Uppsala University

Introduction to Data Science

1MS041

Group Assignment 3

Authors:
Jonas Frankemölle
William Hooper
Yijie Zhou

December 13, 2021

1 Solve Exercise 29: Prove Lemma 6.7

Lemma. 6.7 *Consider a congruential generator $\mathcal{D}$ on $\mathcal{M} = \{0, 1, ..., M-1\}$ with period M, then for any starting point $u_0 \in \mathcal{M}$, define $u_i = D(u_{i-1})$. Then the sequence $v_i = u_i$ mod K for $1 \leq K \leq M$ is pseudoramdom on $\{0, 1, ..., K-1\}$ if M is a multiple of K.*

Proof. By Lemma 6.6, the sequence $u_i = D(u_{i-1})$ is pseudorandom on $\mathcal{M}$. Thus we know

$$\frac{N_n^{\mathcal{M}}(a)}{n} \to \frac{1}{M}$$

for all $a \in \mathcal{M}$, where $N_n^{\mathcal{M}}$ counts the appearances of each a. The superscript $\mathcal{M}$ is added to signify that $a \in \mathcal{M}$ to prevent future ambiguity.

Next, we know that M is a multiple of K, $1 \leq K \leq M$, so $M = bK$ for some $b \in \mathbb{N}$. Thus for the sequence $v_i = u_i \bmod K$,

$$N_n^{\mathcal{K}}(v_i) = \sum_{c=1}^{b} N_n^{\mathcal{M}}(u_i) = bN_n^{\mathcal{M}}(u_i).$$

This is because for each v_i, there exists $u_i = v_i + cK$ for all $c \in \{1, ..., b\}$. To think about this another way, K is splitting M into b partitions of size $|K|$ where each partition contains a value u_i corresponding to each v_i. With these two facts, we can now derive that

$$\frac{N_n^{\mathcal{K}}(a)}{n} = \frac{bN_n^{\mathcal{M}}(a)}{n} \to \frac{b}{M} = \frac{b}{bK} = \frac{1}{K}.$$

Thus the sequence v_i is pseudorandom on $\mathcal{K} = \{0, 1, ..., K-1\}$ as desired. □

2 Solve Exercise 30

Theorem. 6.13 *(Box-Muller). Suppose that $U_1, U_2 \sim \text{Uniform}([0, 1])$, 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 $Z_0, Z_1 \sim \text{Normal}(0, 1)$.

Proof. Consider bivariate normal RV. Z, then the distribution of $Y = |Z|^2$ is χ^2 distributed with 2 degrees of freedom,

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

Then we can calculate U_1 from (1),

$$Z_0^2 + Z_1^2 = -2\ln(U_1)$$
$$U_1 = e^{-\frac{(Z_0^2+Z_1^2)}{2}}$$

Furthermore $W = Z/|Z|$,

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

U_1, U_2 are independent random variables, so based on (1) and (2) W, Y are independent.
Assume (x, y) is uniformly distributed on the unit circle,

$$C = \{(x, y) \in R^2 : x^2 + y^2 = 1\}$$

Since $U_2 \sim$ Uniform([0,1]), $2\pi U_2 \sim$ Uniform([0, 2π]), and $\cos(2\pi U_2) \in [-1, 1], \sin(2\pi U_2) \in [-1, 1], \cos^2(2\pi U_2) + \sin^2(2\pi U_2) = 1$.
Hence, W generated using $(\cos(2\pi U_2)\sin(2\pi U_2))$ is uniform on the unit circle.

Lastly, we need to show that Z_0 and Z_1 are independent. As proposed by the exercise, we can do this by showing that the covariance between the two variables is 0.

$$\mathrm{cov}(Z_0, Z_1) = \mathbb{E}[(Z_0 - \mathbb{E}[Z_0])(Z_1 - \mathbb{E}[Z_1])] = \mathbb{E}[Z_0 Z_1] - \mathbb{E}[Z_0]\mathbb{E}[Z_1].$$

We know $\mathbb{E}[Z_0]\mathbb{E}[Z_1] = 0$ since Z_0 and Z_1 are standard normal, having 0 expectation. Thus, we just have to show that $\mathbb{E}[Z_0 Z_1] = 0$. Expanding, we need to find:

$$\mathbb{E}[-2\ln(U_1)\cos(2\pi U_2)\sin(2\pi U_2)].$$

First, by trig identities we can reduce this to

$$\mathbb{E}[-\ln(U_1)\sin(4\pi U_2)].$$

Next, we are given that U_1 and U_2 are independent, thus functions of these random variables are also independent. We thus get:

$$\mathbb{E}[-\ln(U_1)]\mathbb{E}[\sin(4\pi U_2)]]$$

We can then solve the second part of the equation:

$$\mathbb{E}[\sin(4\pi U_2)] = \int_0^1 x \cdot \sin(4\pi x) = 0$$

. Giving us

$$\mathbb{E}[-\ln(U_1)] \cdot 0 = 0$$

Thus Z_0 and Z_1 are independent.

$\square$

3 Solve Exercise 34: Prove Lemma 7.11

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* $\mu_0 := (\mu_0(s_1), \mu_0(s2), ..., \mu_0(sk))$, *where* $\mu_0(s_i) = \mathbb{P}(X_0 = s_i)$, *and transition matrices*

$$(P_1, P2, ...), P_t := (P_t(s_i, s_j))_{(s_i,s_j)\in\mathbb{X}x\mathbb{X}}, t \in \{1, 2, ...\}$$

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* $\mu_t = \mu_0 P_1 P_2 ... P_t$

Proof. We start by applying the law of total probability

$$\mathbb{P}(X_t = s_i) = \sum_{s_{t-1}} \mathbb{P}(X_t = s_i | X_{t-1} = s_{i-1})\mathbb{P}(X_{t_1} = s_{i-1}) = \sum_{s_{t-1}} P_t(s_{i-1}, s_i)\mathbb{P}(X_{t_1} = s_{i-1})$$

Since t is arbitrary, we can apply it again until we reach X_0.

$$\mathbb{P}(X_t = s_t) = \sum_{s_1,...,s_{t-1}} P_t(s_{i-1}, s_i)...P_1(s_1, s_2)\mathbb{P}(X_0 = s_1)$$

Since this is a sequence of multiplications of different transition matrices, we can rewrite this as:

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

□

4 Solve Exercise 35: Proof of Theorem 7.14

Theorem. *Let $W_1, ..., W \overset{iid}{\sim} F$ such that (p_t, W_t) is a RMR for a transition matrix P_t, for all $t \in \mathbb{N}$. Then if $X_0 \sim \mu_0$, $X_t := p_t(X_{t-1}, W_t), t \in \mathbb{N}$, is a Markov chain with initial distribution μ_0 and transition Matrix P_t at time t.*

Proof. From the definition of a Markov chain we know that

$$\mathbb{P}(X_t = x_t | X_0 = x_0, ..., X_{t-1} = x_{t-1})$$

We insert from the above theorem $X_t := p_t(X_{t-1}, W_t), t \in \mathbb{N}$

$$= \mathbb{P}(p_t(X_{t-1}, W_t) = x_t | X_0 = x_0, ..., X_{t-1} = x_{t-1})$$

$$= \mathbb{P}(p_t(x_{t-1}, W_t) = x_t | X_0 = x_0, ..., X_{t-1} = x_{t-1})$$

W_t is independent of $X_0, ..., X_{t-1}$, therefore

$$\mathbb{P}(p_t(X_{t-1}, W_t) = x_t)$$

W_t is iid with the same distribution as W, thus

$$\mathbb{P}(p_t(X_{t-1}, W) = x_t)$$

Using the definition 7.12 of Random Mapping Representations, we can rewrite to

$$P_t(x_{t-1}, x_t) = P_t$$

Further, it follows from the definition of a Markov chain that $(X_t)_{t\in\mathbb{N}}$ has an initial distribution μ_0. Consequently, the above theorem is proved. □

5 Solve Exercise 37: Prove Proposition 7.23

Proposition. *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* $(X_t)_{t\in\mathbb{Z}_+}$*.*

Proof. π is a reversible distribution for $(X_t)_{t\in\mathbb{Z}_+}$, therefore for every pair states $(x, y) \in \mathbb{X}^2$ and transition matrix P:

$$\pi(x)P(x, y) = \pi(y)P(y, x)$$

To show that π is a stationary distribution, we must show the following 2 conditions:

1) it is a probability distribution: $\pi(x) \geq 0$ for each $x \in \mathbb{X}$ and $\displaystyle\sum_{x\in\mathbb{X}} \pi(x) = 1$

2) it is a fixed point: $\pi P = \pi$ i.e., $\displaystyle\sum_{x\in\mathbb{X}} \pi(x)P(x, y) = \pi(y)$ for each $y \in \mathbb{X}$

1) It follows from the definition of π that $\pi(x) \geq 0$, $\pi(y) \geq 0$, as well as $\sum_{x\in\mathbb{X}} \pi(x) = 1$, $\sum_{x\in\mathbb{X}} \pi(y) = 1$.

To show 2), we rewrite the property of a reversible distribution:

$$\pi(x)P(x, y) = \pi(y)P(y, x)$$
$$\iff \sum_{x\in\mathbb{X}} \pi(x)P(x, y) = \sum_{x\in\mathbb{X}} \pi(y)P(y, x)$$

Because $\pi(y)$ is independent of the sum, we can rewrite

$$\iff \sum_{x\in\mathbb{X}} \pi(x)P(x, y) = \pi(y) \sum_{x\in\mathbb{X}} P(y, x)$$

$\sum_{x\in\mathbb{X}} P(y, x)$ is equivalent to 1, therefore

$$\iff \sum_{x\in\mathbb{X}} \pi(x)P(x, y) = \pi(y) \cdot 1$$
$$\iff \sum_{x\in\mathbb{X}} \pi(x)P(x, y) = \pi(y)$$
$$\iff \pi P = \pi$$

which proofs proposition 7.23.

$\square$

6 Solve Exercise 38: Prove Proposition 7.25

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

Proof. We need to show that the following property holds for any pair of states $(x, y) \in \mathbb{X}^2$:

$$\pi(x)P(x, y) = \pi(y)P(y, x)$$

For arbitrary vertices v_i, v_j, there are 2 possibilities: Either $(v_i, v_j) \in \mathbb{E}$, or $(v_i, v_j) \notin \mathbb{E}$.

First, we consider the case where $(v_i, v_j) \in \mathbb{E}$. For any v_i, v_j:

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

We insert according to the definitions:

$$\iff \frac{deg(v_i)}{d} \cdot \frac{1}{deg(v_i)} = \frac{deg(v_j}{d} \cdot \frac{1}{deg(v_j}$$

$$\iff \frac{1}{d} = \frac{1}{d}$$

Therefore, for any $(v_i, v_j) \in \mathbb{E}$, we can say that $\pi(v_i)P(v_i, v_j) = \pi(v_j)P(v_j, v_i)$ holds without loss of generality.

Now consider the case where $(v_i, v_j) \notin \mathbb{E}$.

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

We insert again according to the definitions:

$$\iff \frac{deg(v_i)}{d} \cdot 0 = \frac{deg(v_j}{d} \cdot 0$$

$$\iff 0 = 0$$

For any $(v_i, v_j) \notin \mathbb{E}$, we can say that $\pi(v_i)P(v_i, v_j) = \pi(v_j)P(v_j, v_i)$ holds without loss of generality.

We can therefore conclude that the random walk on a connected undirected graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ is a reversible Markov chain with reversible distribution π. □

7 Solve Exercise 46

8 Solve Exercise 43

In the above we are mentioning that R needs to be nice enough, Why is that? Does 0-1 loss work? Why?
Furthermore, we used the tower property to derive 8.4 from 8.3, how does this work?

Proof. Part 1: In the above we are mentioning that R needs to be nice enough, Why is that? Does 0-1 loss work? Why?
According to **Definition 8.7.** Given a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, and assume that $Z = ((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)) \overset{\text{IID}}{\sim} F(x, y)$ is a sequence of $\mathbb{R}^{m+1}$ valued random variables taking values in the data space $\mathbb{X} \times \mathbb{Y}$. We define the empirical risk for a function $g : \mathbb{X} \to \mathbb{Y}$ as

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

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

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

The 0-1 loss function is

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

that is, the loss is 1 if y is the wrong value and 0 if it is correct. The pattern recognition problem is the problem of minimizing the functional

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

Hence, 0-1 loss function works.

Part 2: Furthermore, we used the tower property to derive 8.4 from 8.3, how does this work?
Since the testing dataset is independent of the training dataset and hence $\hat{\phi}$ is independent of the testing data, we deduce using Hoeffdings inequality that if $\hat{R}_m(\phi)$ denotes the empirical risk over the testing dataset we have (Provided R is nice enough)

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

According to **Theorem 2.50** (The tower property). Let (X, Y) be a $\mathbb{R}^2$ valued RV.

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

Assume we use 0-1 function, then

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

We use tower property

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

The RHS of expectation is itself, we have

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

and prove

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

□