Homework 1 **Yen-Lin Chen**
ECE 5412 yc2253@cornell.edu
Fall 2019 Due: 09/31/19

Problem 17

The CDF is given as

$$F(x) = \begin{cases} \frac{1-e^{-2x}+2x}{3}, & \text{if } 0 < x \leq 1 \\ \frac{3-e^{-2x}}{3}, & \text{if } 1 < x < \infty \end{cases} \tag{1}$$

The PDF is the following:

$$f(x) = \begin{cases} \frac{2}{5}(e^{-2x}+2), & \text{if } 0 < x \leq 1 \\ \frac{2}{5}e^{-2x}, & \text{if } 1 < x < \infty \end{cases} \tag{2}$$

First we need to generate a random variable that spans the positive real axis. We apply the exponential distribution as a known distribution $q(x) = e^{-x}$ and use acceptance-rejection method to generate the desired distribution using the PDF in Eq. (2).

1. Generate $q(y)$ by using $y = -\log u$ where $u \sim U[0, 1]$.
2. Determine the constants c_1 and c_2 for first and second conditions.

$$c_1 = \max_{0<t\leq 1} \frac{f(t)}{q(t)} = \frac{4e}{5(e-1)}\sqrt{2} \tag{3}$$

$$c_2 = \max_{1<t<\infty} \frac{f(t)}{q(t)} = \frac{2}{5} \tag{4}$$

   We can choose $c = \max(c_1, c_2)$ for both conditions.

3. Generate $u_2 \sim U[0, 1]$ and set $x = y$ if the following holds

$$u_2 < \frac{f(y)}{cq(y)} \tag{5}$$

   otherwise go back to step one.

We generated $n = 1000000$ random samples and plotted the cumulative histogram for the samples as shown in Fig. 1.

Problem 20

We would like to generate random variables from CDF:

$$F(x) = \prod_{i=1}^{n} F_i(x) \tag{6}$$

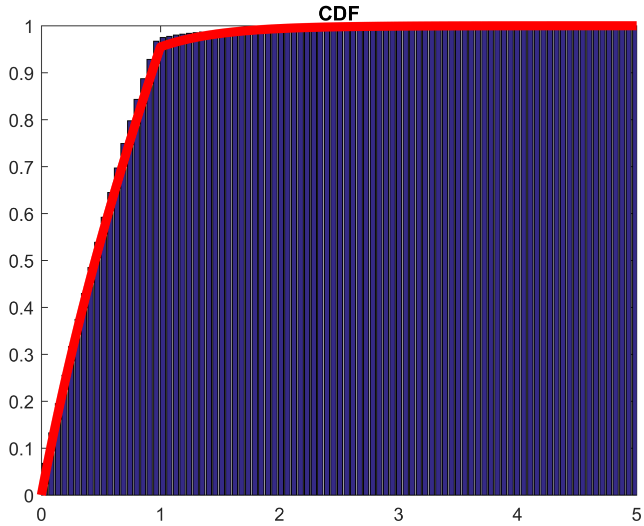

Figure 1: Problem 17: The cumulative histogram of $n = 100000$ samples and the real CDF as in Eq. (1).

where $F_i(x)$'s are known and easy to generate. From Eq. (6), we have the following

$$\begin{aligned} F(t) = P(x \le t) &= \prod_{i=1}^{n} F_i(x) \\ &= \prod_{i=1}^{n} P(x_i \le t) = P(\max_i x_i \le t) \end{aligned} \tag{7}$$

The CDF in Eq. (6) is the distribution of $\max_i x_i$ where $x_i \sim F_i(x_i)$. Therefore we can use the following steps to generate $F(x)$.

1. generate $u \sim U[0, 1]$
2. use u to generate x_i's based on the corresponding CDFs F_i's.
3. set $x = \max_i x_i$

Problem 26

The second-order transition probability $P(X_{k+1}|X_k, X_{k-1})$ where $P : S \bigotimes S^2 \to [0, 1]$. S is the state space of the discrete Markov Chain. And we have

$$\sum_{X_{k-1} \in S} \sum_{X_k \in S} P(X_{k+1}|X_k, X_{k-1}) = 1 \tag{8}$$

The transition probability $P(X_{k+1}|X_k, X_{k-1})$ can be re-written as

$$P(X_{k+1}|X_k, X_{k-1}) = P(X_{k+1}, X_k|X_k, X_{k-1}) \tag{9}$$

since it is conditioned on X_k. Now the transition probability can be viewed as a first order on the new expanded state space: S^2, i.e. $P : S^2 \bigotimes S^2 \to [0, 1]$. Moreover, we have

$$\begin{aligned}\sum_{\{X_k,X_{k-1}\}\in S^2} P(X_{k+1},X_k|X_k,X_{k-1}) &= \sum_{X_{k-1}\in S}\sum_{X_k\in S} P(X_{k+1},X_k|X_k,X_{k-1}) \\ &= \sum_{X_{k-1}\in S}\sum_{X_k\in S} P(X_{k+1}|X_k,X_{k-1}) \\ &= 1\end{aligned} \tag{10}$$

Therefore, a second-order Markov Chain can be expressed as a first-order Markov Chain on the expanded state space: S^2.

Problem 27

(a)

The implementation of inverse transform algorithm is the combination of the MATLAB function $\textit{inverse\_transform}(a, b)$ and $\textit{itmc}(n)$ where $P(state1) = a$ and $P(state1||state2) = b$ is the CDF. n is the number of generated samples, i.e. $n = 500$. The histogram of the generated samples is shown in the left plot of Fig. 2 with stationary distribution indicated by the red dots.

(b)

The implementation of the acceptance rejection algorithm is the combination of the MATLAB function $\textit{acceptance\_rejection}(a, b)$ and $armc(n)$ using the same notation as in (a). The histogram of the generated samples is shown in the middle plot of Fig. 2 with stationary distribution indicated by the red dots.

(c)

With the empirical sequence of the Markov Chain: $(1222231331\dots)$, we can estimate the transition matrix p. We used the MATLAB function *strfind* to look for a specific pattern in the sequence. The transition probability from $i \to j$ is then the number of occurrences of pattern (i, j) divided by number of occurrences of (i), i.e.

$$p_{ij} = \frac{\#(i,j)}{\#(i)} \tag{11}$$

The implementation is the MATLAB function *estimatemc(x)* where x is the sequence of the empirical samples.

The empirical transition matrix p is found to be the following and different sampling might result in slightly different p.

$$p = \begin{bmatrix} 0.1833 & 0.3833 & 0.4333 \\ 0.1209 & 0.8791 & 0.0000 \\ 0.0896 & 0.1045 & 0.8060 \end{bmatrix} \tag{12}$$

(d)

With the transition matrix p we can obtain the empirical stationary distribution as

$$\pi^*_\infty = \begin{bmatrix} 0.1200 & 0.6120 & 0.2680 \end{bmatrix}', \tag{13}$$

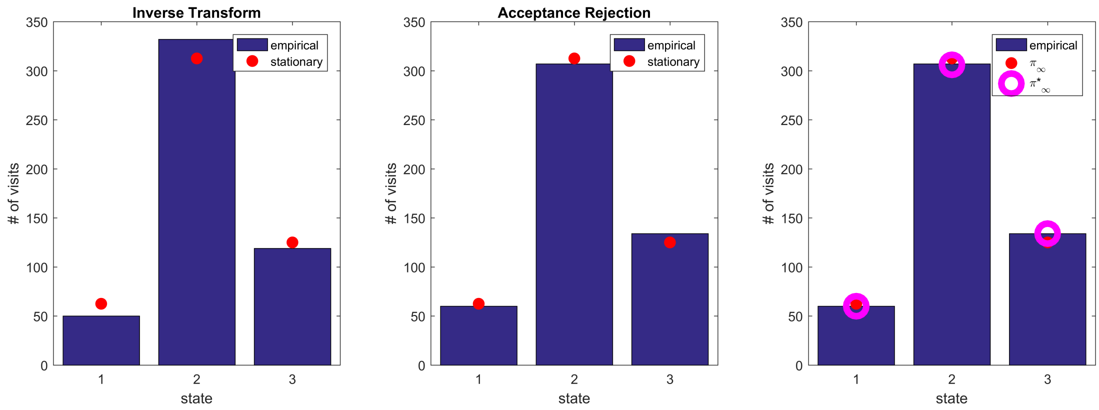

Figure 2: Problem 27: 500 samples of the Marhov Chain using inverse transform (left) and acceptance-rejection (middle). The red dots represent the stationary distribution π_∞ from the transition probability matrix A. The empirical estimator of a the transition matrix from the generated samples (right) where the magenta circles denotes the empirical stationary distribution π^*_∞.

while the real stationary distribution is

$$\pi_\infty = \begin{bmatrix}0.1250 & 0.6250 & 0.2500\end{bmatrix}', \tag{14}$$

Both π^*_∞ and π_∞ are shown in the right plot of Fig. 2 in magenta circles and red dots respectively. Good agreement between empirical and real stationary distributions can be seen.

Problem 28

Without lost of generality, assume the 2-state transition probability is of the following form.

$$P = \begin{bmatrix}1-a & a \\ b & 1-b\end{bmatrix} \tag{15}$$

where $0 \le a, b \le 1$. With the stationary dist ribution π_∞ satisfying

$$\pi_\infty = P'\pi_\infty = \frac{1}{a+b}\begin{bmatrix}b \\ a\end{bmatrix} \tag{16}$$

Therefore, π_∞ is the eigenvector of P' corresponding to eigenvalue 1. The other eigenvalue is $1-a-b$ corresponding to the eigenvector $[1, -1]'$. Given an arbitrary starting probability distribution

$$\pi_0 = \begin{bmatrix}c \\ 1-c\end{bmatrix} = \frac{1}{a+b}\begin{bmatrix}b \\ a\end{bmatrix} + \left(c - \frac{b}{a+b}\right)\begin{bmatrix}1 \\ -1\end{bmatrix} = \pi_\infty + \left(c - \frac{b}{a+b}\right)\begin{bmatrix}1 \\ -1\end{bmatrix} \tag{17}$$

where $0 \le c \le 1$. Therefore, $\pi_k = (P')^k\pi_0$

$$\begin{aligned}\pi_k &= (P')^k\pi_0 \\ &= \pi_\infty + (1-a-b)^k\left(c - \frac{b}{a+b}\right)\begin{bmatrix}1 \\ -1\end{bmatrix}\end{aligned} \tag{18}$$

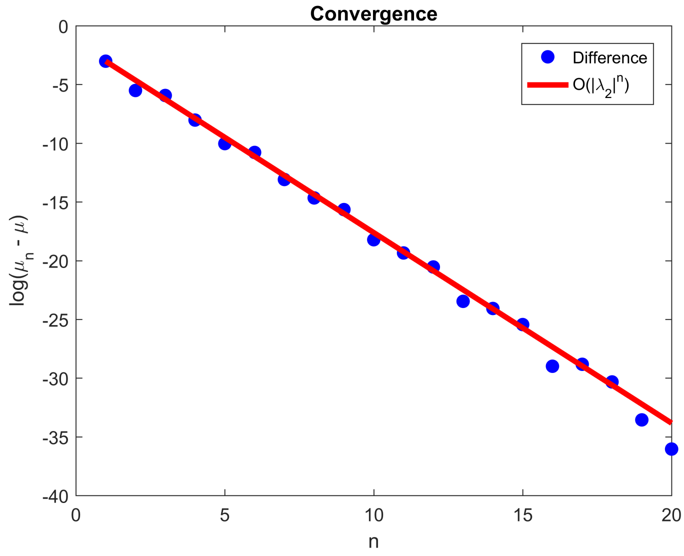

Figure 3: Problem 28: The convergence of d_n is geometrically fast in terms of $|\lambda_2|$.

Therefore, π_k converges to π_∞ geometrically fast in terms of the second largest eigenvalue $1-a-b$. $\forall a, b$ satisfying $0 < a, b < 1$, $0 \leq |1-a-b| < 1$. There are a few situations to consider when $|1-a-b| = 1$.

1. $a = b = 1$: the chain is periodic, π_∞ does not exist unless $c = \frac{1}{2}$.
2. $a = b = 0$: $P = I$ so $\pi_0 = \pi_\infty$.

For a Markov Chain with three states, we can randomly generate the transition probability P and do a numerical experiment. For a state space $S = \{1, 2, 3\}$ with state vector $s = [1, 2, 3]'$, we calculate the stationary average $\mu = s'\pi_\infty$. We initialize a random initial state π_0 and we have $\pi_n = (P')^n \pi_0$. We compute the difference between the average at time n and stationary average:

$$d_n = |s'(\pi_\infty - \pi_n)| \tag{19}$$

And we demonstrate that the difference d_n converges geometrically fast in terms of $|\lambda_2|$: $d_n \sim \mathcal{O}(|\lambda_2|^n)$ as shown in Fig. 3.

Problem 30

Assuming that $i \neq j$, we can remove state j from the state space of the Markov Chain because we want to make sure that for time $k = 1, 2, 3 \ldots n-1$, the chain does not visit state j. The new transition probability is now $R \in \mathbb{R}^{(n-1)\times(n-1)}$ with the j^{th} row and column removed. Moreover, we define the $P_{i\neq j,j} \in \mathbb{R}^{n-1}$ as the transition probability from state $i \neq j$ to state j. The starting state is $x_0 = i$, corresponding to a one-hot vector $\pi_0 \in \mathbb{R}^{n-1}$ with the i^{th} element equal to 1. Let's list some of the quantities to see the underlying rules.

$$P(x_1 = j | x_0 = i) = P_{ij} = P'_{i\neq j,j}\pi_0 = P'_{i\neq j,j} I \pi_0 \tag{20}$$

$$P(x_2 = j, x_1 \neq j | x_0 = i) = P'_{i\neq j,j} R' \pi_0 \tag{21}$$

$$P(x_3 = j, x_2 \neq j, x_1 \neq j | x_0 = i) = P'_{i \neq j, j}(R')^2 \pi_0 \tag{22}$$

Therefore, we have

$$P(x_n = j, x_{n-1} \neq j, \ldots, x_1 \neq j | x_0 = i) = P'_{i \neq j, j}(R')^{n-1} \pi_0 \tag{23}$$

The probability of ending up in state j given starting at state i is then

$$\begin{aligned} p_{i \to j} &= \sum_{n=1}^{\infty} P(x_n = j, x_{n-1} \neq j, \ldots, x_1 \neq j | x_0 = i) \\ &= \sum_{n=1}^{\infty} P'_{i \neq j, j}(R')^{n-1} \pi_0 \\ &= P'_{i \neq j, j} \left[\sum_{n=1}^{\infty} (R')^{n-1} \right] \pi_0 \\ &= P'_{i \neq j, j}(I - R')^{-1} \pi_0 \end{aligned} \tag{24}$$

where I is the identity matrix. The time it takes for the first visit of state j given starting at state i, T_{ij} is

$$\begin{aligned} T_{ij} &= \sum_{n=1}^{\infty} n P(x_n = j, x_{n-1} \neq j, \ldots, x_1 \neq j | x_0 = i) \\ &= \sum_{n=1}^{\infty} P'_{i \neq j, j} n (R')^{n-1} \pi_0 \\ &= P'_{i \neq j, j} \left[\sum_{n=1}^{\infty} n (R')^{n-1} \right] \pi_0 \\ &= P'_{i \neq j, j} \left[(I - R')^{-1} \right]^2 \pi_0 \end{aligned} \tag{25}$$

Problem 33

We start by randomly generate the transition probability matrix A and π_k for a regular Markov Chain with $N = 1000$ states. We used the acceptance-rejection algorithm to generate integer i_m from distribution π_k. We estimated the inner product $\pi_{k+1} = A'\pi$ as

$$\pi_{k+1} = A'\pi_k \approx \pi^*_{k+1} = \frac{1}{M} \sum_{m=1}^{M} A'(i_m, :) \tag{26}$$

We are interested in the estimation error between π_{k+1} and π^*_{k+1} so we use the L1-norm of their difference as a metric for the accuracy of the estimation: $||\pi^*_{k+1} - \pi_{k+1}||_1$. We use 100 different M values from 100 to 10000 and plot the error and the estimation time of Eq. (26) as shown on the left and right of Fig. 4 respectively. The error stabilizes for $M \leq 5000$ in our experiment. As expected, the estimation time is of linear complexity of $\mathcal{M}$.

For the fastest estimation, $M = 100$, the estimation time is 1.5 ms while the computation time for $A'\pi_k$ is 0.693 ms. Moreover, for $M = 5000$ where the error converges, the estimation time is 27.5 ms. All of the estimation time is greater than the computation time for direct product. This might be due to the fact that $N = 1000$ here is not large enough to see the speeding up of the estimation. MATLAB uses their version of *BLAS* which starts to see the significance of $\mathcal{O}(N^2)$ for matrix-vector product after $N > 1e5$ according to my personal experience.

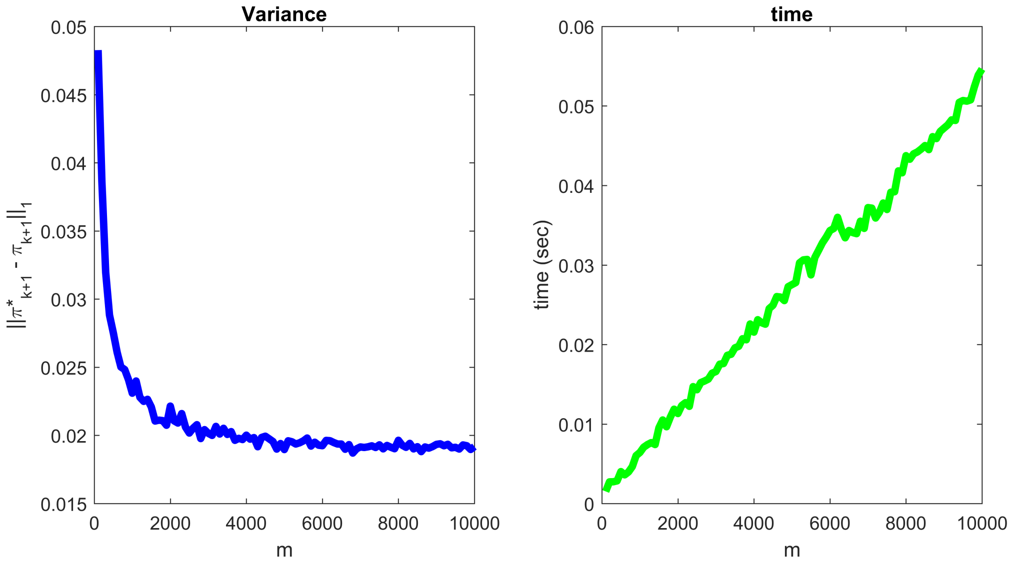

Figure 4: Problem 33: The L1-norm error between estimation and real π_{k+1} (left) and the computation time of Eq. (26) using different M values (right).

Problem 34

In this problem, we use the Standard & Poor 500 (S&P500) Index from 2014/9/24 to 2019/9/24 to compute the return for different time periods. The stock market data was obtained from *Yahoo Finance* with data frequency of one day and is shown in the first plot of Fig. 5. The log return at time k is

$$r_k = \log\left(\frac{p_k}{p_{k-1}}\right) \tag{27}$$

where p_k and p_{k-1} are the stock price at time k and $k-1$ respectively. We tested 8 different periods: 1 day, 3 days, 1 week, 2 weeks, 1 month, 1 season, 1 year and 2 years. The returns are plotted in percentage and fit by a normal distribution as shown in the panels of Fig. 5. We performed the one-sample Kolmogorov-Smirnov test by using the *kstest* function in MATLAB. The p-values are reported in the titles. In general, the *kstest* suggests that non of the return distributions is normal.

An interesting point is that the mean of the short-period return is zero for periods less than a month. Long-term returns have positive means which might due to the bouncing-back from the slump in February 2017.

Attachments

Problem 17: *CDF.m, PDF.m, generate.m, q17.m*
Problem 27: *acceptance_rejection.m, armc.m, inverse_transform.m, itmc.m, estimatemc.m, q27.m*
Problem 28: *q28.m*
Problem 33: *acc_rej.m, q33.m*
Problem 34: *sp5ydaily.csv, getreturns.m, q34.m*

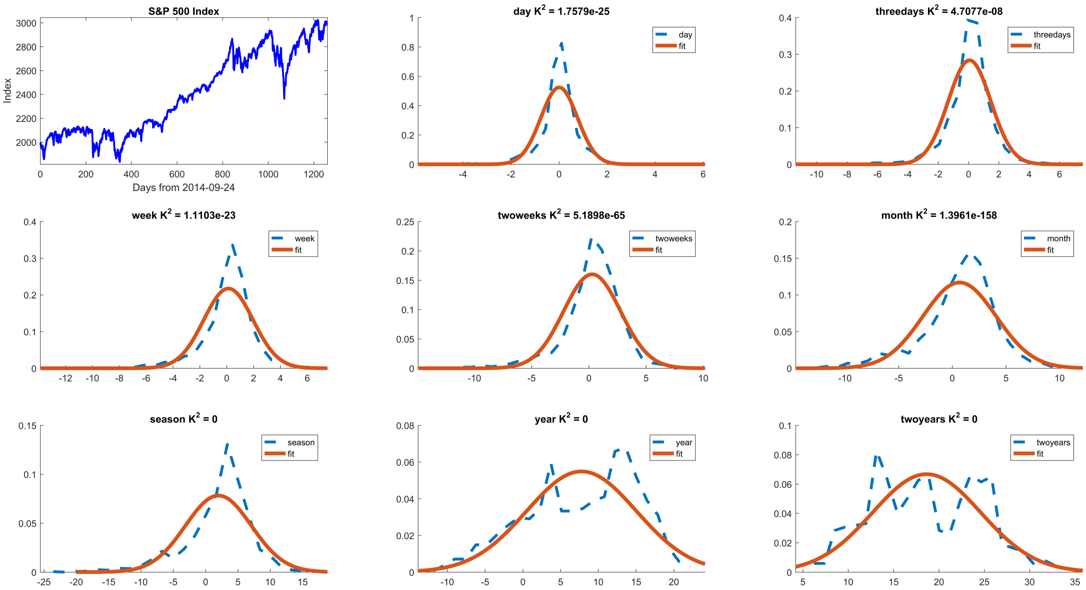

Figure 5: Problem 34: The S&P 500 Index and the return distributions from 8 different time periods where the x-axis is the percentage and y-axis is the probability density.