

Homework 6: Hidden Markov Models

UC Irvine CS275P: Graphical Models & Statistical Learning

Homework due at 11:59pm on May 29, 2025

Full Name:

Viveka Agrawal

UCINetID (e.g., ucinetid@uci.edu):

vivekaa @ uci.edu

Question 0: (5 points)

Canvas includes detailed guidelines on how homework solutions should be formatted for submission to Gradescope. To receive full credit on this question, be sure that you read and follow these guidelines carefully! In particular, please take care that:

- There are two Gradescope assignments for Homework 6. For the first assignment, submit a PDF containing your answers for questions 1-2. For the second assignment, submit your Python code for question 2, which will be checked by an autograder.
- For the PDF submission, you must fill in this PDF template with your answers. All pages must remain in their original order when uploading your assignment to Gradescope. *Do not rearrange, add, or remove any pages from the provided PDF template.*
- For multiple-choice questions, ensure that you *completely fill in the bubble next to your selected answer*, and provide an explanation or justification to support your answer.
- Enter your final answers (to a precision of 3 decimal places) in the answer boxes, and write your explanations and/or math justifications in the space provided below each question. *Always show the work needed to produce your answers, not just the final result.* Math may be handwritten, but if possible, please type any sentences.
- For question 2, use the provided Python file. After adding your code to the template, submit that file separately to the second Gradescope assignment. In addition to uploading your Python code, be sure you provide answers and explanations in the designated areas of the PDF. *Do not include code for question 2 in the PDF.*
- If any question asks you to create a plot, insert an image showing the plot in the PDF.
- Check that your answers (especially scanned math) are readable within Gradescope, and that content has not been cut-off by the page margins.

Question 1: (25 points)

We first examine a simple hidden Markov model (HMM). We observe a sequence of rolls of a four-sided die at an “occasionally dishonest casino”, where at time t the observed outcome $x_t \in \{1, 2, 3, 4\}$. At each of these times, the casino can be in one of two states $z_t \in \{1, 2\}$. When $z_t = 1$ the casino uses a fair die, while when $z_t = 2$ the die is biased so that rolling a 1 is more likely. In particular:

$$p(x_t = 1 | z_t = 1) = p(x_t = 2 | z_t = 1) = p(x_t = 3 | z_t = 1) = p(x_t = 4 | z_t = 1) = 0.25,$$

$$p(x_t = 1 | z_t = 2) = 0.7, \quad p(x_t = 2 | z_t = 2) = p(x_t = 3 | z_t = 2) = p(x_t = 4 | z_t = 2) = 0.1.$$

Assume that the casino has an equal probability of starting in either state at time $t = 1$, so that $p(z_1 = 1) = p(z_1 = 2) = 0.5$. The casino usually uses the same die for multiple iterations, but occasionally switches states according to the following probabilities:

$$p(z_{t+1} = 1 | z_t = 1) = 0.8, \quad p(z_{t+1} = 2 | z_t = 2) = 0.9.$$

The other transition probabilities you will need are the complements of these. For all parts below, include work showing how you computed your answer. *Hint: For all of the questions below, you should not need to enumerate all possible state sequences.*

- a) Under the HMM generative model, what is $p(z_1 = z_2 = z_3)$, the probability that the same die is used for the first three rolls?

$$p(z_1 = z_2 = z_3) = \boxed{0.725}$$

$$\begin{aligned} p(z_1 = z_2 = z_3) &= p(z_1 = 1, z_2 = 1, z_3 = 1) + \\ &\quad p(z_1 = 2, z_2 = 2, z_3 = 2) \\ &= p(z_1 = 1)p(z_2 = 1 | z_1 = 1)p(z_3 = 1 | z_2 = 1) + \\ &\quad p(z_1 = 2)p(z_2 = 2 | z_1 = 2)p(z_3 = 2 | z_2 = 2) \\ &= (0.5)(0.8)(0.8) + (0.5)(0.9)(0.9) \\ &= (0.5)(0.64 + 0.81) = 0.725 \end{aligned}$$

b) What is $p(x_2 = 1 | x_1 = 1)$, the probability that the second roll is a 1 given that the first roll was also 1? Is this different from $p(x_2 = 1)$, the probability that the second roll is a 1 given no information about the first roll?

$$p(x_2 = 1 | x_1 = 1) = \boxed{0.572}$$

$$\begin{aligned} p(z_2 = 1) &= p(z_1 = 1)p(z_2 = 1 | z_1 = 1) + \\ &\quad p(z_1 = 2)p(z_2 = 1 | z_1 = 2) \quad p(x_2 = 1) = \boxed{0.498} \\ &= (0.5)(0.8) + (0.5)(0.1) = 0.45 \end{aligned}$$

$$p(z_2 = 2) = 1 - 0.45 = 0.55$$

$$\begin{aligned} p(x_2 = 1) &= p(x_2 = 1 | z_2 = 1)p(z_2 = 1) + \\ &\quad p(x_2 = 1 | z_2 = 2)p(z_2 = 2) \\ &= (0.25)(0.45) + (0.7)(0.55) = 0.475 \end{aligned}$$

$$p(z_1 | x_1 = 1) \propto p(z_1)p(x_1 = 1 | z_1)$$

$$\left. \begin{aligned} p(z_1 = 1 | x_1 = 1) &\propto (0.5)(0.25) = 0.125 \\ p(z_1 = 2 | x_1 = 1) &\propto (0.5)(0.7) = 0.35 \end{aligned} \right\} 0.475$$

$$p(z_1 = 1 | x_1 = 1) = \frac{0.125}{0.475} \approx 0.263$$

$$p(z_1 = 2 | x_1 = 1) = \frac{0.35}{0.475} \approx 0.737$$

$$\begin{aligned} p(x_2 = 1 | x_1 = 1) &= 0.263((0.8)(0.25) + (0.2)(0.7)) + \\ &\quad 0.737((0.1)(0.25) + (0.9)(0.7)) \\ &= 0.572 \end{aligned}$$

c) Suppose that we observe the first two rolls. What is $p(z_1 = 1 \mid x_1 = 2, x_2 = 4)$, the probability that the casino used the fair die in the first roll?

$$p(z_1 = 1 \mid x_1 = 2, x_2 = 4) = \boxed{0.827}$$

$$p(z_1 = 1 \mid x_1, x_2) \propto \sum_{z_2} p(z_1) p(x_1 \mid z_1) p(z_2 \mid z_1) p(x_2 \mid z_2)$$

$$= p(x_1 = 2 \mid z_1 = 1) ((p(x_2 = 4 \mid z_2 = 1) p(z_2 = 1 \mid z_1 = 1) \\ + p(x_2 = 4 \mid z_2 = 2) p(z_2 = 2 \mid z_1 = 1)))$$

$$= 0.25((0.25)(0.8) + (0.1)(0.2)) = 0.055$$

$$p(x_1 = 2, x_2 = 4 \mid z_1 = 2) = 0.1((0.25)(0.1) + (0.1)(0.9)) \\ = 0.0115$$

$$p(x_1 = 2, x_2 = 4) = (0.055)(0.5) + (0.0115)(0.5) \\ = 0.03325$$

$$p(z_1 = 1 \mid x_1 = 2, x_2 = 4) = \frac{(0.055)(0.5)}{0.03325}$$

$$= \frac{0.0275}{0.03325}$$

$$= 0.827$$

- d) Consider a sequence of T rolls of the die, and let $z = (z_1, \dots, z_T)$ and $x = (x_1, \dots, x_T)$. What state sequence \hat{z} and observation sequence \hat{x} jointly maximize $p(z, x)$? You do not need to compute the numerical value of $p(\hat{z}, \hat{x})$, but justify your answer.

Maximize $p(z, x)$ by
finding most likely
state and most likely
observations given those states.

$$\hat{z} = (\hat{z}_1, \dots, \hat{z}_T) = (2, 2, \dots, 2)$$

$$\hat{x} = (\hat{x}_1, \dots, \hat{x}_T) = (1, 1, \dots, 1)$$

Maximizing $P(z_1)$: $P(z_1=2) = P(z_1=1) = 0.5$
Since both are equally likely, there is no preference
for z_1 .

Maximizing $P(z_{t+1}|z_t)$:

$$P(z_{t+1}=1|z_t=1) = 0.8 \quad P(z_{t+1}=2|z_t=1) = 0.2$$

$$P(z_{t+1}=1|z_t=2) = 0.1 \quad P(z_{t+1}=2|z_t=2) = 0.9$$

Since highest probability is 0.9 in state 2, choose $z_2=2$.

Maximizing $P(x_t|z_t)$:

If $z_t=1$, all $P(x_t|z_t=1) = 0.25$, so there is no preference.

If $z_t=2$, $P(x_t=1|z_t=2) = 0.7$ and everything else is 0.1,
so $x_t=1$ is preferred.

Question 2: (45 points)

We next learn hidden Markov models (HMMs) of the statistics of English text. In this application, each discrete “time” point corresponds to a single letter. For training, we use a chapter from Lewis Carroll’s *Alice’s Adventures in Wonderland*, available in `aliceTrainRaw.txt`. To simplify the modeling task, we first converted letters to lower-case and removed all punctuation. The resulting text, stored in `aliceTrain.txt`, is a sequence composed of 27 distinct characters (26 letters, as well as whitespace encoded via an underscore ‘_’).

In many applications of HMMs, there is insufficient data or computational resources to select the model order via cross-validation. In these situations, the state dimension is often selected via either the *Akaike information criterion (AIC)* or *Bayesian information criterion (BIC)*. Let $x = (x_1, \dots, x_T)$ denote the observed training sequence, $z = (z_1, \dots, z_T)$ a hidden state sequence, and $\hat{\theta}_M$ an ML estimate of the parameters for an HMM with M states:

$$\hat{\theta}_M = \arg \max_{\theta_M} p(x | \theta_M) = \arg \max_{\theta_M} \sum_z p(x | z, \theta_M) p(z | \theta_M), \quad x_t \in \{1, \dots, M\}.$$

For this model, the AIC and BIC take the following form:

$$\begin{aligned} \text{AIC}_M &= \log p(x | \hat{\theta}_M) - d(M), \\ \text{BIC}_M &= \log p(x | \hat{\theta}_M) - \frac{1}{2}d(M) \log(T). \end{aligned}$$

Here, $d(M)$ is the *number* of parameters (degrees of freedom) for an HMM with M states. The “best” model is then the one for which AIC_M or BIC_M is largest.

The `CategoricalHMM` class from the `hmmlearn` Python package provides an implementation of the expectation maximization (EM) algorithm for ML parameter estimation of HMMs for discrete data. *WARNING: Training HMMs of multiple orders via the EM algorithm may take an hour of computation time, so start early!*

- a) Use the `CategoricalHMM.fit` method, with default initialization, to learn HMMs with state dimensions of $M = 1, 5, 10, 15, 20, 30, 40, 50, 60$. (With $M = 1$ the HMM is a unigram model, which assumes that characters are independent.) For each of these models use a single random initialization, and run the EM algorithm for at most 500 iterations, or until the change in log-likelihood falls below 10^{-6} . Compute and store the log-likelihood which each model assigns to the training sequence. Save these models for later sections. (You do not need to show anything for this part.)

- b) Derive a formula for the number of parameters $d(M)$ in an HMM with M states, and observations taking W discrete values. Remember to account for normalization constraints: a discrete distribution on 4 events has only 3 degrees of freedom, since the probabilities must sum to one. Which criterion favors simpler models?

$$d(M) = (M-1) + M(M-1) + M(W-1)$$

Which model selection criterion favors simpler models?

- AIC BIC

Derivation of $d(M)$ and explanation for which criterion favors simpler models:

Count number of independent parameters for the following 3 components of HMM:

1. Initial state distribution ($M-1$)

- discrete distributions over M number of states
- since it must sum to 1, it can only have $M-1$ degrees of freedom

2. Transition Matrix ($M(M-1)$)

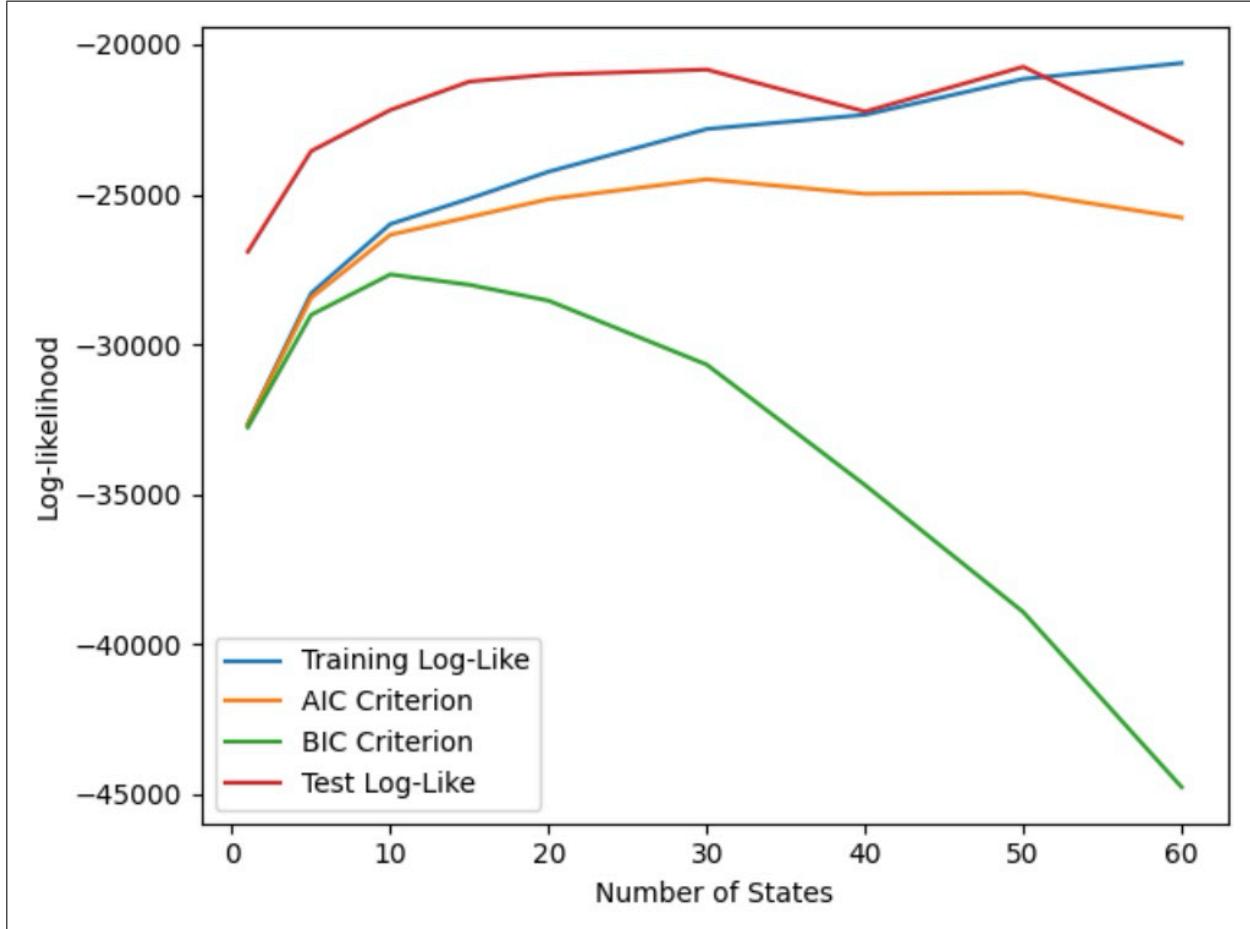
- matrix of size $M \times M$, each row is a prob distribution
- since it must sum to 1, it can only have $M-1$ degrees of freedom

3. Emission Matrix ($M(W-1)$)

- each hidden state emits one of W observable characters
- each M row has $W-1$ degrees of freedom

AIC penalizes model complexity linearly with $d(M)$ while BIC penalizes model complexity more strongly based on data size, so therefore BIC favors simpler models more strongly.

c) To test our learned HMMs, we use the text from a different chapter of Alice's Adventures in Wonderland, available in `aliceTest.txt`. Using `CategoricalHMM.score`, evaluate the test chapter's log-likelihood for each HMM learned in part (a). Plot these test log-likelihoods versus M , together with the training log-likelihood $\log p(x | \hat{\theta}_M)$, AIC_M , and BIC_M versus M for the HMMs learned in part (a). Which model selection criterion better predicted test performance?



Which model selection criterion better predicted test performance?

- AIC
- BIC

- d) Using the methods `CategoricalHMM.sample` and `num_to_text`, generate a random 500-character sequence from four different HMMs: the model with no sequential dependence ($M = 1$), the model with the highest BIC_M , the model with the highest AIC_M , and the most complex model ($M = 60$). Compare and contrast these sequences. What aspects of English text do they capture? What do they miss?

```

M = 1 (No sequential dependence):
t wlat s tedw nyh hsirt hv iidyeerhtsuvint a nmnaln nhtwes laa ia seieskbow
utntahotenyett uwliy়psitseiectt o odednv anau nychtg r teo ad htteoaoarhrm ko r
nilhhaauwoewshoecpioitohee nehehds ehstto ediat e e e d b tt onsoyhes es a owhbtg e
c esyue ihhtteuoelw tp e t m tgonvs wkptneaeyi o e hmlv iai toyuhl eyt cnue wwaowaor
ietttemncalnlrme eawie neq nieedw e od aanttneloiest de yni ebsttoe c t ee ni etsa
ecdwohss aee nmeiae d dn tun ittia h hlntss hede iw ca wnbmiwd sals wv o

M = 30 (Best AIC):
thon any the himc cish tyoulsroathe yon whe lumethor it to taoed a his ot the lume sh
ed tly and aaluthaky saomk taeanshe saing ne noce so tou thon svilk that whe boowore
reblot yonnh pisked srany it sremld mase i jot the no i somhutkk weu taily hudt asveo
sh he peilr u pece mutly it thed thing ing ther wurre a coonne thed os you sef spiny
leme ate thanlink af ole yosful wound alink at kher waw yow ou semellir b thery co wa
l lawhaaloulice slaip kher fo lacteasish cowuktt wo eainued sralinme to thi

M = 10 (Best BIC):
tht mivptheaclines or bes thal al emildnd theithe rotre mainf sunl tho careisensode a
ngiheryave norwottvdese ges lict thog the theavass i ce naos jufher ther mithivenchein
ak it goscher thod heged slemched for ent ind ate wang enlice pud thintgsiold bathe
of feithin sur the as whe ant ke geimet he ty id sh bon antr is wud onshe o wud au wa
ifo laren boso tlemd hit tithe th dasnurwleng il se yaem tetrif slrs ilte dotreter as
t noxe inl ankl i se petg mous te ded tet aluthial the afsisunwonnd nanse w

M = 60 (Most complex):
thent sreadome what thing aligh dored yor an the ruch teabyery aneaid the eare and as
a nurted tice then the treapomioh he nan mabch oud has pedli therer you is inan ande
d thiog leep a menle aaich the gite auice can it you sary oft dot and cime wotter pev
iogs thann and what harty som iss she dorincle ilich hay und the sas wo this thed at
thin tei ruke wame has sighnmhen mlanck to the tring tea rusking and hain nhitlice tw
iog wage druriouse to ther as than liitet thoulh the haid shawted twing as

```

Discussion of what aspects of English text are captured and missed by the HMMs:

$M=1$ model only shows basic letter frequencies and produces incoherent text that does not have a meaningful structure. $M=10$ model (BIC) shows frequent character sequences and short fragments of words but still is incoherent. $M=30$ model (AIC) shows plausible tokens which resemble words and fluent text, meaning that it has a better grasp of local dependencies. $M=60$ model has the most natural looking output with some fragments that look like real sentences, but it still does not have grammatical structure and could potentially be overfitting the data. Increasing the number of hidden states seems to improve local consistency.

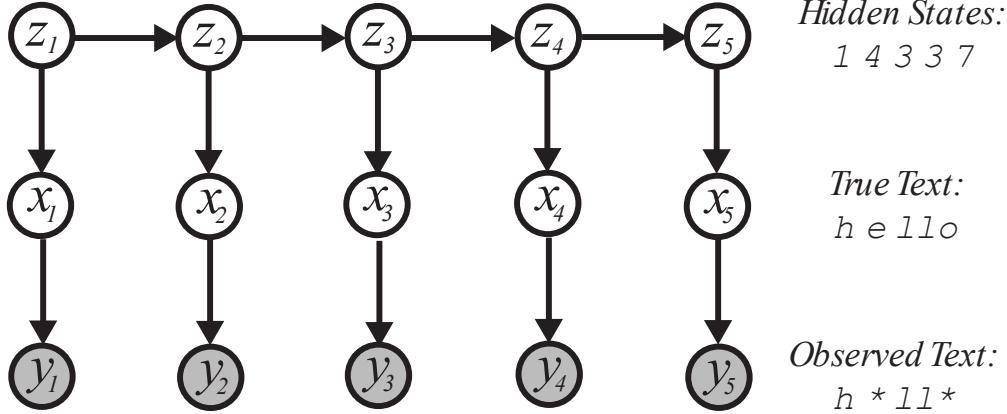


Figure 1: Graphical model illustrating an HMM with hidden states z_t which generate letters x_t . We observe a sequence y in which some of these letters have been *erased*.

In addition to computing likelihoods, HMMs lead to an efficient forward-backward algorithm which estimates the posterior probabilities of unobserved state sequences. We will now use this algorithm to estimate characters which have been *erased* from a text document.

Let $z_t \in \{1, \dots, M\}$ denote the hidden state at position t , and x_t the “true” character at position t of some document. Suppose that instead of observing x , we observe an alternative sequence y in which some letters have been erased. We assume that each letter is independently erased with probability ϵ , so that

$$p(Y_t = x_t | X_t = x_t) = 1 - \epsilon, \quad p(Y_t = * | X_t = x_t) = \epsilon.$$

Here, ‘*’ is a special erasure symbol. Figure 1 shows a graphical model describing this generative process. Note that we never observe an “incorrect” letter; y_t is always either identical to x_t , or the erasure symbol ‘*’. The demo code shows how to make a noisy version of the test sequence `aliceTest.txt` by randomly erasing letters with probability $\epsilon = 0.2$.

- e) Using the method `CategoricalHMM.predict_proba`, compute the posterior distributions $p(z_t | y)$ for the four models from part (d). To do this, exploit the fact that

$$p(y_t | z_t) = \sum_{x_t} p(y_t | x_t) p(x_t | z_t).$$

This implies that if we marginalize over the possible true letters x_t , we recover a standard HMM in which the observations y_t are independent given the hidden state sequence z . (You do not need to show anything for this part.)

- f) Suppose that we observe an erasure at position t , so that $p(y_t = * | x_t) = \epsilon$ remains constant as x_t is varied (since erasures provide no information about the underlying letter). Using the factored form of the generative model (as illustrated by the graph in Fig. 1), and the form of the observation model, show that the posterior distribution of x_t is

$$p(x_t | y, y_t = *) = \sum_{z_t} p(x_t, z_t | y, y_t = *) = \sum_{z_t} p(x_t | z_t) p(z_t | y),$$

where $p(z_t | y)$ is the posterior distribution of z_t given the full noisy sequence y .

Derivation of equation for posterior distribution of x_t :

1. Marginalization

$$p(x_t | y, y_t = *) = \sum_{z_t} p(x_t, z_t | y, y_t = *)$$

2. Chain Rule

$$= \sum_{z_t} p(x_t | z_t, y, y_t = *) \cdot p(z_t | y, y_t = *)$$

3. Conditional Independence assumption

$$= \sum_{z_t} p(x_t | z_t) \cdot p(z_t | y)$$

- g) Using the marginal distributions $p(z_t | y)$ from part (e), and the equation from part (f), determine the most likely missing letter for each erasure. Or, equivalently, implement the decision rule which minimizes the expected number of incorrect characters. Note that when we observe a letter $y_t \neq *$ at position t , this implies that $x_t = y_t$.

Most likely missing letter at position t , expressed in terms of $p(x_t | y)$ from part (f):

Original test text
length = 9567

Number of erased
characters = 1895

$\hat{x}_t =$ Erasure Rate = 0.198

h) Determine the percentage of missing letters which were correctly estimated by each model from part (d). What would chance performance be for this task? Print the first 500 characters of the denoised text produced by each model.

Accuracy for randomly choosing one of the 27 letters:

0.037

Accuracy for HMM with $M = 1$:

0.190

Accuracy for HMM with M selected by BIC:

0.396

Accuracy for HMM with M selected by AIC:

0.487

Accuracy for HMM with $M = 60$:

0.534

First 500 characters of denoised text for each of the four HMM models:

```
Denoising with M = 1 (No sequential dependence):
Accuracy on erased characters: 0.190 (360 out of 1895)
First 500 characters of denoised text:
the king and queen of heart w r se ted on he th one w en t ey arrived wit a gre
at crowd sembled about hem all sort o l ttle bi ds and b asts a well a th whd
l p ck of cards the k ave was sta ing efore the in chains ith a soldie n ac
ide to ua d im nd near the king as th hite ab it wi a rumpet i on hand
d a sc oll of archment in th o h n the very d e the c urt was ta le it
a e ish o tart upon it the lo ked s ood tha it m e alic u t

Denoising with M = 30 (Best AIC):
Accuracy on erased characters: 0.487 (923 out of 1895)
First 500 characters of denoised text:
the king and sueen of hearte wore serted on there thaone when they arrived wite a gre
at crowd a sembled about them all sort tou lttle bicks and boasts at well at the who
le pick of cards the koave was stalling before thed in chains tith a soldie in mace
tide to tuand him and near the king has the hite hab it wine a srumpet it ong hand a
nd a scooll of marchment in the o ht in the very a d ce it the court was a tattle wit
e a sarte lish ot tart upon it thed lotked so wood that it mace alice mutt

Denoising with M = 10 (Best BIC):
Accuracy on erased characters: 0.396 (750 out of 1895)
First 500 characters of denoised text:
the king and tueen of hearte wore se ted on here theone when they arrived wite a gre
at crowd asembed about them all sort on lottle binds and bhasts an well an the who
lt pack of cards the khave was stan ing tefore ther in chains sith a soldie an sach
ide to tuand sim and near the king has the thite sab it wine a trumpet in ond hand a
nd a sc oll of harchment in the othe an the very and te ae the churt was antatle hit
e a te he ish on tart eupon it ther lonked se tood than it more alice au t

Denoising with M = 60 (Most complex):
Accuracy on erased characters: 0.534 (1012 out of 1895)
First 500 characters of denoised text:
the king and aueen of hearte ware serted on their throne when they arrived wite a gre
at crowd a sembled about them all sort ou lttle binds and beasts at well at the who
le pick of cards the keave was starting before thes in chains tith a soldier in tace
tide to buaid im and near the king has the hite hab it wice a drumpet it one hand a
nd a scuoll of harchment in the othe in the very a dore it the court was a tarle tit
e a sarce tish ou tart bupon it thes losked so tood that it mare alice outt
```