Assignment 3, ITDS

Amir Parsian, Brigitta Selb, Lingkai Zhu, Weilin Zhang

November 2021

1 Introduction

Q1: Prove lemma 6.7

Consider a congruential generator D on $\mathcal{M} = \{0, 1, ..., M-1\}$ with period M, then for any starting point μ_0 in $\mathcal{M}$, define $u_i = D(u_{i-1})$ then the sequence $v_i = u_i \bmod K$ for $1 \leq K \leq M$ is pseudorandom on $\{0, 1, ..., K-1\}$ if M is a multiple of K.

Solution

To prove this lemma we need to first prove lemma 6.6.
The length of $\mathcal{M}$ is M and the period of the congruential generator is M. This means that all members of $\mathcal{M}$ are included in each period and there is no repetition (otherwise the period becomes less than M). Since the numbers in each period is unique we will have $\frac{N(a)}{n} = \frac{1}{M}$ for one period (note that for one period $N(a) = 1$ and $n = M$). In long term the sequence repeat itself by a period of M and therefore $\frac{N(a)}{n}$ approaches $\frac{1}{M}$. From definition of 6.3 this sequence is pseudorandom on $\mathcal{M}$.

Now, for lemma 6.7, we have:

$M = i \cdot K$ where $i \in \mathbb{N}$.

We can partition $\mathcal{M}$ into i subsets ($\mathcal{M}_1$, $\mathcal{M}_2$, ... , $\mathcal{M}_i$). Each of this subsets have a length of K. We partition $\mathcal{M}$ in a way that $\mathcal{M}_1 = \{0, 1, ..., K-1\}$, $\mathcal{M}_2 = \{K, K+1, ..., 2K-1\}$, ...

Note that:

1. we obtain elements of $\mathcal{M}_1$ by subtracting multiples of K from members of $\mathcal{M}_2$, $\mathcal{M}_3$, ...

2. m mod $K = (m + iK)$ mod K

From 1 and 2, we can say that each member of $\mathcal{M}_1$ is repeated i times when m mod K is calculated for $m \in \mathcal{M}$. This means:
$\frac{N(a)}{n} = i\frac{1}{M}$ which gives $\frac{N(a)}{n} = \frac{1}{K}$

Note that the sequence of $v_i = u_i$ mod K has a period of M (since M is multiple of K). Therefore, $\frac{N(a)}{n}$ approaches $1/K$ which by definition is pseudorandom on $\mathcal{M}_1$.

Q2: Theorem 6.13 (Box-Muller)

Suppose that U_1 and U_2 are independent random variables and $U_1, U_2 \sim$ Uniform([0,1]), then

$$Z_0 = \sqrt{-2\ln U_1}\cos(2\pi U_2)$$
$$Z_1 = \sqrt{-2\ln U_1}\sin(2\pi U_2)$$

are independent random variables, and $Z_0, Z_1 \sim$ Normal(0,1).

Solution

Since $|Z| = \sqrt{Z_0^2 + Z_1^2}$, Let

$$\begin{aligned} Y = |Z|^2 &= Z_0^2 + Z_1^2 \\ &= (\sqrt{-2\ln U_1}\cos(2\pi U_2))^2 + (\sqrt{-2\ln U_1}\sin(2\pi U_2))^2 \\ &= -2\ln U_1 \qquad (1) \end{aligned}$$

According to (1), we could get $U_1 = e^{-(Z_0^2+Z_1^2)/2}$.

Since Z = $(Z_0$, $Z_1)$, Let

$$\begin{aligned} W = \frac{Z}{|Z|} &= \frac{(Z_0, Z_1)}{\sqrt{Z_0^2 + Z_1^2}} \\ &= \frac{(\sqrt{-2\ln U_1}\cos(2\pi U_2), \sqrt{-2\ln U_1}\sin(2\pi U_2))}{\sqrt{-2\ln U_1}} \\ &= (\cos(2\pi U_2), \sin(2\pi U_2)) \qquad (2) \end{aligned}$$

According to (1) and (2), Y and W are independent since U_1, U_2 are independent.
Suppose (X,Y) is uniformly distributed over the unit circle, then

$$S_{XY} = \{(x,y) : x^2 + y^2 \leq 1\}.$$

Since $U_2 \sim$ Uniform([0,1]), we could have $2\pi U_2 \sim$ Uniform([0, 2π]), in this case,

$$\cos(2\pi U_2) \in [-1,1], \sin(2\pi U_2) \in [-1,1]$$

and because

$$\cos^2(2\pi U_2) + \sin^2(2\pi U_2) = 1$$

then with (2) we could say, $W \subseteq S_{XY}$, so W is uniform on the unit circle.
To prove Z_0, Z_1 are independent random variables, and $Z_0, Z_1 \sim$ Normal(0,1), we need to prove they are Gaussian and their covariance is zero.
First, to prove Z_0, Z_1 are Gaussian, from

$$\frac{Z_1}{Z_0} = \frac{\sqrt{-2\ln U_1}\sin(2\pi U_2)}{\sqrt{-2\ln U_1}\cos(2\pi U_2)} = \tan(2\pi U_2),$$

we could have

$$U_2 = \frac{1}{2\pi}\arctan(\frac{Z_1}{Z_0})$$

according to the Jacobian of the polar transformation

$$\frac{\partial(U_1,U_2)}{\partial(Z_0,Z_1)} = \begin{vmatrix} \frac{\partial U_1}{\partial Z_0} & \frac{\partial U_1}{\partial Z_1} \\ \frac{\partial U_2}{\partial Z_0} & \frac{\partial U_2}{\partial Z_1} \end{vmatrix} = \frac{\partial U_1}{\partial Z_0}\frac{\partial U_2}{\partial Z_1} - \frac{\partial U_2}{\partial Z_0}\frac{\partial U_1}{\partial Z_1} = -\frac{1}{2\pi}e^{-(Z_0^2+Z_1^2)/2}$$

since U_1 and U_2 are independent random variables and $U_1, U_2 \sim$ Uniform([0,1]),the probability density function $f(U_1,U_2) = f(U_1)f(U_2) = 1$
by the formula for the random variable transformation, we have

$$f(Z_0,Z_1) = f(U_1,U_2)\left|\frac{\partial(U_1,U_2)}{\partial(Z_0,Z_1)}\right| = \frac{1}{2\pi}e^{-(Z_0^2+Z_1^2)/2}$$

According to the definition of multivariate normal distribution, (Z_0, Z_1) is two-dimensional normal distribution. Hence $Z_0, Z_1 \sim$ Normal(0,1).
Second, to prove the covariance of Z_0, Z_1 is zero,

$$Cov(Z_0,Z_1) = E[(Z_0 - E(Z_0)(Z_1 - E(Z_1)] = E[Z_0Z_1] - E[Z_0]E[Z_1] \qquad (3)$$

since Y and W are independent,

$$\begin{aligned} E[Z_0Z_1] &= E[\sqrt{-2\ln U_1}\cos(2\pi U_2)\cdot\sqrt{-2\ln U_1}\sin(2\pi U_2)] \\ &= E[(-2\ln U_1)\cdot\cos(2\pi U_2)\cdot\sin(2\pi U_2)] \\ &= E[-2\ln U_1]E[\cos(2\pi U_2)\sin(2\pi U_2)] \end{aligned}$$

$$\begin{aligned} E[Z_0]E[Z_1] &= E[\sqrt{-2\ln U_1}\cos(2\pi U_2)]E[\sqrt{-2\ln U_1}\sin(2\pi U_2)] \\ &= E[\sqrt{-2\ln U_1}]^2 E[\cos(2\pi U_2)]E[\sin(2\pi U_2)] \end{aligned}$$

With the definition, if X is a continuous random variable

$$E[g(X)] = \int g(x) f_X(x) dx$$

Because $4\pi U_2 \sim$ Uniform$(0, 4\pi)$, the probability density function $f(4\pi U_2) = \frac{1}{4\pi}$,

$$\begin{aligned} E[\cos(2\pi U_2)\sin(2\pi U_2)] &= \frac{1}{2} E[\sin(4\pi U_2)] \\ &= \frac{1}{2} \int_0^{4\pi} \frac{1}{4\pi} sin(4\pi U_2) d(4\pi U_2) \\ &= 0 \end{aligned}$$

then $E[Z_0 Z_1] = 0$.
Because $2\pi U_2 \sim$ Uniform$(0, 2\pi)$, the probability density function $f(4\pi U_2) = \frac{1}{2\pi}$,

$$E[\cos(2\pi U_2)] = \int_0^{2\pi} \frac{1}{2\pi} sin(2\pi U_2) d(2\pi U_2) = 0$$

then $E[Z_0]E[Z_1] = 0$.
So in (3), $Cov(Z_0, Z_1) = E[Z_0 Z_1] - E[Z_0]E[Z_1] = 0$, which means that Z_0 and Z_1 are uncorrelated.
Now, we have $Z_0, Z_1 \sim$ Normal(0,1) and their covariance is zero, then we could get Z_0, Z_1 are independent random variables.

Q3: Prove Lemma 7.11

For a finite inhomogeneous Markov chain $(X_t)_{t \in \mathbb{Z}_+}$ with state space $\mathbb{X} = \{s_1, s_2, ..., s_k\}$, initial distribution

$$\mu_0 := (\mu_0(s_1), \mu_0(s_2), ..., \mu_0(s_k)),$$

where $\mu_0(s_i) = \mathbb{P}(X_0 = s_i)$, and transition matrices

$$(P_1, P_2, ...), P_t := (P_t(s_i, s_j))_{(s_i, s_j) \in \mathbb{X} \times \mathbb{X}}, t \in \{1, 2, ...\}$$

we have for any $t \in \mathbb{Z}_+$ that the distribution at time t given by:

$$\mu_t := (\mu_t(s_1), \mu_t(s_2), ..., \mu_t(s_k)),$$

where $\mu_t(s_i) = \mathbb{P}(X_t = s_i)$, satisfies:

$$\mu_t = \mu_0 P_1 P_2 \cdot \cdot \cdot P_t$$

Solution

We know that $\mu_t := (\mu_t(s_1), \mu_t(s_2), ..., \mu_t(s_k))$ is a row vector where

$$\mu_t(s_i) = \mathbb{P}(X_t = s_i)$$

The simulation steps:

Step1: Draw $X_0 \sim \mu_0$. Thus, $\mathbb{P}(X_0 = s_i) = \mu_0(s_i)$.

Step2: Denote the outcome of step 1 by s_i. Draw $X_1 \sim (P_1, P_2, ...)$, which can also be written as $\mathbb{P}(X_1 = s_j | X_0 = s_i) = p_{s_i s_j}$.

Step3: Suppose the outcome of step 2 is S_j. Draw $X_2 \sim (P_1, P_2, ...)$, which can also be written as $\mathbb{P}(X_2 = s_k | X_1 = s_j) = p_{s_j s_k}$.

And so on.

A consequence of the above simulation is the following:

$$\begin{aligned}
\mu_t(s_k) &= \mathbb{P}(X_k = x_k) \\
&= \sum_{s_{k-1}} \mathbb{P}(X_k = s_k | X_{k-1} = s_{k-1}) \mathbb{P}(X_{k-1} = s_{k-1}) \\
&= \sum_{s_{k-1}} p_{s_k s_{k-1}} \mathbb{P}(X_{k-1} = s_{k-1}) \\
&= \sum_{s_{k-1} s_{k-2}} p_{s_k s_{k-1}} \mathbb{P}(X_{k-1} = s_{k-1} | X_{k-2} = s_{k-2}) \mathbb{P}(X_{k-2} = s_{k-2}) \\
&= \sum_{s_{k-1} s_{k-2}} p_{s_k s_{k-1}} p_{s_{k-2} s_{k-1}} \mathbb{P}(X_{k-2} = s_{k-2})
\end{aligned}$$

Apply until we reach X_0, we have

$$\begin{aligned}
\mathbb{P}(X_k = x_k) &= \sum_{s_{k-1} s_{k-2} \ldots s_1} p_{s_k s_{k-1}} p_{s_{k-2} s_{k-1}} \ldots p_{s_2 s_1} \mathbb{P}(X_0 = s_1) \\
&= \mu_0 P_1 P_2 \cdots P_t
\end{aligned}$$

Proof of Theorem 7.14

Theorem 7.14. Let $W_1, ..., \overset{IID}{\sim} F$ such that (ρ, W_t) is a RMR for a transition matrix P_t, for all $t \in \mathbb{N}$. Then if $X_0 \sim \mu_0$,

$$X_t := \rho_t(X_{t-1}, W_t), t \in \mathbb{N}$$

is a Markov chain with initial distribution μ_0 and transition matrix P_t at time t.

Solution

The simulation steps:

Step1: Draw $X_0 \sim \mu_0$. Thus, $\mathbb{P}(X_0 = x_0) = \mu_0(x_0)$.

Step2: Denote the outcome of step 1 by x_0. Draw $X_1 \sim \mathbf{P}$, which can also be written as $\mathbb{P}(X_1 = x_1|X_0 = x_0) = p_{x_0x_1} = \mathbf{P}(x_0, x_1) = \mathbb{P}(\{\rho(x_0, W_1) = x_1\}) = \mathbb{P}(X_1 = x_1)$.

Step3: Suppose the outcome of step 2 is x_1. Draw $X_2 \sim \mathbf{P}$, which can also be written as $\mathbb{P}(X_2 = x_2|X_1 = x_1) = p_{x_1x_2} = \mathbf{P}(x_1, x_2) = \mathbb{P}(\{(\rho(x_1, W_2) = x_2\}) = \mathbb{P}(X_2 = x_2)$.

....

Step t: Suppose the outcome of step $t-1$ is x_{t-1}. Draw $X_t \sim \mathbf{P}$, which can also be written as $\mathbb{P}(X_t = x_t|X_{t-1} = x_{t-1}) = p_{x_{t-1}x_t} = \mathbf{P}(x_{t-1}, x_t) = \mathbb{P}(\{(\rho(x_{t-1}, W_t) = x_t\}) = \mathbb{P}(X_t = x_t)$

$$\begin{aligned}
\mathbb{P}(\{(\rho(x_{t-1}, W_t) = x_t\}) &= \mathbb{P}(X_t = x_t) \\
&= \sum_{x_{t-1}} \mathbb{P}(X_t = x_t|X_{t-1} = x_{t-1})\mathbb{P}(X_{t-1} = x_{t-1}) \\
&= \sum_{x_{t-1}} p_{x_{t-1}x_t}\mathbb{P}(X_{t-1} = x_{t-1}) \\
&= \sum_{x_{t-1}x_{t-2}} p_{x_{t-1}x_t}\mathbb{P}(X_{t-1} = x_{t-1}|X_{t-2} = x_{t-2})\mathbb{P}(X_{t-2} = x_{t-2}) \\
&= \sum_{x_{t-1}x_{t-2}} p_{x_{t-1}x_t}p_{x_{t-2}x_{t-1}}\mathbb{P}(X_{t-2} = x_{t-2}) \\
&= \sum_{x_{t-1}x_{t-2}...x_0} p_{x_{t-1}x_t}p_{x_{t-2}x_{t-1}}...p_{x_1x_0}\mathbb{P}(X_0 = x_0) \\
&= \mu_0 P^t = \mu_0 P_t
\end{aligned}$$

Q5: Prove Proposition 7.23

(A reversible π is a stationary π). Let $(X_t)_{t\in\mathbb{Z}_+}$ be a Markov chain on $\mathbb{X} = s_1, s_2, ..., s_k$ with transition matrix P. If π is a reversible distribution for $(X_t)_{t\in\mathbb{Z}_+}$ then π is a stationary distribution for $(X_t)_{t\in\mathbb{Z}_+}$.

Solution:

Proof. For a reversible π to be stationary it must satisfy the two conditions given in **Definition 7.20**:

1. a probability distribution: $\pi(x) \geq$ for each $x \in \mathbb{X}$ and $\sum_{x\in\mathbb{X}} \pi(x) = 1$, and
2. a fixed point: $\pi P = \pi$ i.e.,$\sum_{x\in\mathbb{X}} \pi(x)P(x,y) = \pi(y)$ for each $y \in \mathbb{X}$

The first condition is already satisfied as the definition of a reversibility, **Definition 7.22**, states that π is a probability distribution therefore, $\pi(x) \geq$ for each $x \in \mathbb{X}$ and $\sum_{x\in\mathbb{X}} \pi(x) = 1$

For the second condition we know that for a reversible distribution the following is defined

$$\begin{aligned}
\pi(x)P(x,y) &= \pi(y)P(y,x) \\
\sum_{x\epsilon\mathbb{X}} \pi(x)P(x,y) &= \sum_{x\in\mathbb{X}} \pi(y)P(y,x) \\
&= \pi(y) \sum_{x\in\mathbb{X}} P(y,x) \\
&= \pi(y)
\end{aligned}$$

Thus satisfying the second condition, $\pi P = \pi$.

Therefore, π is reversible and π is stationary. □

Q6: Prove Prove Proposition 7.25 by directly showing that π is reversible

The random walk on a connected undirected graph $\mathbb{G} = (\mathbb{V}; \mathbb{E})$, with vertex set $\mathbb{V} := \{v_1, v_2, ..., v_k\}$ and degree sum $d = \sum_{i=1}^{k} deg(v_i)$ is a reversible Markov chain with the reversible distribution π given by:

$$\pi = (\frac{deg(v_1)}{d}, \frac{deg(v_2)}{d}, ..., \frac{deg(v_k)}{d})$$

Solution:

From definition 7.22, we need to show that:

$$\pi(v_i)P(v_i, v_j) = \pi(v_j)P(v_j, v_i)$$

From **Model 6** we have:

$P(v_i, v_j) = \frac{1}{deg(v_i)}$ for $(v_i, v_j) \in \mathbb{E}$. Therefore:

$$\pi(v_i)P(v_i, v_j) = \frac{1}{deg(v_i)} \cdot \frac{deg(v_i)}{d} \Rightarrow \pi(v_i)P(v_i, v_j) = \frac{1}{d}$$

$$\pi(v_j)P(v_j, v_i) = \frac{1}{deg(v_j)} \cdot \frac{deg(v_j)}{d} \Rightarrow \pi(v_j)P(v_j, v_i) = \frac{1}{d}$$

which is independent from i and j and therefore, $\pi(v_i)P(v_i, v_j) = \pi(v_j)P(v_j, v_i)$.

Q7:

Part1: In the above we are mentioning that R needs to be nice enough, Why is that? Does 0-1 loss work? Why?

We define $\hat{\phi}$ the empirical risk minimizer on the training dataset as

$$\hat{R_n}(\hat{\phi}) = min_{\phi \in \mathbb{M}} \hat{R}_n(\phi)$$

which comes from definition of the empirical risk:

$$\hat{R}_n = \hat{R}_n(Z; g) = \frac{1}{n} \sum_{i}^{n} L(Y_i, g(X_i))$$

This means that any plug in estimator of a linear functional is actually the sum of independent RVs. If $L(Y_i, g(X_i))$ is nice enough, such as sub-Gaussian or sub-exponential, then we can utilize the concentration inequalities described in notebook section 3.1.

We know that 0-1 loss is defined as

$$\mathbb{1}_{y \neq g(x)} = \begin{cases} 1 & if\ y \neq g(x) \\ 0 & if\ y = g(x) \end{cases}$$

which means that the loss is 1 if y is the wrong value and 0 if it is correct. Then the pattern recognition problem is equivalent to the problem of minimizing the functional

$$R(g) = \int L(y, g(x)) dF(x, y) = \mathbb{E}[L(Y, g(X))] = \frac{1}{n} \sum_{i}^{n} L(Y_i, g(X_i)) = \mathbb{P}(\{Y \neq g(X)\})$$

The above equation shows that 0-1 loss function works.

Part2: Furthermore, we used the tower property to derive 8.4 from 8.3, how does this work?

We know that given training set T_n, the probability of empirical risk over the test dataset deviating from ground truth is bounded by:

$$\mathbb{P}(|\hat{R}_m(\hat{\phi}) - R(\hat{\phi})| > \epsilon | T_n) < 2e^{-C\epsilon^2 n}$$

We know from Theorem 2.50 that the tower property can be written as:

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$$

Suppose we are using 0-1 loss function, then Then 8.3 can be re-written as

$$\mathbb{P}(|\hat{R}_m(\hat{\phi}) - R(\hat{\phi})| > \epsilon | T_n) = \mathbb{E}[\mathbb{1}_{\{|\hat{R}_m(\hat{\phi}) - R(\hat{\phi})| > \epsilon\}} | T_n]$$

Then we can apply tower property to the above equation and get

$$\mathbb{E}[\mathbb{E}[\mathbb{1}_{\{|\hat{R}_m(\hat{\phi}) - R(\hat{\phi})| > \epsilon\}} | T_n]] = \mathbb{E}[\mathbb{1}_{\{|\hat{R}_m(\hat{\phi}) - R(\hat{\phi})| > \epsilon\}}]$$

Because the the expectation of the RHS is itself, we get

$$\mathbb{E}[\mathbb{1}_{\{|\hat{R}_m(\hat{\phi}) - R(\hat{\phi})| > \epsilon\}}] < 2e^{-C\epsilon^2 n}$$

which is

$$\mathbb{P}(|\hat{R}_m(\hat{\phi}) - R(\hat{\phi})| > \epsilon) < 2e^{-C\epsilon^2 n}$$