Assignment 02

Temporal Subsampling of a Discrete Time Markov Process

(a)

$$
\begin{aligned}
p(x_n, x_{n-1}, \dots, x_2) &= \sum_{x_1} p(x_n, x_{n-1}, \dots, x_1) \\
&= \sum_{x_1} p(x_1)p(x_2|x_1)\dots p(x_n|x_{n-1}) \\
&= \prod_{i=3}^{n} p(x_i|x_{i-1}) \times \sum_{x_1} p(x_2|x_1)p(x_1) \\
&= \prod_{i=3}^{n} p(x_i|x_{i-1}) \times p(x_2)
\end{aligned}
$$

Repeat the above process $n-k-1$ times and we have

$$
p(x_n, x_{n-1}, \dots, x_{n-k}) = p(x_{n-k}) \prod_{i=n-k+1}^{n} p(x_i|x_{i-1})
$$

(b)

$$
\begin{aligned}
p(x_1, x_3, x_4) &= \sum_{x_2} p(x_1, x_2, x_3, x_4) \\
&= \sum_{x_2} p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_3) \\
&= p(x_1)p(x_4|x_3) \sum_{x_2} p(x_3|x_2)p(x_2|x_1) \\
&= p(x_1)p(x_3|x_1)p(x_4|x_3)
\end{aligned}
$$

(c)

From

$$
\begin{cases}
p(x_1, x_3, x_4) = p(x_1, x_3)p(x_4|x_1, x_3) \\
p(x_1, x_3, x_4) = p(x_1)p(x_3|x_1)p(x_4|x_3) \\
p(x_1, x_3) = p(x_1)p(x_3|x_1)
\end{cases}
$$

we can conclude that

$$
p(x_4|x_3, x_1) = p(x_4|x_3)
$$

(d)

$$
\begin{aligned}
p(x_1, x_2, x_4) &= \sum_{x_3} p(x_1, x_2, x_3, x_4) \\
&= \sum_{x_3} p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_3) \\
&= p(x_1)p(x_2|x_1) \sum_{x_3} p(x_4|x_3)p(x_3|x_2) \\
&= p(x_1)p(x_2|x_1)p(x_4|x_2)
\end{aligned}
$$

(e)

From

$$\begin{cases} p(x_1, x_2, x_4) = p(x_1)p(x_2|x_1)p(x_4|x_2) \\ p(x_1, x_2, x_4) = p(x_1, x_2)p(x_4|x_1, x_2) \\ p(x_1, x_2) = p(x_1)p(x_2|x_1) \end{cases}$$

we can conclude that

$$p(x_4|x_2, x_1) = p(x_4|x_2)$$

(f)

For any subset of random variables from a Markov process, the joint probability can be factored as a product of conditional probabilities, where each conditional probability depends only on the immediate predecessor in the selected subset.

Markov Models for Text: Seuss and Saki

(a)

```
lmhg ixub xrgq whlf ovxd jlgs qaol qqbb lccl ggjk
mrdz qwta uovx liei rjdu rwqu gnuk emzg axyl spgg
tgja senk nkke qxas lgtx uxxi jots tbjb iinn gesh
ielg htzx umhx nlfc ippm cvba jirm juer xslb yfsz
zmtb tcze vhvb ibbv erqb fbyr qrpk pzxp inry tjeq
pguh keov atkm dqyz hbvl twui ognq gcbc dtsa tmkk
kzay jsjb ipwr bjqc xjey bfoc ryzd mhrj pxzo xdej
spgi unwx kaah mety lhym gypn jwty mflo pteb jadr
jnvz owmr ekce tggc twfy smgh mccf jmus laru uqel
paut mdeu lawo fspo oqdu ghao zedh infu bemt oyap
```

(b)

```
sidc iokn nrte tman lypl ylnr nnmd bare uyns ulko
teua uwlr otlr udot tnon moim sspo nyye olku ppdi
isww uuwy mdoe fnit uioi leoh anir sple eaep okiw
dpho dlen reyt nhri okeo yrtu nmya lloe wmur cohk
scna tmny acwu suer flyt tnye esds oeei rtnt iitr
prmo nmrl lili tnoo maiv eymr aneo myyl dceo pkyr
eade ieas nowe rnoh leai lyoo leho eoeh ocio laoh
oooc ilsa oeta wnnt loey ekbs cera etai itnl mdrt
hiee mhad iwum etrl aniw oert womi aoit nnel etii
aytt moie odee fiti iita btoy tmsw roki edit hddy
```

(c)

```
yrer omem udou coma noun doth tere ouyo dest noul
nomy heem llik rewi youl heer othe atou thev ther
keth meee aith like sche ithe reen wifr itom erke
thea woul erew rote ther thit nota ouyw noul iker
idld evea ceve vill linu sthe inde ithe enth dere
leth them idor oule otsp roto rere ouro otam yoth
omer eath ncot myoo erke rote erea eend eeno eren
thea ilyo inor ikee math itre thet ithe keme ular
nyot thin mami oure ulik mare dero here lith oull
aman oull youn rere urot youd rere arke ilit hend
```

(d)

```
ourr notr uldq ldum ther eere otch otzw like eall
ayhs like spam midd ldse ther deth ndzg ould uldv
lddl itsd ould lewd ould spam read onch otiu erea
even hath reat notg deat otan ould dolk ewdh toqq
then dopb thin amid myua here cene ikeg even ikek
otya beej enec like tany tsfr enee ould from youl
ourq rude sche like that uldo ikew ould ould itsu
nche will ywhe otlf ther ulde spam then notg your
ould here ikev thin hene topa omet eepn ould amid
here otxd eate urlc like uldp love meth amid thin
```

(e)

```
Letter Probabilities from Saki Story (10x10 Grid)
==================================================
evea ndte ntcs thoo odxa mesm olea uase snwe ueen
wara ueme isee taol seei ttia htgs arur enoe itna
npne bdht wooa uobe mfca nalo wasa tyoe rvep tiri
iidt hses uhea hnet drni sthi ehsh bire pnaw itea
eifm rtct bhen tsoh ltee eseo east wsno aett trra
teem onos nqte moig ntto rsih ttga kryb wotc eooy
tlcs veot dhet seol htat ooeh athn besr nror lmsk
eihw aelb tste rtir eaea troi trds srst tnll htrn
tcsd sioo senw ceea idau eayu ikrl aoos ildz ltet
ewat erea dlot hent icss ifrn hcmd reeo yird eero
```

```
1st Order Markov from Saki Story (10x10 Grid)
==================================================
yilf hasp okic lers menr lyon ingt athe hoar ousu
gipr icon thou fayi iptz orsa uggu asul nthe ngho
ndut oren anth engh neel wldl rele rago oldg rdid
carv dsul huru igit unor ongo isto sheg trsp frsh
orel alyo vese sthe isid dshe enda syen swhi hewa
anem arye anil edos otin hera ther herr cyed nuto
heni heat beng oudy edes nesu fand onth ryea medr
onde ange ysom ndss eser rnth indt best uthe wngr
urie dira erul ther dico thor eldi ousi ssth ldsu
itho ndif ovou rota ener ones wher nera esul iser
```

```
2nd Order Markov from Saki Story (10x10 Grid)
==================================================
iess hade dece ther asto butb fros ther alse acer
erse ther call houl edge andf hend earo gres sher
cern isee kenc ways hent asai rone vold bout erld
sher ndow then atin tare ther ribl osto plam rone
unds sher rrie ster wing gern vest ofal eree llac
athe wast liti opic esse ingr yout ways nged erea
wily yser onig elin mose died outi sped fron akee
seed orno thav ttiv hers mbus ngen ttle erno atzx
ther horu whis adde yeak sell heds mene cerc iste
seds acho dief nusi ince lver ding esso thar deas
```

(f)

1. Equiprobable Letters (Part a):

- Generated words are completely random
- No linguistic patterns or structure
- Expected to have very few valid English words

2. Letter Probabilities (Part b):
   - Words reflect the frequency distribution of letters in the source text
   - More realistic letter combinations than random
   - Still lacks sequential dependencies between letters
3. First Order Markov Chain (Part c):
   - Considers the probability of the next letter given the current letter
   - Captures basic sequential patterns in the language
   - Should produce more realistic letter sequences
4. Second Order Markov Chain (Part d):
   - Considers the probability of the next letter given the previous two letters
   - Captures more complex linguistic patterns
   - Should produce the most realistic letter sequences
5. Comparison between Spamiam and Saki Story:
   - Spamiam: Short, repetitive text with limited vocabulary
   - Saki Story: Longer, more diverse text with richer vocabulary
   - Saki Story should produce more varied and realistic word patterns, but with the current amount of data, Spamiam produces more legal English words due to its repetitive nature
6. Expected Improvements with Higher Order Models:
   - Higher order models capture more complex dependencies
   - Should produce more valid English words
   - Better approximation of natural language patterns

(g)

Applying the entropy formula for Markov chains we easily get

Spamiam.txt Entropy Rates:

- Zero-order (letter probabilities): 4.0927 bits/letter
- First-order (Markov chain): 1.9503 bits/letter
- Second-order (Markov chain): 1.3253 bits/letter

Saki Story.txt Entropy Rates:

- Zero-order (letter probabilities): 4.1533 bits/letter
- First-order (Markov chain): 3.0546 bits/letter
- Second-order (Markov chain): 1.9179 bits/letter

Analysis:

1. Entropy rates generally decrease with higher order models
2. Saki Story has higher entropy than Spamiam due to more diverse vocabulary
3. Higher order models capture more structure, reducing uncertainty
4. The entropy rate approaches the true language entropy as order increases