APMA 1941D Project 1: Decoding a Substitution Cipher using MCMC

Sheridan Feucht

February 2023

Contents

1 Description of Project

1.1 Problem Setup

We have been given three texts that have been scrambled with a substitution cipher. Precisely, for each text, there is some scrambler function $\sigma \in S_{|A|}$, a permutation on A symbols that converts a string of English $a_1a_2...a_n$ to a scrambled code $b_1b_2...b_n$ where $b_i = \sigma(a_i)$. Section 1.5 of the lecture notes defines our best guess for this unknown permutation σ_* as

$$\sigma_* = \text{argmax}_{\sigma \in S_{|A|}} L(\sigma) \tag{1}$$

where the likelihood function for σ is defined in Equation (2), with Q being a ground truth transition probability matrix (which we obtain from a training corpus).

$$L(\sigma) = \mathbb{P}_{true}(\sigma^{-1}(b_1)) \prod_{j=1}^{n-1} Q(\sigma^{-1}(b_j), \sigma^{-1}(b_{j+1})) \tag{2}$$

In order to find $\sigma*$, our goal is to find the solution to (1), which involves finding the σ that maximizes (2). The problem is that the space of possible permutations is intractable to search through, motivating a Markov-chain based random search as described in the next section.

1.2 Methods

To find this value, we use an MCMC method, the Metropolis algorithm as described in Diaconis (2009).

1. Start with a preliminary guess for σ.
2. Compute Pl(σ). Here, define Pl (plausibility) as the probability of the backwards-translated text with the current guess for σ. This probability is calculated as Equation (2).
3. Get a new permutation $\sigma*$ by making a random transposition of the values that σ assigns to two symbols.
4. Compute Pl($\sigma*$).
   (a) If it is larger than Pl(σ), accept $\sigma*$.
   (b) If it is smaller than Pl(σ), accept $\sigma*$ regardless with probability Pl($\sigma*$)/Pl(σ). Otherwise stay at σ.
   (c) Repeat starting from Step 2.

During implementation, we actually cast this as a minimization problem by computing the negative likelihood, and also work in log space to avoid underflow from repeated multiplication of probabilities. We also add another parameter

β that allows for tuning of the probability in Step 4(b). We trained for 10,000 steps, which was overkill for most runs. We scraped English transition probabilities from the 19th-century Russian novel *War and Peace* by Leo Tolstoy (about 3 million characters).

2 Analysis and Discussion

2.1 Time to Converge

My implementation of the algorithm took a minute or two to run for 10000 steps. It was usually able to completely decode the text. However, for several random seeds, the model would get stuck at some local minimum, unable to climb out of that hole despite the extra coin-toss probability in place to prevent this exact issue. Perhaps I would have to finetune β to determine if that would address this issue (I was using $\beta = 1$ as mentioned in the lecture notes).

Below is a visualization of time taken to converge across seeds for each of the three texts given to us. We can clearly see a difference in performance based on the contents of the decoded text. For the first text, which was an excerpt from an English translation of a 19th-century Russian novel, convergence is relatively fast (roughly 2000 steps) compared to time taken to decode the Affordable Care Act (roughly 3000 steps or more). The third text, a piece of Irish fiction *Ulysses* from 1922, seems to be in between the other two in terms of difficulty. This difference in time to converge based on text content motivates the next set of experiments in Section 2.2.

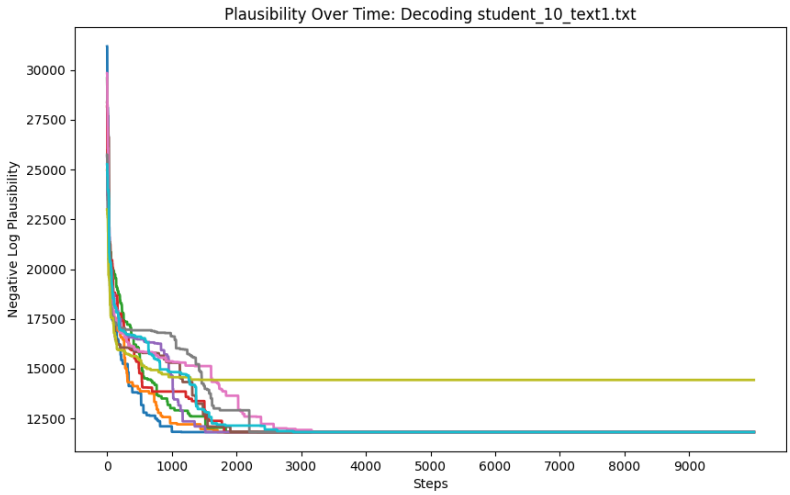

Figure 1: Number of steps taken to converge for model on the first text (*Dead Souls*). Each line shows the algorithm being run with a different seed (with 10 different seeds). Most runs converge at around 2000 steps.

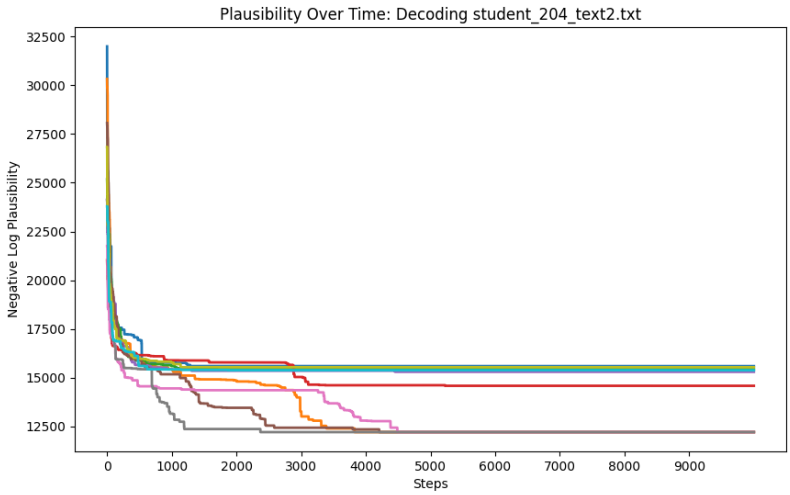

Figure 2: Number of steps taken to converge for model on the second text, part of the Affordable Care Act. Each line shows the algorithm being run with a different seed (with 10 different seeds). Here, runs generally take longer to converge than the model did in 1.

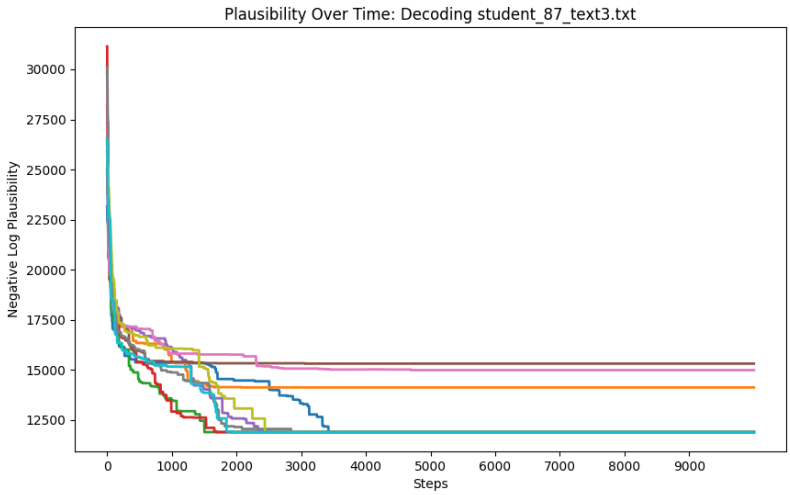

Figure 3: Number of steps taken to converge for model on the third text, *Ulysses* (1922). Each line shows the algorithm being run with a different seed (with 10 different seeds). The model also seems to take longer to converge for this document, but with slightly fewer runs getting stuck at local minima.

2.2 Does Training Corpus Matter?

As seen in Section 2.1, the model seems to be sensitive to the domain from which its training data is sampled when it comes to efficiency in decoding any given text. We compare the model using transition probabilities from several sources:

- *War and Peace* by Leo Tolstoy (3,013,732 characters)
- A small health policy document, U.S Public Law 111-163 111th Congress (137,995 characters)
- A large collection of U.S. health policy documents (3,109,557 characters)
- Works by James Joyce including *Finnegan's Wake*, *Dubliners*, and *A Portrait of the Artist as a Young Man*, but not including *Ulysses*. (2,073,117 characters)

Literature data is taken from Project Gutenberg and preprocessed, lowercasing text and removing special characters. U.S. policy data is taken from GovInfo, maintained by the U.S. Government Publishing Office. It is preprocessed in the same way.

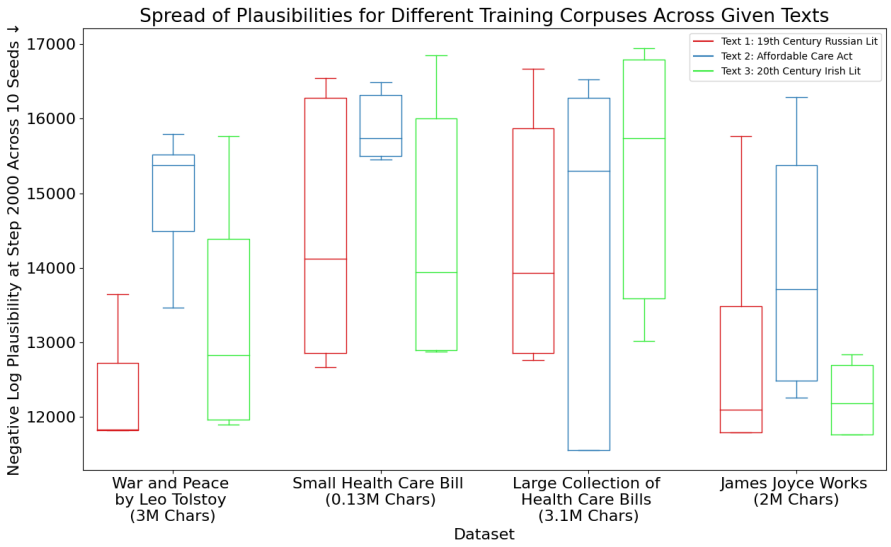

Figure 4: Plotting negative log plausibility across ten seeds for each training corpus and each text (lower is better). The leftmost red boxplot shows that negative log plausibilities were consistently low when *War and Peace* was used to decode Text 1, an excerpt from another 19th-century Russian novel. This success is likely because the train and test data were drawn from similar distributions. In contrast, the blue boxplot next to it shows that *War and Peace* is not as useful for decoding an excerpt from the Affordable Care Act. Looking at the two middle groups of boxplots, we see that plausibility has a high variance for runs that use transition probabilities mined from United States health care bills, regardless of corpus size. Finally, on the right we see that a training corpus consisting of James Joyce's work improves performance when decoding his most famous book, *Ulysses* (see the rightmost green boxplot).

As seen in Figure 4, performance degrades when trained on any type of health care bill data, regardless of dataset size (although a large dataset can *sometimes* do well when decoding other health care bills). This may be because U.S. health care bills contain a lot of numbers and special characters interspersed with legal jargon. After preprocessing, much of the data becomes almost unreadable. Thus there may be too much noise in the data to provide as much significant information about English transition probabilities.

2.3 Fine-tuning β: An Inconclusive Glance

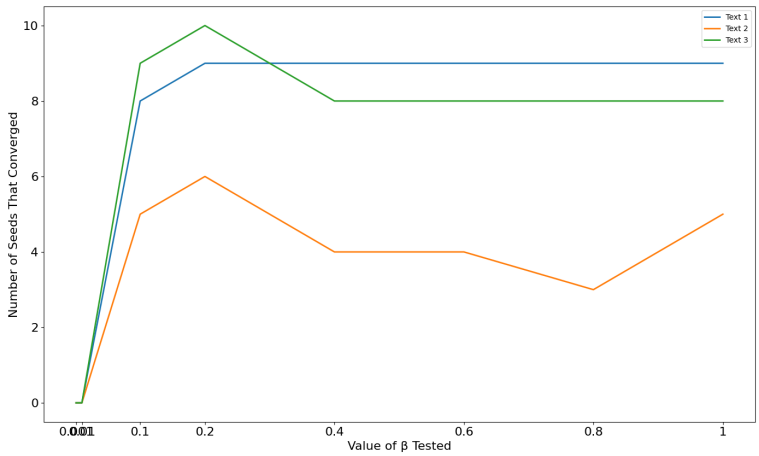

Figure 5: Figure showing number of runs that converged (out of 10) for different values of β. Each color shows performance on a different text.

This experiment was motivated by the fact that the algorithm was often getting stuck at local minimums. Intuitively, decreasing the value of beta would make it more and more likely that the model would escape any local minima, as the probability of going "uphill" would increase. I would expect to see an increase in number of runs converging as β decreases, up until some point where this increased tendency to "go the wrong way" becomes counterproductive. We do see a possible trend here, but I recognize that testing ten seeds for each value of beta is not enough data to show anything conclusive. Regardless, I still thought this graph posed an interesting research question, and would want to explore further given more computational resources.

3 Implementation Details

See repository on Github for full implementation details and data [here].

3.1 Code Sample: Mining for Transition Probabilities

```
import numpy as np
CHARS = list(ascii_lowercase) + [" "]
```

```
"""Helper function to obtain the transition matrix M, where each row has
    probabilities of what characters should come next. e.g. M[0][1]
    will have probability of 'b' coming after 'a'.
"""
def get_M(text):
    matrix = []
    for a in CHARS:
        row = []
        for b in CHARS:
            row.append(text.count(a+b))
        matrix.append(row)
    matrix = np.array(matrix) + 1.0 # pseudo-laplace smoothing
    matrix /= matrix.sum(axis=1, keepdims=True)
    return matrix
```

3.2 Code Sample: MCMC Algorithm to Decode Text

```
"""function that takes in a transition matrix M and a scrambled text and
    uses the Metropolis MCMC method to unscramble the text. Returns the
    optimal f: scrambled->unscrambled
"""
def mcmc(M, scrambled, beta=BETA):
    f = np.random.permutation(len(CHARS)) # start with preliminary guess
        for f (randomized array mapping)
    plausibilities = []

    i = 0
    while True:
        # compute Pl(f) using formula above
        pl_f = neglog_plausibility(M, scrambled, f)

        # change to f* by doing a random change ## f[[0, 26]] = f[[26,
            0]]
        idx1, idx2 = np.random.choice(np.arange(len(CHARS)), size=(2,),
            replace=False)
        fstar = np.copy(f)
        fstar[[idx1, idx2]] = fstar[[idx2, idx1]]

        # compute -logPl(f*) and then compare it to -logPl(f). We want
            to minimize
        pl_fstar = neglog_plausibility(M, scrambled, fstar)
        if pl_fstar < pl_f:
            f = fstar
        elif pl_fstar >= pl_f:
            flip = random.uniform(0, 1)
            if flip <= np.exp(-beta * (pl_fstar - pl_f)):
                f = fstar # accept the new one with this probability
            else:
```

```
            pass

        # document what's happening and break if needed
        plausibilities.append(pl_f)
        if i % 100 == 0:
            print(f"Step {i}: Plausibility {pl_f}")
        if i % 2000 == 0 and i != 0:
            print(unscramble(scrambled, f))

        i += 1
        if i > STOP:
            break

    return f, plausibilities
```

3.3 Results

3.3.1 Decoded Text 1

Decoded version of `student_10_text1.txt`. From *Dead Souls* by Nikolai Vasilievich Gogol (1842).

```
ashed himself rubbed himself from head to foot with a wet sponge a
    performance executed only on sundays and the day in question
    happened to be a sunday shaved his face with such care that his
    cheeks issued of absolutely satin like smoothness and polish donned
    first his bilberry coloured spotted frockcoat and then his bearskin
    overcoat descended the staircase attended throughout by the waiter
    and entered his britchka with a loud rattle the vehicle left the
    inn yard and issued into the street a passing priest doffed his cap
    and a few urchins in grimy shirts shouted gentleman please give a
    poor orphan a trifle presently the driver noticed that a sturdy
    young rascal was on the point of climbing onto the splashboard
    wherefore he cracked his whip and the britchka leapt forward with
    increased speed over the cobblestones at last with a feeling of
    relief the travellers caught sight of macadam ahead which promised
    an end both to the cobblestones and to sundry other annoyances and
    sure enough after his head had been bumped a few more times against
    the boot of the conveyance chichikov found himself bowling over
    softer ground on the town receding into the distance the sides of
    the road began to be varied with the usual hillocks fir trees
    clumps of young pine trees with old scarred trunks bushes of wild
    juniper and so forth presently there came into view also strings of
    country villas which with their carved supports and grey roofs the
    latter looking like pendent embroidered tablecloths resembled
    rather bundles of old faggots likewise the customary peasants
    dressed in sheepskin jackets could be seen yawning on benches
    before their huts while their womenfolk fat of feature and swathed
    of bosom gazed out of upper windows and the windows below displayed
```

here a peering calf and there the unsightly jaws of a pig in short the view was one of the familiar type after passing the fifteenth verst stone chichikov suddenly recollected that according to manilov fifteen versts was the exact distance between his country house and the town but the sixteenth verst stone flew by and the said country house was still nowhere to be seen in fact but for the circumstance that the travellers happened to encounter a couple of peasants they would have come on their errand in vain to a query as to whether the country house known as zamanilovka was anywhere in the neighbourhood the peasants replied by doffing their caps after which one of them who seemed to boast of a little more intelligence than his companion and who wore a wedge shaped beard made answer perhaps you mean manilovka not zamanilovka yes yes manilovka manilovka eh well you must continue for another verst and then you will see it straight before you on the right on the right re echoed the coachman yes on the right affirmed the peasant you are on the proper road for manilovka but zamanilovka well there is no such place the house you mean is called manilovka because manilovka is its name but no house at all is called zamanilovka the house you mean stands there on that hill and is a stone house in which a gentleman lives and its name is manilovka but zamanilovka does not stand hereabouts nor ever has stood so the travellers proceeded in search of manilovka and after driving an additional two versts arrived at a spot whence there branched off a by road yet two three or four versts of the by road had been covered before they saw the least sign of a two storied stone mansion then it was that chichikov suddenly recollected that when a friend has invited one to visit his country house and has said that the distance thereto is fifteen versts the distance is sure to turn out to be at least thirty not many people would have admired the situation of manilovs abode for it stood on an isolated rise and was open to every wind that blew on the slope of the rise lay closely mown turf while disposed here and there after the english fashion were flower beds containing clumps of lilac and yellow acacia also there were a few insignificant groups of slender leaved pointed tipped birch trees with under two of the latter an arbour having a shabby green cupola some blue painted wooden supports and the inscription this is the temple of solitary thought lower down the slope lay a green coated pond green coated ponds constitute a frequent spectacle in the gardens of russian landowners and lastly from the foot of the declivity there stretched a line of mouldy log built huts which for some obscure reason or another our hero set himself to count up to two hundred or more did he count but nowhere could he perceive a single leaf of vegetation or a single stick of timber the only thing to greet the eye was the logs of which the huts were constructed nevertheless the scene was to a certain extent enlivened by the spectacle of two peasant women who with clothes picturesquely tucked up were wading knee deep in the pond and dragging behind them with wooden handles a ragged fishing net in the meshes of which t

3.3.2 Decoded Text 2

Decoded version of `student_204_text2.txt`. This is an excerpt from the Affordable Care Act.

```
ed to reduce premium costs for an entity described in subsection abi or
    to reduce premium contributions co payments deductibles co
    insurance or other out of pocket costs for plan participants such
    payments shall not be used as general revenues for an entity
    described in subsection abi the secretary shall develop a mechanism
    to monitor the appropriate use of such payments by such entities
    payments not treated as income payments received under this
    subsection shall not be included in determining the gross income of
    an entity described in subsection abi that is maintaining or
    currently contributing to a participating employment based plan
    appeals the secretary shall establish a an appeals process to
    permit participating employment based plans to appeal a
    determination of the secretary with respect to claims submitted
    under this section and b procedures to protect against fraud waste
    and abuse under the program d audits the secretary shall conduct
    annual audits of claims data submitted by participating employment
    based plans under this section to ensure that such plans are in
    compliance with the rejuirements of this section e funding there is
    appropriated to the secretary out of any moneys in the treasury not
    otherwise appropriated to carry out the program under this section
    such funds shall be available without fiscal year limitation f
    limitation the secretary has the authority to stop taking
    applications for participation in the program based on the
    availability of funding under subsection e page stat sec immediate
    note deadlines usc information that allows consumers to identify
    affordable coverage options a internet portal to affordable
    coverage options immediate establishment not later than quly the
    secretary in consultation with the states shall establish a
    mechanism including an internet website through which a resident of
    any state may identify affordable health insurance coverage options
    in that state connecting to affordable coverage an internet website
    established under paragraph shall to the extent practicable provide
    ways for residents of any state to receive information on at least
    the following coverage options a health insurance coverage offered
    by health insurance issuers other than coverage that provides
    reimbursement only for the treatment or mitigation of i a single
    disease or condition or ii an unreasonably limited set of diseases
    or conditions as determined by the secretary b medicaid coverage
    under title xix of the social security act c coverage under title
    xxi of the social security act d a state health benefits high risk
    pool to the extent that such high risk pool is offered in such
    state and e coverage under a high risk pool under section b
```

```
enhancing comparative purchasing options in general not later note
deadline standard format than days after the date of enactment of
this act the secretary shall develop a standardized format to be
used for the presentation of information relating to the coverage
options described in subsection a such format shall at a minimum
rejuire the inclusion of information on the percentage of total
premium revenue expended on nonclinical costs as reported under
section a of the public health service act eligibility availability
premium rates and cost sharing with respect to such coverage
options and be consistent with the standards adopted for the
uniform explanation of coverage as provided for in section of the
public health service act use of format the secretary shall utilize
the format developed under paragraph in compiling information
concerning coverage options on the internet website established
under subsection a c authority to contract the secretary may carry
out this section through contracts entered into with jualified
entities sec administrative simplification a purpose of
administrative simplification section of the health insurance
portability and accountability act of usc d note is amended by
inserting uniform before standards and by inserting and to reduce
the clerical burden on patients health care providers and health
plans before the period at the end b operating rules for health
information transactions page stat definition of operating rules
section of the social security act usc d is amended by adding at
the end the following operating rules the term operating rules
means the necessary business rules and guidelines for the
electronic exchange of information that are not defined by a
standard or its implementation specifications as adopted for
purposes of this part transaction standards operating rules and
compliance section of the social security act usc d is amended a in
subsection a by adding at the end the following new subparagraph q
electronic funds transfers b in subsection a by adding at the end
the following new paragraph rejuirements for financial and
administrative transactions a in general the standards and
associated operating rules adopted by the secretary shall i to the
extent feasible and appropriate enable determination o
```

3.3.3 Decoded Text 3

Decoded version of `student_87_text3.txt`. This seems to be an excerpt from *Ulysses* by the Irish writer James Joyce (1922).

```
nuzzling thirstily her clove of orange shouts rang shrill from the boys
    playfield and a whirring whistle again a goal i am among them
    among their battling bodies in a medley the joust of life you mean
    that knockkneed mothers darling who seems to be slightly crawsick
    jousts time shocked rebounds shock by shock jousts slush and
    uproar of battles the frozen deathspew of the slain a shout of
    spearspikes baited with mens bloodied guts now then mr deasy said
```

rising he came to the table pinning together his sheets stephen stood up i have put the matter into a nutshell mr deasy said its about the foot and mouth disease just look through it there can be no two opinions on the matter may i trespass on your valuable space that doctrine of laissez faire which so often in our history our cattle trade the way of all our old industries liverpool ring which jockeyed the galway harbour scheme european conflagration grain supplies through the narrow waters of the channel the pluterperfect imperturbability of the department of agriculture pardoned a classical allusion cassandra by a woman who was no better than she should be to come to the point at issue i dont mince words do i mr deasy asked as stephen read on foot and mouth disease known as kochs preparation serum and virus percentage of salted horses rinderpest emperors horses at mrzsteg lower austria veterinary surgeons mr henry blackwood price courteous offer a fair trial dictates of common sense allimportant question in every sense of the word take the bull by the horns thanking you for the hospitality of your columns i want that to be printed and read mr deasy said you will see at the next outbreak they will put an embargo on irish cattle and it can be cured it is cured my cousin blackwood price writes to me it is regularly treated and cured in austria by cattledoctors there they offer to come over here i am trying to work up influence with the department now im going to try publicity i am surrounded by difficulties by intrigues by backstairs influence by he raised his forefinger and beat the air oldly before his voice spoke mark my words mr dedalus he said england is in the hands of the jews in all the highest places her finance her press and they are the signs of a nations decay wherever they gather they eat up the nations vital strength i have seen it coming these years as sure as we are standing here the jew merchants are already at their work of destruction old england is dying he stepped swiftly off his eyes coming to blue life as they passed a broad sunbeam he faced about and back again dying he said again if not dead by now the harlots cry from street to street shall weave old englands windingsheet his eyes open wide in vision stared sternly across the sunbeam in which he halted a merchant stephen said is one who buys cheap and sells dear jew or gentile is he not they sinned against the light mr deasy said gravely and you can see the darkness in their eyes and that is why they are wanderers on the earth to this day on the steps of the paris stock exchange the goldskinned men quoting prices on their gemmed fingers gabble of geese they swarmed loud uncouth about the temple their heads thickplotting under maladroit silk hats not theirs these clothes this speech these gestures their full slow eyes belied the words the gestures eager and unoffending but knew the rancours massed about them and knew their zeal was vain vain patience to heap and hoard time surely would scatter all a hoard heaped by the roadside plundered and passing on their eyes knew their years of wandering and patient knew the dishonours of their flesh who has not stephen said what do you mean mr deasy asked he

```
came forward a pace and stood by the table his underjaw fell
sideways open uncertainly is this old wisdom he waits to hear from
me history stephen said is a nightmare from which i am trying to
awake from the playfield the boys raised a shout a whirring
whistle goal what if that nightmare gave you a back kick the ways
of the creator are not our ways mr deasy said all human history
moves towards one great goal the manifestation of god stephen
jerked his thumb towards the window saying that is god hooray ay
whrrwhee what mr deasy asked a shout in the street stephen
answered shrugging his shoulders mr deasy looked down and held for
awhile the wings of his nose tweaked between his fingers looking
up again he set them free i am happier than you are he said we
have committed many errors and many sins a woman brought sin into
the world for a woman who was no better than she should be helen
the runaway wife of menelaus ten years the greeks made war on troy
a faithless wife first brought the strangers to our shore here
macmurroughs wife and her leman orourke prince of breffni a woman
too brought parnell low many errors many failures but not the one
sin i am a struggler now at the end of my days but i will fight
for the right till the end for ulster will fight and ulster will
be right stephen raised th
```

---

4 References

The Markov Chain Monte Carlo Revolution by Persi Diaconis, Bulletin of the American Mathematical Society 46.2 (2009): 179-205.

Pattern Theory: Old and New by Govind Menon, APMA 1941D Lecture Notes (2022).