} J = \frac{\partial J}{\partial \theta_i} = \frac{e^{\theta_i}}{\sum_j e^{\theta_j}} - 1 = \hat{y}_i - 1$

$\nabla_{\theta_{j \neq i}} J = \frac{\partial J}{\partial \theta_{j \neq i}} = \frac{e^{\theta_j}}{\sum_j e^{\theta_j}} = \hat{y_j}$

$$\implies \nabla_\theta J = \frac{\partial J}{\partial \theta} = \hat{y} - y$$

1.b

$$J = CE(y, \hat{y}) = -\sum_i y_i log(\hat{y}_i)$$

$$h = \sigma(xW_1 + b_1)$$

$$\hat{y} = \text{softmax}(hW_2 + b_2)$$

denote

$$p = hW_2 + b_2$$

$$q = xW_1 + b_1$$

then

$$\frac{\partial J}{\partial x} = \frac{\partial J}{\partial p} \cdot \frac{\partial p}{\partial h} \cdot \frac{\partial h}{\partial q} \cdot \frac{\partial q}{\partial x}$$

1. $\frac{\partial J}{\partial p} = \hat{y} - y$
2. $\frac{\partial p}{\partial h} = \frac{\partial (hW_2 + b_2)}{\partial h} = W_2$
3. $\frac{\partial h}{\partial q} = \frac{\partial \sigma(q)}{\partial q} = \frac{\partial \frac{1}{1+exp(-q)}}{\partial q} = \frac{exp(-q)}{(1+exp(-q))^2} = \frac{1}{1+exp(-q)} \cdot \frac{exp(-q)}{1+exp(-q)} = h \cdot (1-h)$
4. $\frac{\partial q}{\partial x} = \frac{\partial (xW_1 + b_1)}{\partial x} = W_1$

$$\implies \frac{\partial J}{\partial x} = (\hat{y} - y) \cdot W_2 \cdot h \cdot (1-h) \cdot W_1$$

1.c

(code)

1.d

(code)

dev perplexity : 112.81714028757537

2 Theoretical Inquiry of a Simple RNN Language Model

2.a

$$e^{(t)} = x^{(t)}L$$

$$h^{(t)} = \sigma(h^{(t-1)}H + e^{(t)}I + b_1)$$

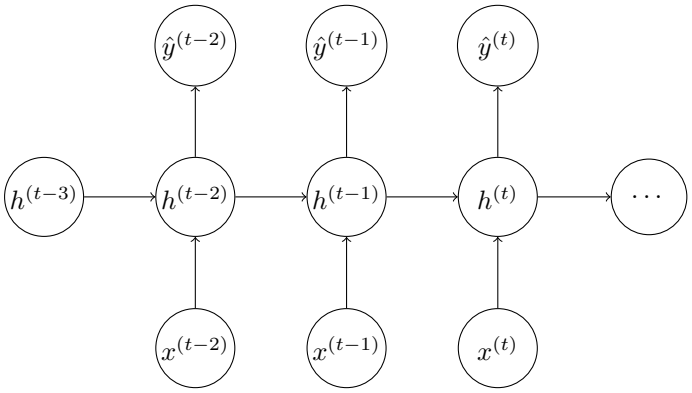
$$\hat{y}^{(t)} = \text{softmax}(h^{(t)}U + b_2)$$

$$J^{(t)}(\theta) = CE(y^{(t)}, \hat{y}^{(t)}) = -\sum_{i=1}^{|V|}(y_i^{(t)} \cdot log(\hat{y_i}^{(t)}))$$

- $\frac{\partial J^{(t)}}{\partial U} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial U} = (\hat{y}^{(t)} - y^{(t)}) \cdot h^{(t)}$
- $\frac{\partial J^{(t)}}{\partial L_{x^{(t)}}} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial e^{(t)}} \cdot \frac{\partial e^{(t)}}{\partial L_{x^{(t)}}} = (\hat{y}^{(t)} - y^{(t)}) \cdot U \cdot h^{(t)} \cdot (1 - h^{(t)})$
- $\frac{\partial J^{(t)}}{\partial I}|_{(t)} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial e^{(t)}} \cdot \frac{\partial e^{(t)}}{\partial I}|_{(t)} = (\hat{y}^{(t)} - y^{(t)}) \cdot U \cdot h^{(t)} \cdot (1 - h^{(t)}) \cdot x^{(t)}L$
- $\frac{\partial J^{(t)}}{\partial H}|_{(t)} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial e^{(t)}} \cdot \frac{\partial e^{(t)}}{\partial H}|_{(t)} = (\hat{y}^{(t)} - y^{(t)}) \cdot U \cdot h^{(t)} \cdot (1 - h^{(t)}) \cdot h^{(t-1)}$
- $\delta^{(t-1)} = \frac{\partial J^{(t)}}{\partial h^{(t-1)}} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial h^{(t-1)}} = (\hat{y}^{(t)} - y^{(t)}) \cdot U \cdot h^{(t)} \cdot (1 - h^{(t)}) \cdot H$

2.b

2.b.i Unrolled network for 3 time-steps



2.b.ii "back-propagation-through-time" gradients

- $\frac{\partial J^{(t)}}{\partial L_{x^{(t-1)}}} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial e^{(t)}} \cdot \frac{\partial e^{(t)}}{\partial L_{x^{(t-1)}}} = (\hat{y}^{(t)} - y^{(t)}) \cdot U \cdot h^{(t)} \cdot (1 - h^{(t)}) \cdot W \cdot h^{(t-1)}$
- $\frac{\partial J^{(t)}}{\partial H}|_{(t-1)} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial e^{(t)}} \cdot \frac{\partial e^{(t)}}{\partial H}|_{(t-1)} = (\hat{y}^{(t)} - y^{(t)}) \cdot U \cdot h^{(t)} \cdot (1 - h^{(t)}) \cdot W \cdot h^{(t-2)}$
- $\frac{\partial J^{(t)}}{\partial I}|_{(t-1)} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial e^{(t)}} \cdot \frac{\partial e^{(t)}}{\partial I}|_{(t-1)} = (\hat{y}^{(t)} - y^{(t)}) \cdot U \cdot h^{(t)} \cdot (1 - h^{(t)}) \cdot W \cdot h^{(t-3)}$
- $\frac{\partial J^{(t)}}{\partial b_1}|_{(t-1)} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial e^{(t)}} \cdot \frac{\partial e^{(t)}}{\partial b_1}|_{(t-1)} = (\hat{y}^{(t)} - y^{(t)}) \cdot U \cdot h^{(t)} \cdot (1 - h^{(t)})$

3 Generating Shakespeare Using a Character-level Language Model

3.a

Character-based language model advantage:

- It can generate or recognize words that are not present in the training vocabulary by composing them from known characters. This makes them more robust when dealing with rare or specialized terms that may not be present in the training data of a word-based model.
- Better identifing and correcting misspaelled and typos.

Word-based language model advantage:

- Not 'making up' words.
- Better context and semantic Understanding: Word-based models capture the meaning of a text at a higher level of abstraction. MOre coherent and contextually appropriate responses.
- Computational Efficiency: Word-based models are computationally more efficient than character-based models, as they deal with larger units of text.

3.b Training-loss plot

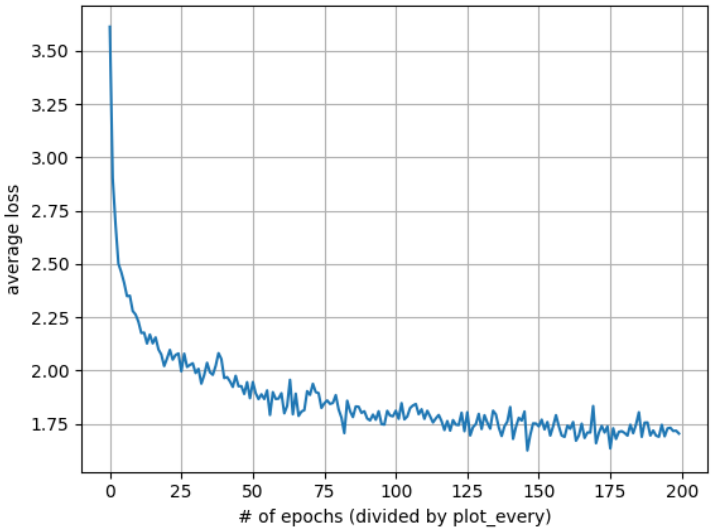


Figure 1: Training-loss

4 Perplexity

4.a

we will show:

$$2^{\frac{1}{M}\sum_{i=1}^{M} log_2 P(s_i|s_1...s_{i-1})} = e^{\frac{1}{M}\sum_{i=1}^{M} lnP(s_i|s_1...s_{i-1})}$$

proof:

$$2^{\frac{1}{M}\sum_{i=1}^{M} log_2 P(s_i|s_1...s_{i-1})} = 2^{\frac{1}{M}log_2 P(s_1|s_1)} 2^{\frac{1}{M}log_2 P(s_2|s_1 s_2)} ... 2^{\frac{1}{M}log_2 P(s_M|s_1...s_M)} =$$

$$= P(s_1|s_1)^{\frac{1}{M}} P(s_2|s_1 s_2)^{\frac{1}{M}} ... P(s_M|s_1...s_M)^{\frac{1}{M}} =$$

$$= e^{\frac{1}{M}lnP(s_1|s_1)} e^{\frac{1}{M}lnP(s_2|s_1 s_2)} ... e^{\frac{1}{M}lnP(s_M|s_1...s_M)} = e^{\frac{1}{M}\sum_{i=1}^{M} lnP(s_i|s_1...s_{i-1})}$$

4.b

bi-gram model

- **Shakespeare perplexity:** 7.504480691937161
- **Wikipedia perplexity:** 30.17933100905446

Char-based bi-gram model

- **Shakespeare perplexity:** 6.769117782221575
- **Wikipedia perplexity:** 16.2687740098086

4.c

The first model is a word-based bi-gram model, and the second model is a char-based bi-gram model. Though we were expecting the char-based model to be better on Shakespeare passege and the word-based model to be better on Wikipedia passege (in terms of perplexity), we were surprised to see that both models did better on Shakespeare passege. The char-based model is more robust to rare words and was trained on Shakespeare corpus, so the results are not surprising. On the other hand, the word-based model wasn't trained on Shakespeare corpus, so we were expecting it to do better on Wikipedia passege. We can think about few reasons for that:

- Consistency of Language: "Shakespeare's language" is more consistent and constrained compared to the diversed range of topics and writing styles found in a single passage from Wikipedia. So based on the context and words in a peragraph, both models might predict better on Shakespeare's "strang" words.
- The Wikipedia

5 Deep Averaging Networks

5.a

(5.a)

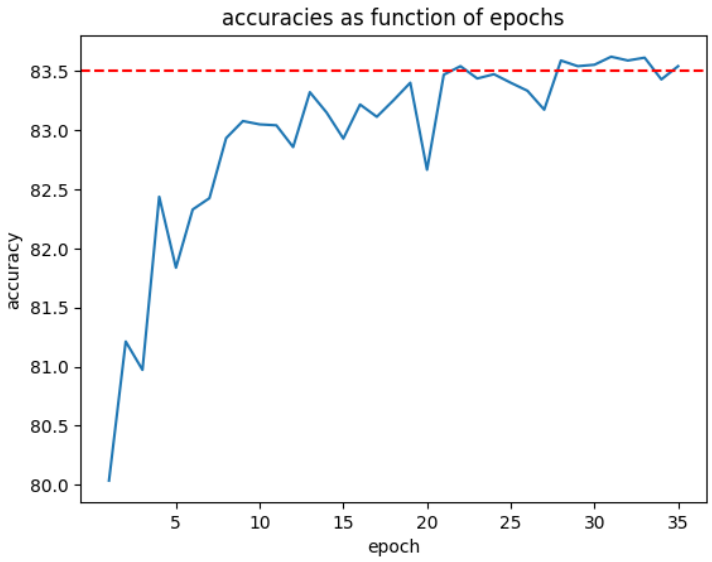


Figure 2: Accuracy as a function of number of epochs

5.b

(5.b)

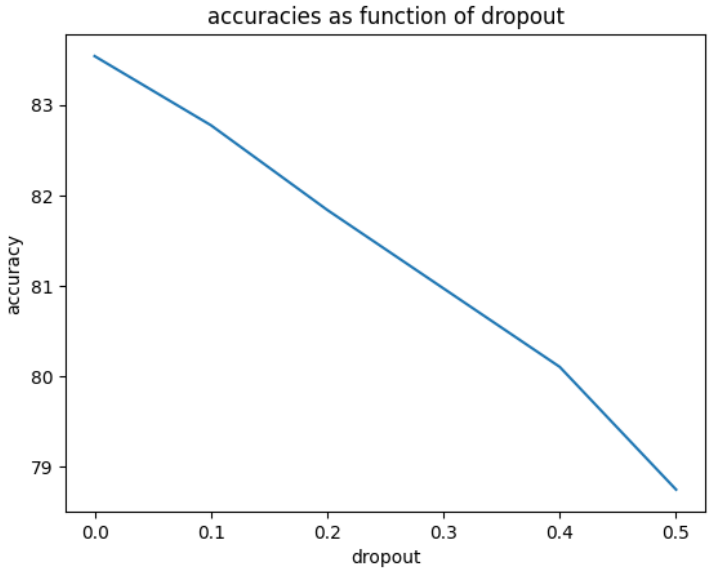


Figure 3: Accuracy as a function of dropout rate

5.c

(5.c)

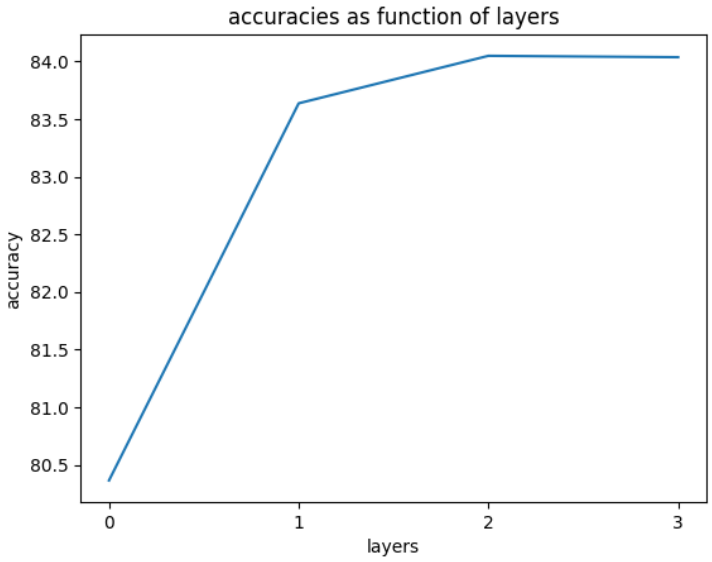


Figure 4: Accuracy as a function of number of hidden layers

5.d

(5.d)

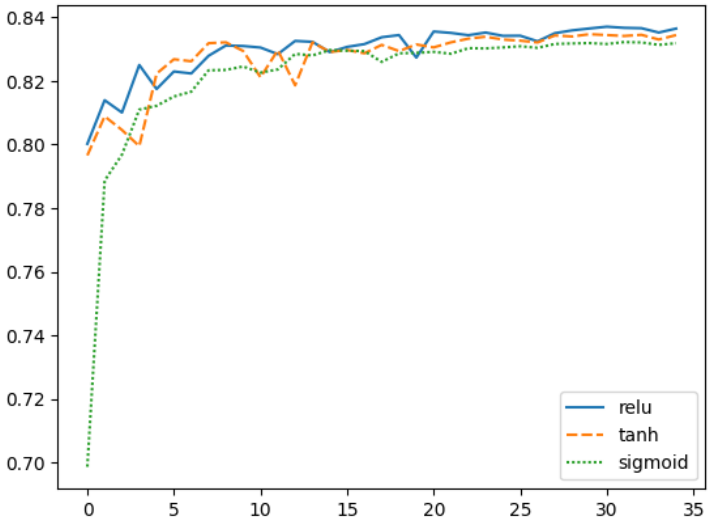


Figure 5: Accuracy across epoch

5.e

1. "This is the latest entry in the long series of films with the French agent, O.S.S. 117 (the French answer to James Bond). The series was launched in the early 1950's, and spawned at least eight

films (none of which was ever released in the U.S.). 'O.S.S.117:Cairo,Nest Of Spies' is a breezy little comedy that should not...repeat NOT, be taken too seriously. Our protagonist finds himself in the middle of a spy chase in Egypt (with Morroco doing stand in for Egypt) to find out about a long lost friend. What follows is the standard James Bond/Inspector Cloussou kind of antics. Although our man is something of an overt xenophobe,sexist,homophobe, it's treated as pure farce (as I said, don't take it too seriously). Although there is a bit of rough language & cartoon violence, it's basically okay for older kids (ages 12 & up). As previously stated in the subject line, just sit back,pass the popcorn & just enjoy."

Predicted: 0 (negative)
Actual: 1 (positive)
Explanation: the words "should not...repeat NOT", "pure farce", "xenophobe,sexist,homophobe", "rough language & cartoon violence" and "don't take it too seriously" are negative words.

2. "This is a really sad, and touching movie! It deals with the subject of child abuse. Itś really sad, but mostly a true story, because it happens everyday. Elijah Wood and Joseph Mazzello play the two children or Lorraine Bracco, a single mother who just tries to make a home for them. While living with her parents, a man, who likes to be called "The King" comes into their life. He hits the youngest boy, Bobby, but the two brothers vow not to tell their mother. But finally she finds out, after the Bobby is hurt badly. The end kind of ruined it for me, because it is so totally unbelievable. But, except for that, I love the movie."

   Predicted: 0 (negative)
   Actual: 1 (positive)
   Explanation: the words "really sad", "hurt badly", "ruined" and "unbelievable" are negative words.

3. "These days, writers, directors and producers are relying more and more on the "surprise" ending. The old art of bringing a movie to closure, taking all of the information we have learned through out the movie and bringing it to a nice complete ending, has been lost. Now what we have is a movie that, no matter how complex, detailed, or frivolous, can be wrapped up in 5 minutes. It was all in his/her head. That explanation is the directorś safety net. If all else fails, or if the writing wasnt́ that good, or if we ran out of money to complete the movie, we can always say "it was all in his/her head" and end the movie that way. The audience will buy it because, well, none of us are psychologists, and none of us are suffering from schizophrenia (not that we know about) so we take the story and believe it. After all, the mind is a powerful thing. Some movies have pulled it off. But those movies are the reason why we are getting more and more of these crap endings. Every director/writer now thinks they can pull it off because, well, Fight Club did it and it made a lot of money. So we get movies like The Machinist, Secret Window, Identity, and this movie (just to name a few)."

   Predicted: 1 (positive)
   Actual: 0 (negative)
   Explanation: the words ""surprise" ending", "nice complete ending", "The audience will buy it", "pull it off" are positive words.

4. "It could be easy to complain about the quality of this movie (you don't have to throw cartloads of money at a movie to make it good, nor will it guarantee that it is worth watching) but I think

that is totally missing the point. If your expecting fast cars, T&A or a movie that will spell itself out for you then don't watch this, you'll be disappointed and dumbfounded.¡br /¿¡br /¿This movie was thoroughly enjoyable, kept us on the edge of our seats and made us really think. The writer obviously put a lot of thought and research behind this movie and it shows through the end, just remember to keep an open mind. ¡br /¿¡br /¿Note: the school scenes were all filmed at McMaster University and most of the rest was done in Toronto."

Predicted: 0 (negative)
Actual: 1 (positive)
Explanation: the words "totally missing the point", "disappointed and dumbfounded" are negative words. The begining of the review use negative words, but the end of the review use positive words.

5. "Intended as light entertainment, this film is indeed successful as such during its first half, but then succumbs to a rapidly foundering script that drops it down. Harry (Judd Nelson), a "reformed" burglar, and Daphne (Gina Gershon), an aspiring actress, are employed as live window mannequins at a department store where one evening they are late in leaving and are locked within, whereupon they witness, from their less than protective glass observation point, an apparent homicide occurring on the street. The ostensible murderer, Miles Raymond (Nick Mancuso), a local sculptor, returns the following day to observe the mannequins since he realizes that they are the only possible witnesses to the prior nightś violent event and, when one of the posing pair "flinches", the fun begins. Daphne and Harry report their observations at a local police station, but when the detective taking a crime report remembers Harryś criminal background, he becomes cynical. There are a great many ways in which a film can become hackneyed, and this one manages to utilize most of them, including an obligatory slow motion bedroom scene of passion. A low budget affair shot in Vancouver, even police procedural aspects are displayed by rote. The always capable Gershon tries to make something of her role, but Mancuso is incredibly histrionic, bizarrely so, as he attacks his lines with an obvious loose rein. Although the film sags into nonsense, cinematographer Glen MacPherson prefers to not follow suit, as he sets up with camera and lighting some splendidly realised compositions that a viewer may focus upon while ignoring plot holes and witless dialogue. A well-crafted score, appropriately based upon the action, is contributed by Hal Beckett. The mentioned dialogue is initially somewhat fresh and delivered well in a bantering manner by Nelson and Gershon, but in a subsequent context of flawed continuity and logic, predictability takes over. The direction reflects a lack of original ideas or point of view, and post-production flaws set the work back farther than should be expected for a basic thriller."

Predicted: 1 (positive)
Actual: 0 (negative)
Explanation: the words "entertainment", "successful", "fun begins", "splendidly realised", "well-crafted", "delivered well", "fresh", "A well-crafted score" and " are positive words. Long and detailed review with many actor's names.

6 Right-to-left vs left-to-right Estimation

From the chain rule, we know that $P(x_n)P(x_{n-1}|x_n)...P(x_0|x_1) = P(x_n, x_{n-1}, ..., x_0)$.

From Bayes theorem, we know that $P(B)P(A|B) = P(B|A)P(A)$.

we will prove by induction:

- Base case: $P(x_0)P(x_1|x_0) = P(x_0, x_1) = P(x_1)P(x_0|x_1)$
- Inductive step: Assume $P(x_n)P(x_{n-1}|x_n)...P(x_0|x_1) = P(x_0)P(x_1|x_0)...P(x_n|x_{n-1})$.
  we want to show :

$$P(x_{n+1})P(x_n|x_{n+1})P(x_{n-1}|x_n)...P(x_0|x_1) = P(x_{n+1}|x_n)P(x_n|x_{n-1})...P(x_1|x_0)P(x_0)$$

  $P(x_{n+1})P(x_n|x_{n+1})P(x_{n-1}|x_n)...P(x_0|x_1) =$
  (according to Bayes: $P(x_{n+1})P(x_n|x_{n+1}) = P(x_n)P(x_{n+1}|x_n)$)
  $P(x_n)P(x_{n+1}|x_n)P(x_{n-1}|x_n)...P(x_0|x_1) = P(x_{n+1}|x_n)P(x_n)P(x_{n-1}|x_n)...P(x_0|x_1) =$
  (from the inductive hypothesis)
  $= P(x_{n+1}|x_n)P(x_0)P(x_1|x_0)...P(x_n|x_{n-1}) = P(x_{n+1}|x_n)P(x_n|x_{n-1})...P(x_1|x_0)P(x_0)$