Andrew Wu 6.437 Project Part I.

Problem 1:

a) For observed cipher text y, we model english as a Markov chain, so the likelihood is given by

$$L(y) = P(X_1 = f^{-1}(y_1)) \cdot M_{f^{-1}(y_2), f^{-1}(y_1)} \cdot M_{f^{-1}(y_3), f^{-1}(y_2)} \cdots$$

$$p_{y|f}(y|f) = \underbrace{P(X_1 = f^{-1}(y_1))}_{P_{f^{-1}(y_1)}} \cdot \prod_{i=1}^{n-1} M_{f^{-1}(y_{i+1}), f^{-1}(y_i)}$$

Since $M^{-1}_{f^{-1}(y_{i+1}), f^{-1}(y_i)}$ is the probability of moving from $X_i \to X_{i+1}$, which is equivalent to the probability or moving from $y_i \to y_{i+1}$, since y is just x transformed by the cipher f.

b) By Baye's Rule $p_{f|y}(f|y) = \dfrac{p(y|f) \cdot p(f)}{p(y)}$

we model the set of ciphers as uniformly drawn from from all permutations of the alphabet, so $p(f) = \frac{1}{28!}$, since $|A| = 28$.

$p(y)$ is the marginalization over f, so

$$P_{f|y}(f|y) = \frac{1}{28!} \cdot \frac{P(X_1 = f^{-1}(y_1)) \cdot \prod_{i=1}^{N-1} M[f^{-1}(y_{i+1}), f^{-1}(y_i)]}{\sum_{f'} P(X_1 = f'^{-1}(y_1)) \cdot \prod_{i=1}^{N-1} M[f'^{-1}(y_{i+1}), f'^{-1}(y_i)]}$$

c) $\hat{f}_{MAP}$ is the cipher that maximizes this quantity. Only the numerator, $p_{y|f}(y|f)$ depends on f, but since M, f are discrete, we have to compute:

$$\hat{f}_{MAP} = \arg\max_f P(X_1 = f^{-1}(y_1)) \cdot \prod_{i=1}^{N-1} M[f^{-1}(y_{i+1}), f^{-1}(y_i)]$$

for every f, which is infeasible over $28!$ possible f.

Problem 2

a) Assuming the ciphers are uniformly distributed over 28! permutations, then for

f_1, f_2. The number of ciphers f_2 which differ in precisely 2 locations is $\binom{28}{2}$, since we pick 2 mappings in f_1 and swap them

Thus, f_2 differs from f_1 with probability

$$\boxed{\frac{\binom{28}{2}}{28!}} \quad \frac{28\cdot 27}{2\cdot 28!} = \frac{1}{2\cdot 26!}$$

b) Following the hint, we start with the proposal distribution over the set of ciphers f, where

~~$V(f'|f) = \frac{\binom{28}{2}}{28!}$ if f' and f differ in exactly~~

$$V(f'|f) = \begin{cases} \frac{1}{\binom{28}{2}} & \text{if } f' \text{ and } f \text{ differ in exactly 2 symbol assignments} \\ 0 & \text{otherwise.} \end{cases}$$

Then, we define a markov chain with transition probabilities

$$W(f'|f) = V(f'|f)\cdot \alpha(f\to f') \quad \text{from MH}$$
$$= V(f'|f)\cdot \min\left\{1, \frac{P_{y|f}(y|f')\cdot V(f|f')}{P_{y|f}(y|f)\cdot V(f'|f)}\right\}$$

But $V(f|f') = V(f'|f)$ by our definition.

So our Markov Chain has transition probability matrix over the set of ciphers f given by

$$W(f'|f) = V(f'|f) \cdot \frac{P_{Y|f}(y|f')}{P_{Y|f}(y|f)}$$

$$V(f'|f) = \frac{1}{\binom{28}{2}}$$

where $P_{Y|f}$ was defined before as $P(f^{-1}(y_1)) \prod_{i=1}^{N-1} M_{f^{-1}(y_{i+1}), f^{-1}(y_i)}$

By M-H, ~~this~~ since $P_{f|y} \sim P_{Y|f}$, the normalized stationary distribution is precisely the posterior, $P_{F|y}$.

c) With the above proposal distribution, the MCMC-based MH algorithm is then:

Initialize an arbitrary decoder f as a permutation of A.

for k iterations:

Draw some f' from $V(f'|f^i)$

acceptance factor $a = \min\left(1, \frac{P_{Y|f}(y|f')}{P_{Y|f}(y|f^i)}\right)$

X is drawn from ~~Bernoulli~~ Uniform $[0,1]$

If $X < a$, we accept:

$f^{i+1} = f'$

Otherwise

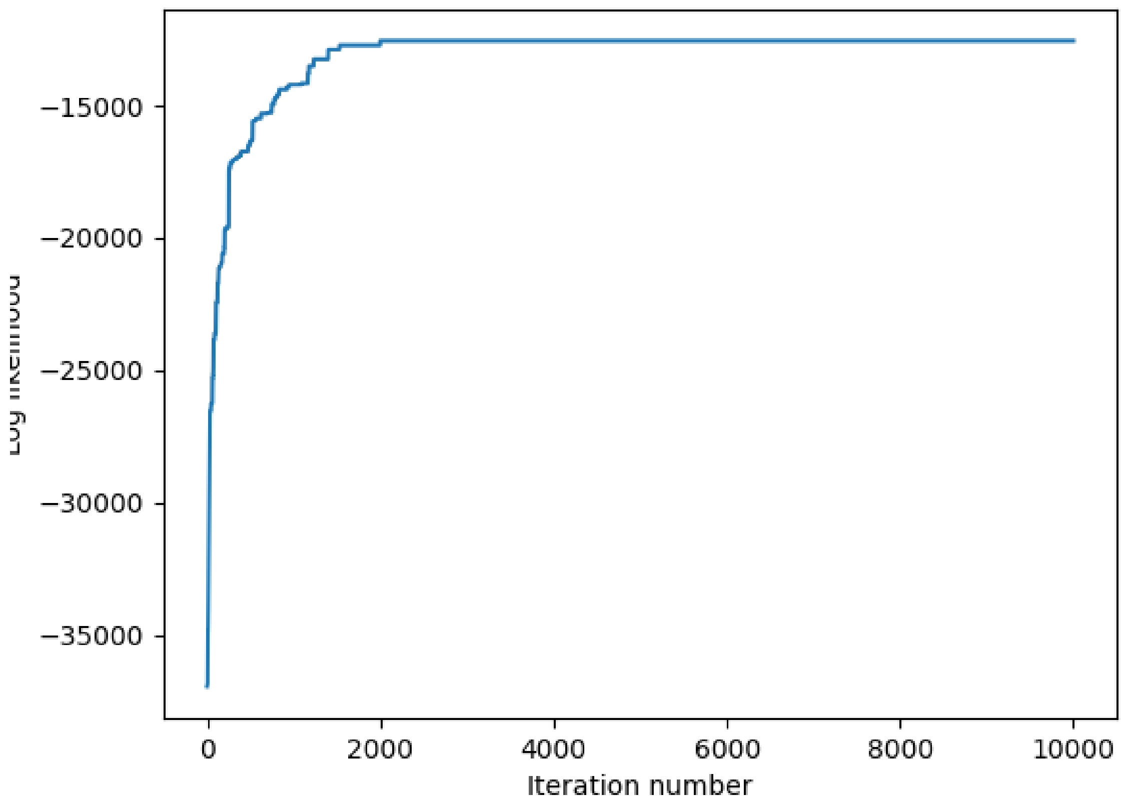
$f^{i+1} = f^i$

Return f^k, the converged decoder f.

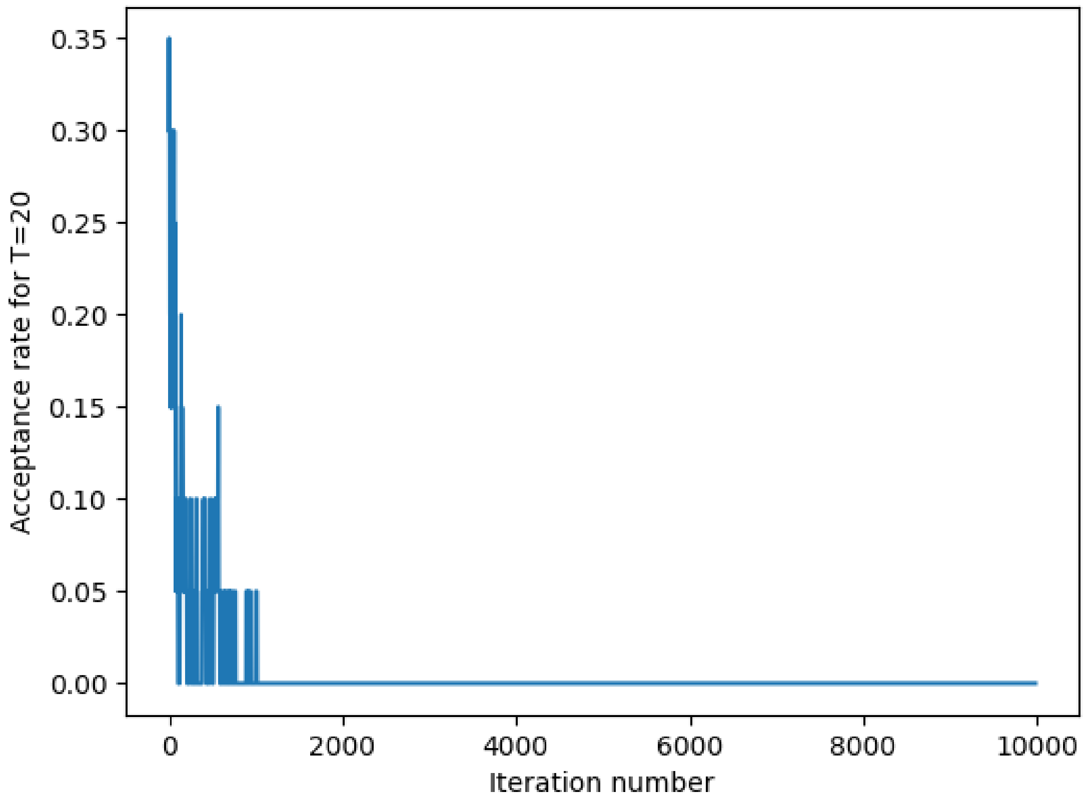
6.437 Metropolis Hasting Algorithm Implementation

Problem 3

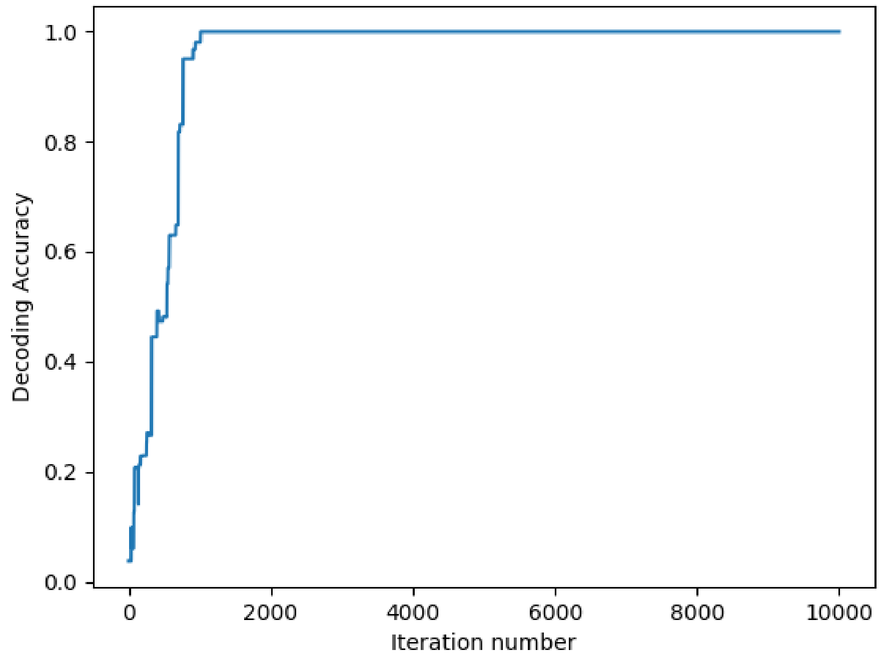
a) The log-likelihood of the accepted state plotted as a function of the iteration count is shown below.



b) The acceptance rate for the choice of T = 20 is shown below (note the total number of iterations is 10000).



c) The decoding accuracy vs iteration is plotted below.



d) Truncating the input text seems to not have much impact on the final decoding accuracy, as shown below, which can still reproduce the text relatively accurately. This is because as long as the input length is sufficient, the Markov Chain assumption holds, so regardless of start location the decoding is accurate.

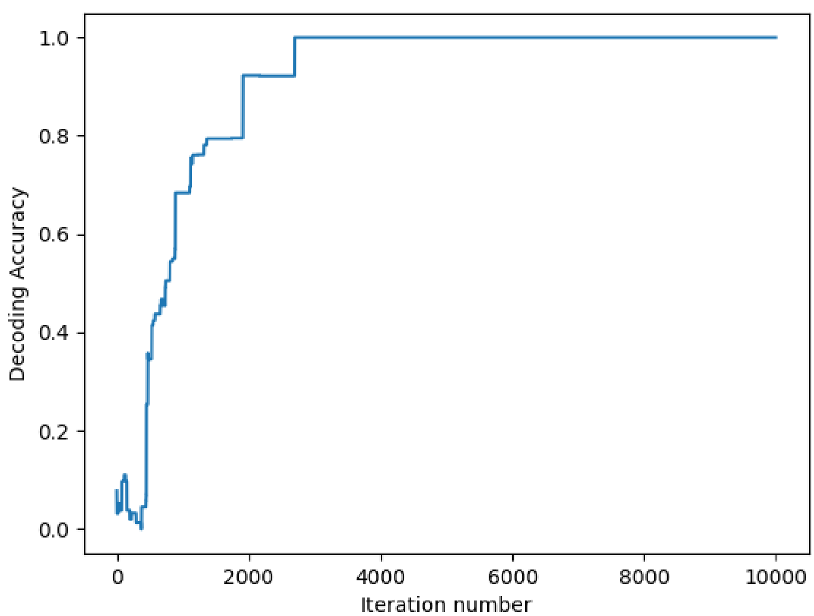
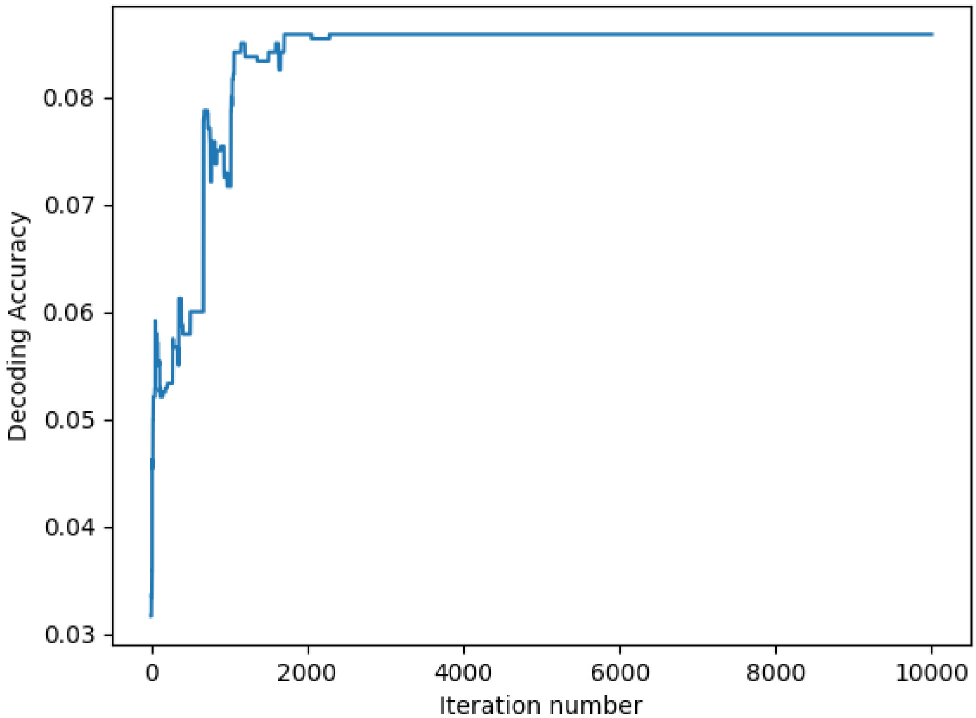
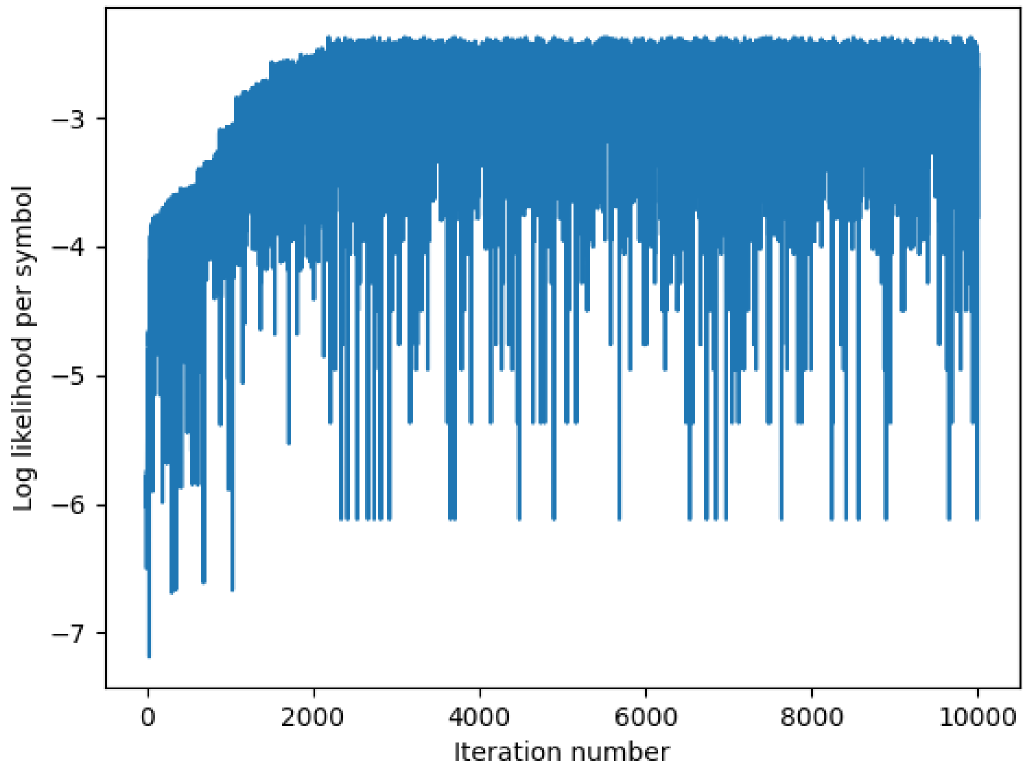


Figure 1 Decoding Accuracy when only the first half of the cipher text is used as input.

However, reducing the input size too small removes the asymptotic properties of the Markov chain, so when only 100 characters are used as input, we get the following, where the accuracy is limited to 0.08. When the input size is too small, the likelihood of the “true” cipher is no longer exponentially larger than other ciphers, and thus the maximum likelihood can converge to some suboptimal choices.



e) The log-likelihood per symbol over iterations is plotted below:



The final equilibrium likelihood value is -2.46. Computing the entropy using the probability distribution over the alphabet given as data, we compute a (negative) entropy of -2.853, which is relatively close to the equilibrium value. This is precisely what we expect, since the true decoded distribution should match in likelihood/Entropy with English since it is English. That is, the distribution of symbols within the decoded sequence of English should be close to the "true" distribution/entropy of real English.