MIT 6.437: Inference and Information **Spring 2018**

Project 1

Student Name: Jon Paul Janet

Problem 1

a Since f is one-to-one:

$$\begin{aligned} y_i &= f(x_i) \quad \forall i \in \{1, 2, \ldots, n\} \\ p_{\mathbf{x}}\left(\mathbf{x} = \{x_1, x_2, \ldots x_n\}\right) &= P(x_1, x_2, \ldots, x_n) \\ &= P(x = x_1) P(x_2, \ldots, x_n | x_1 = x_1) \\ &= p_x\left(x_1\right) p_{x_2|x_3}\left(x_2|x_3\right) p_{x_3|x_4}\left(x_3|x_4\right) \ldots p_{x_n|x_{n-1}}\left(x_n|x_{n-1}\right) \\ &= p_x\left(x_1\right) M_{(x_2),(x_3)} M_{(x_3),(x_4)} \ldots M_{(x_{n-1}),(x_n)} \end{aligned}$$

Where $M_{(x_i),(x_j)}$ represents the transition probability from the symbol of x_i to the symbol of x_j. The probability of y can then be expressed, knowing f:

$$\begin{aligned} p_{y_1|f}\left(y_1|f\right) &= \sum_{\substack{x' : f(x') = y_1 \\ x' \in \mathcal{A}}} p_x(x') = p_x\left(f^{-1}(y_1)\right) \\ p_{\mathbf{y}|f}\left(\mathbf{y}|f\right) &= p_{\mathbf{x}}\left(f^{-1}(\mathbf{y})\right) \\ &= p_x\left(f^{-1}(\mathbf{y})_1\right) M_{(f^{-1}(\mathbf{y})_2),(f^{-1}(\mathbf{y})_3)} M_{(f^{-1}(\mathbf{y})_3),(f^{-1}(\mathbf{y})_4)} \ldots M_{(f^{-1}(\mathbf{y})_{n-1}),(f^{-1}(\mathbf{y})_n)} \end{aligned}$$

---

b

$$\begin{aligned} p_{f|\mathbf{y}}\left(f|\mathbf{y}\right) &= \frac{p_{\mathbf{y}|f}\left(\mathbf{y}|f\right) p_f\left(f\right)}{p_{\mathbf{y}}\left(\mathbf{y}\right)} \\ p_{\mathbf{y}}\left(\mathbf{y}\right) &= \sum_{\mathbf{y}} p_{\mathbf{y}|f}\left(\mathbf{y}|f\right) p_f\left(f\right) \end{aligned}$$

As it will be useful later:

$$\begin{aligned} p_{f|\mathbf{y}}\left(f|\mathbf{y}\right) &= \frac{p_{\mathbf{y}|f}\left(\mathbf{y}|f\right) p_f\left(f\right)}{p_{\mathbf{y}}\left(\mathbf{y}\right)} \\ \implies \log p_{f|\mathbf{y}}\left(f|\mathbf{y}\right) &= \log p_{\mathbf{y}|f}\left(\mathbf{y}|f\right) + \log p_f\left(f\right) - \log p_{\mathbf{y}}\left(\mathbf{y}\right) \\ &= \left[\log p_x\left(f^{-1}(\mathbf{y})_1\right) + \sum_{i=1}^{n} M_{(f^{-1}(\mathbf{y})_{i-1}),(f^{-1}(\mathbf{y})_i)}\right] + \log|\mathcal{A}| - \log p_{\mathbf{y}}\left(\mathbf{y}\right) \end{aligned}$$

Here, we can compute $p_{\mathbf{y}|f}(\mathbf{y}|f)$ for any f in terms of the given M and we assume a uniform prior on f across the space of all 28-key permutations $\mathcal{F}$, i.e. $p_f(f) = \frac{1}{28!}$. The map can be defined as

$$\begin{aligned}\hat{f}_{MAP}(\mathbf{y}) &= \arg\max_{f\in\mathcal{F}} \left[p_{f|\mathbf{y}}(f|\mathbf{y})\right] \\ &= \arg\max_{f\in\mathcal{F}} \left[\frac{p_{\mathbf{y}|f}(\mathbf{y}|f)\, p_f(f)}{p_{\mathbf{y}}(\mathbf{y})}\right] \\ &= \arg\max_{f\in\mathcal{F}} \left[p_{\mathbf{y}|f}(\mathbf{y}|f)\, p_f(f)\right] = \arg\max_{f\in\mathcal{F}} \left[\log p_{\mathbf{y}|f}(\mathbf{y}|f)\, p_f(f)\right]\end{aligned}$$

---

c The above minimization is overt the set $\mathcal{F}$ which is discrete and very large (the 28-permutohedron), with $|\mathcal{F}| = 28!$, meaning the the optimization is NP-hard and unfeasibly-high dimensional

---

Problem 2

a let $f^{(1)}$, $f^{(2)}$ correspond to two randomly drawn permuations from $\mathcal{F}$

$$\mathbb{P}\left(f_i^{(1)} \neq f_i^{(2)},\ f_j^{(1)} \neq f_j^{(2)},\ f_k = f_k\ \forall k \neq i,j\right)$$

Clearly there are $\binom{N}{2}$ pairs of i,j that can be interchanged starting from f^1, and $N!-1$ other permutations to select $f^{(2)}$ from. Therefore for two uniformly selected $f^{(1)}, f^{(2)} \in \mathcal{F}$:

$$\mathbb{P}\left(f_i^{(1)} \neq f_i^{(2)},\ f_j^{(1)} \neq f_j^{(2)},\ f_k = f_k\ \forall k \neq i,j\right) = \frac{\binom{N}{2}}{N!-1}$$

Let the set of $1-$pair exchanged sequences from a given $f \in \mathcal{F}$ be denoted $\mathcal{F}_1(f)$, with $|\mathcal{F}_1(f)| = \binom{N}{2} << |\mathcal{F}| = N!$. This also us to define a uniform proposal density $V(f'|f)$:

$$V_{f'|f}(f'|f) = \begin{cases} \binom{N}{2}^{-1} & f' \in \mathcal{F}_1(f) \\ 0 & \text{otherwise} \end{cases}$$

Note that this transition probability is symmetric, i.e. $V_{f|f}(f^{(k+1)}|f^{(k)}) = V_{f|f}(f^{(k)}|f^{(k+1)})$ an so the detailed balance equations will be satisfied for this choice.

b The MHMC algorithm is as follows:

1 Generate $f^{(0)}$ from $\mathcal{F}$ according to $p_f(f)$ (uniform)

2 for $k = 0, 1, \ldots, n$:

i generate f' from $\mathcal{F}_1$ according to $V(f'|f^{(k)})$ (uniform on $\mathcal{F}_1(f^{(k)})$

ii calculate $\alpha = \min\left[1, \frac{p_{f|\mathbf{y}}(f'|\mathbf{y})V(f^{(k)}|f')}{p_{f|\mathbf{y}}\left(f^{(k)}|\mathbf{y}\right)V(f'|f^{(k)})}\right] = \min\left[1, \frac{p_{f|\mathbf{y}}(f'|\mathbf{y})}{p_{f|\mathbf{y}}\left(f^{(k)}|\mathbf{y}\right)}\right]$.

iii generate a from a Bernouli distribution with parameter α, $a \sim \mathcal{B}(\alpha)$.

iv If $a = 1$, $f^{(k+1)} = f'$, else $f^{(k+1)} = f^{(k)}$

v repeat until $k = n$

Algorithm 1 My algorithm

```
3  1: procedure MyProcedure
   2:     stringlen ← length of string
   3:     i ← patlen
   4: top:
   5:     if i > stringlen then return false
   6:     end if
   7:     j ← patlen
   8: loop:
   9:     if string(i) = path(j) then
  10:         j ← j − 1.
  11:         i ← i − 1.
  12:         goto loop.
  13:         close;
  14:     end if
  15:     i ← i + max(delta1(string(i)), delta2(j)).
  16:     goto top.
  17: end procedure
```

Problem 3

The inference task was repeated for 50 random initializations and run for 5000 iterations each time.

a)

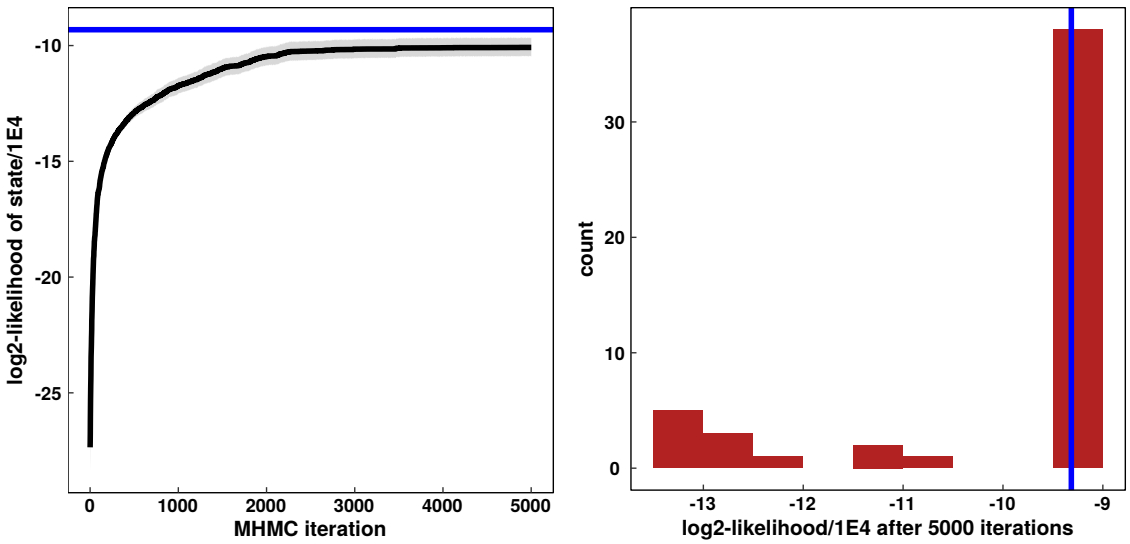

Figure 1.1. (L) Log-likelihood in bits as a function of MHMC iteration, showing average over 50 random initializations and 95% confidence interval based on non-parametric bootstrap with 1000 draws. The blue line shows the log-likelihood of the true cipher. (R) Histogram of log-likelihood after 5000 iterations showing that the correct minimum is obtained in 76% of random initializations.

b)

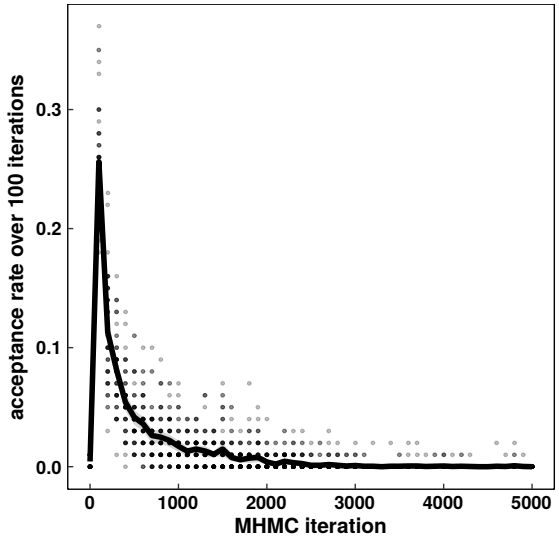

Figure 1.2. Average accept rate over 100 iterations as a function of MHMC iteration, showing average over 50 random initializations and 95% confidence interval based on nonparametric bootstrap. Individual runs shown as dots.

---

c)

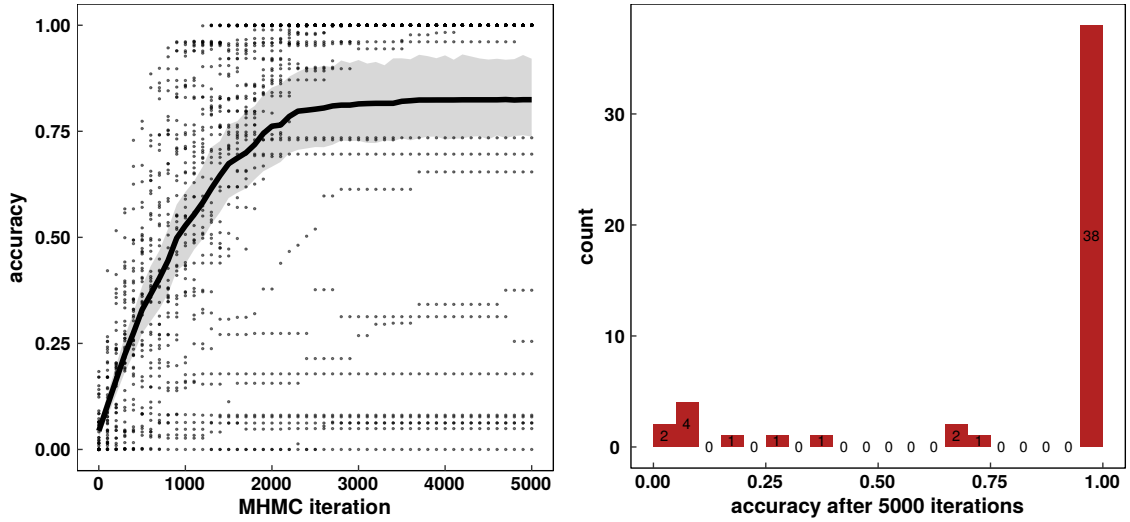

Figure 1.3. (L) Accuracy in fraction of correctly decoded characters as a function of MHMC iteration, showing average over 50 random initializations and 95% confidence interval based on non-parametric bootstrap with 1000 draws. Individual runs shown as dots. (R) Final accuracy after 5000 iterations showing bin counts for 50 random initializations. The correct answer is obtained in 76% of random initializations.

---

d) The inference task was repeated, varying the length of the input ciphertext from 500 to 27320 characters (full length). These trials were repeated using 10 random initializations each. Reducing sequence length makes the inference task more difficult, although the random initial conditions and low number of repeats (in particular for $L = 8000$) make the trend less obvious, as evidenced by the large confidence interval obtained. This is expected since the longer the text, the more sharply peaked the log-likelihood will be around ciphers that produce common letter orders. Additionally, the inference is based on aggregate 2-gram statistics and it is expected local deviations from these averages will be more extreme (i.e. weak law of large numbers).

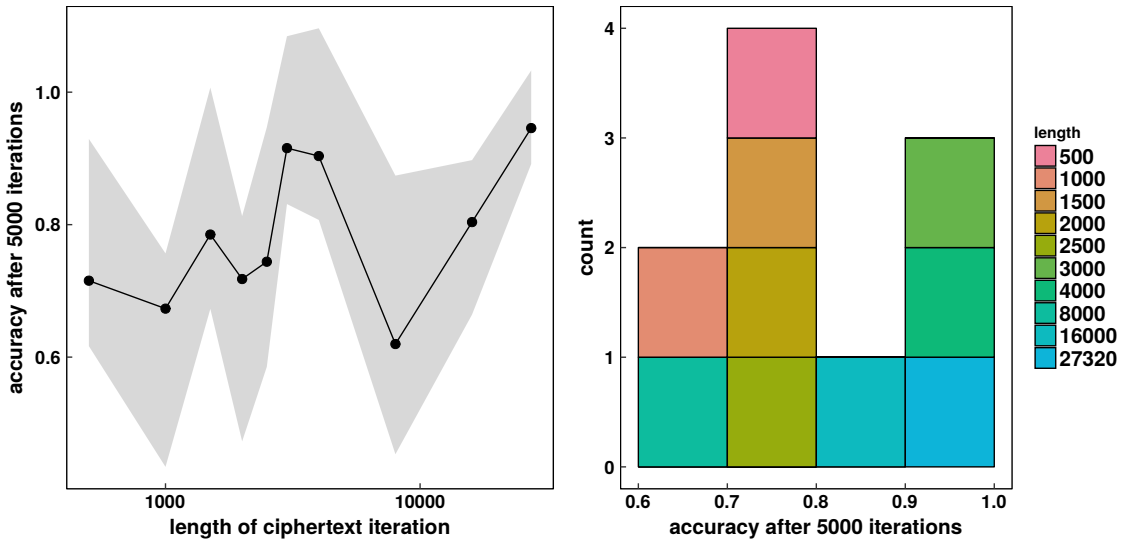

Figure 1.4. (L) Accuracy in fraction of after 5000 MHMC iterations as a function of ciphertext length, showing average over 10 random initializations and 95% confidence interval based on non-parametric bootstrap with 1000 draws. (R) Final accuracy after 5000 iterations grouped by ciphertext length, showing bin counts for 10 random initializations. While the $L = 8000$ case is atypical, there is a clear trend towards obtaining high accuracy for longer chains (cool colours concentrate to the right).

e) The (negative) log-likelihood in bits per character of the decoded plaintext, which is 27320 characters, increases with iteration and maximizes at a very similar value of 3.41 compared to the 2-gram entropy value of 3.32 suggested by Shannon[1] for a 27-symbol alphabet.

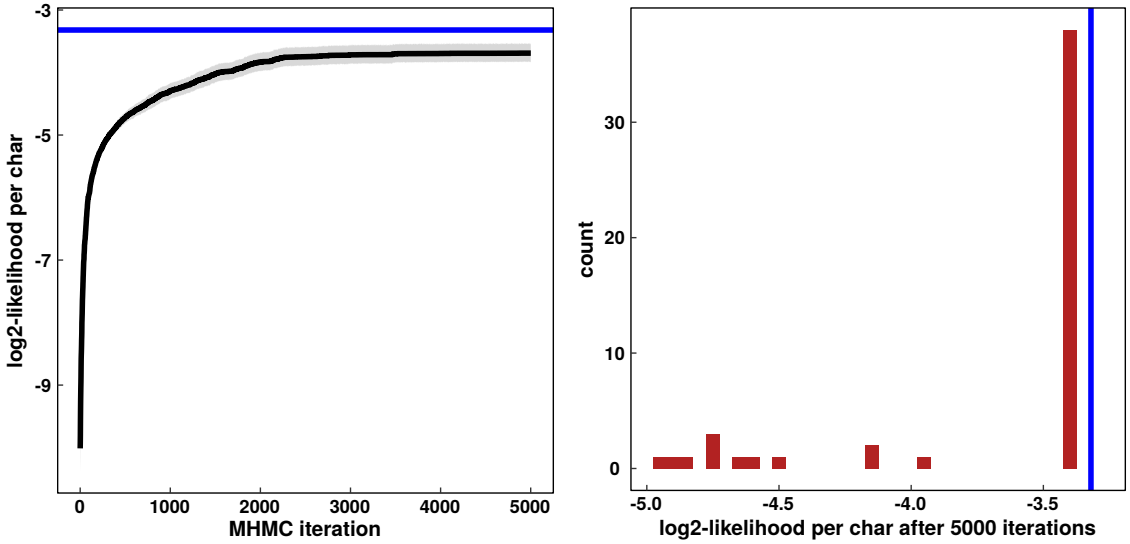

Figure 1.5. (L) Log-likelihood in bits per character as a function of MHMC iteration, showing average over 50 random initializations and 95% confidence interval based on non-parametric bootstrap with 1000 draws. The blue line shows the Shannon 2-gram entropy[1]. (R) Histogram of log-likelihood per character after 5000 iterations showing that most chains converge to to a value near the Shannon 2-gram entropy[1].

Bibliography

[1] C. E. Shannon. Prediction and entropy of printed english. *Bell Labs Technical Journal*, 30(1):50–64, 1951.