POST GRADUATE DEGREE PROGRAMME (CBCS) IN

MATHEMATICS

SEMESTER IV

SELF LEARNING MATERIAL

PAPER : DSE 4.2
(Pure and Applied Streams)

Advanced Operations Research II

Directorate of Open and Distance Learning
University of Kalyani
Kalyani, Nadia
West Bengal, India

Content Writers

Advanced Operations Research II	Dr. Sahidul Islam Associate Professor Department of Mathematics University of Kalyani

April, 2024

Directorate of Open and Distance Learning, University of Kalyani

Published by the Directorate of Open and Distance Learning

University of Kalyani, 741235, West Bengal

Director's Message

Satisfying the varied needs of distance learners, overcoming the obstacle of Distance and reaching the unreached students are the three fold functions catered by Open and Distance Learning (ODL) systems. The onus lies on writers, editors, production professionals and other personnel involved in the process to overcome the challenges inherent to curriculum design and production of relevant Self Learning Materials (SLMs). At the University of Kalyani a dedicated team under the able guidance of the Hon'ble Vice-Chancellorhas invested its best efforts, professionally and in keeping with the demands of Post Graduate CBCS Programmes in Distance Mode to devise a self-sufficient curriculum for each course offered by the Directorate of Open and Distance Learning (DODL), University of Kalyani.

Development of printed SLMs for students admitted to the DODL within a limited time to cater to the academic requirements of the Course as per standards set by Distance Education Bureau of the University Grants Commission, New Delhi, India under Open and Distance Mode UGC Regulations, 2020 had been our endeavor. We are happy to have achieved our goal.

Utmost care and precision have been ensured in the development of the SLMs, making them useful to the learners, besides avoiding errors as far as practicable. Further suggestions from the stakeholders in this would be welcome.

During the production-process of the SLMs, the team continuously received positive stimulations and feedback from Professor **(Dr.) Amalendu Bhunia, Hon'ble Vice-Chancellor, University of Kalyani**, who kindly accorded directions, encouragements and suggestions, offered constructive criticism to develop it with in proper requirements. We gracefully, acknowledge his inspiration and guidance.

Sincere gratitude is due to the respective chairpersons as well as each and every member of PGBOS (DODL), University of Kalyani. Heartfelt thanks are also due to the Course Writers-faculty members at the DODL, subject-experts serving at University Post Graduate departments and also to the authors and academicians whose academic contributions have enriched the SLMs. We humbly acknowledge their valuable academic contributions. I would especially like to convey gratitude to all other University dignitaries and personnel involved either at the conceptual or operational level of the DODL of University of Kalyani.

Their persistent and coordinated efforts have resulted in the compilation of comprehensive, learner-friendly, flexible texts that meet the curriculum requirements of the Post Graduate Programme through Distance Mode.

All the Self Learning Materials are self written and collected from e-book, journals and websites.

Director

Directorate of Open and Distance Learning

University of Kalyani

Post Graduate Board of Studies
Department of Mathematics
Directorate of Open and Distance Learning
University of Kalyani

Sl No.	Name & Designation	Role
1	Dr. Samares Pal, Professor & Head Department of Mathematics, University of Kalyani	Chairperson
2	Dr. Pulak Sahoo, Professor Department of Mathematics, University of Kalyani	Member
3	Dr. Sahidul Islam, Associate Professor Department of Mathematics, University of Kalyani	Member
4	Dr. Sushanta Kumar Mohanta, Professor Department of Mathematics, West Bengal State University	External Nominated Member
5	Ms. Audrija Choudhury, Assistant Professor Department of Mathematics Directorate of Open and Distance Learning University of Kalyani	Member
6	Director Directorate of Open and Distance Learning University of Kalyani	Convener

Discipline Specific Elective Paper

DSE 4.2
Marks : 100 (SEE : 80; IA : 20)

Advanced Operations Research II
(Applied and Pure Streams)

Syllabus

- **Unit 1:** Reliability: Elements of Reliability theory, failure rate, extreme value distribution.
- **Unit 2:** Analysis of stochastically falling equipments including the reliability function, reliability and growth model.
- **Unit 3:** Information Theory: Information concept, expected information.
- **Unit 4:** Entropy and properties of entropy function.
- **Unit 5:** Bivariate Information theory,
- **Unit 6:** Economic relations involving conditional probabilities,
- **Unit 7:** Coding theory: Communication system, encoding and decoding.
- **Unit 8:** Shannon-Fano encoding procedure.
- **Unit 9:** Haffman encoding, noiseless coding theory, noisy coding.
- **Unit 10:** Family of codes, Hamming code.
- **Unit 11:** Golay code, BCH codes, Reed-Muller code, Perfect code, codes and design.
- **Unit 12:** Linear codes and their dual, weight distribution.
- **Unit 13:** Markovian Decision Process: Ergodic matrices, regular matrices.
- **Unit 14:** Imbedded Markov Chain method for Steady State solution.
- **Unit 15:** Posynomial, Signomial, Degree of difficulty, Unconstrained minimization problems, Solution using Differential Calculus, Solution seeking Arithmetic-Geometric inequality, Primal dual relationship & sufficiency conditions in the unconstrained case,
- **Unit 16:** Constrained minimization, Solution of a constrained Geometric Programming problem, Geometric programming with mixed inequality constrains, Complementary Geometric programming.
- **Unit 17:** A brief introduction to Inventory Control, Single-item deterministic models without shortages.

- **Unit 18:** Single-item deterministic models with shortages Dynamic Demand Inventory Models.
- **Unit 19:** Multi-item inventory models with the limitations on warehouse capacity
- **Unit 20:** Models with price breaks, single-item stochastic models without Set-up cost and with Set-up cost, Average inventory capacity, Capital investment.

Contents

Unit 1

Course Structure

- Reliability: Elements of Reliability theory, failure rate, extreme value distribution

---

1.1 Introduction

Reliability is the probability of a device performing its purpose adequately for the period of time intended under the operating conditions encountered. The definition brings into focus, four important factors, namely,

- the reliability of a device is expressed as a probability;
- the device is required to give adequate performance;
- the duration of adequate performance is specified;
- the environment or operating conditions are prescribed.

Some of the important aspects of reliability are:

a) Reliability is a function of time. We could not expect an almost wornout light bulb to be as reliable as one recently put into service.

b) Reliability is a function of conditions to use. In very severe environments, we expect to encounter frequent system breakdowns than in normal environments.

c) Reliability is expected as a probability which helps us to quantify it and think of optimizing system reliability.

1.2 Reliability

Definition 1.2.1. Hazard Rate/Failure Rate: Failure rate is the ratio of the number of failures during a particular unit interval to the average population during that interval. Thus the failure rate for the ith interval is

$$\frac{n_i}{\frac{1}{2}\left[\left(N-\sum_{k=1}^{i-1} n_k\right)+\left(N-\sum_{k=1}^{i} n_k\right)\right]},$$

where n_i is the number of failures during the ith interval and N is the total number of components.

Definition 1.2.2. Failure Density: The failure density in a particular unit interval is the ratio of the number of failures during that interval to the number of components. So the failure density during the ith interval is

$$\frac{n_i}{N} = f_{d_i}.$$

Let l be the last interval after which there are no intervals. Then

$$f_{d_l} = \frac{n_l}{N}.$$

Thus,

$$f_{d_1} + f_{d_2} + \cdots + f_{d_l} = \frac{1}{N}(n_1 + n_2 + \cdots + n_l) = \frac{N}{N} = 1.$$

Hence the sum of values entered in column 5 is 1 (Table 1.1).

Definition 1.2.3. Reliability: Reliability(R), is the ratio of the number of survivals at any given time to the total initial population. That is, reliability at ith time is

$$R(i) = \frac{s_i}{N},$$

s_i is the number of survivals during the ith interval.

Definition 1.2.4. Probability of failure: The concept of probability of failure is similar to that of the concept of probability of survival. This is the ratio of the number of units failed within a certain time to the total population.

Hence, the probability of failure within ith time is

$$\frac{n_1 + n_2 + \cdots + n_i}{N} \quad \text{or} \quad \frac{F_i}{N},$$

so that the probability of failure at ith time plus reliability at ith time is

$$\frac{F_i}{N} + \frac{s_i}{N} = 1$$

(since $F_i + s_i = N$), that is, probability of failure and reliability at the same time is always 1.

Definition 1.2.5. Mean Failure Rate(h)**:** If Z_1 is the failure rate for the first unit of time, Z_2 is the failure rate for the second unit of time, ..., Z_T is the failure rate for the Tth unit of time, then the mean failure rate for T times will be

$$h(T) = \frac{Z_1 + Z_2 + \cdots + Z_T}{T}.$$

The mean failure rate is also obtained from the formula

$$\frac{1}{T}\left[\frac{N(0) - N(T)}{N(0)}\right],$$

where $N(0)$ is the total population at $t = 0$ and $N(T)$ is the total population remaining at time $t = T$.

Definition 1.2.6. Mean time to failure (MTTF): In general, if t_1 is the time to failure for the first specimen, t_2 is the time to failure for the second specimen, ..., t_N is the time to failure for the Nth specimen, then the MTTF for N specimens is

$$\frac{t_1 + t_2 + \cdots + t_N}{N}.$$

Time(t)	Number of failures(n)	Cumulative failures(F)	Number of Survivals(S)	Failure density(f_d)	Failure/ Hazard rate(Z)	Reliability
0	0	0	1000	0	0	1
1	130	130	870	$\frac{130}{1000} = 0.130$	$\frac{130}{\frac{1000+870}{2}} = 0.139$	$1 - 0.130 = 0.870$
2	83	213	787	0.083	$\frac{83}{\frac{870+787}{2}} = 0.100$	$1 - (0.130 + 0.083) = 0.787$
3	75	288	712	0.075	$\frac{75}{\frac{787+712}{2}} = 0.100$	0.712
4	68	356	644	0.068	$\frac{68}{\frac{712+644}{2}} = 0.100$	0.644
5	62	418	582	0.062	$\frac{62}{\frac{644+582}{2}} = 0.101$	0.582
6	56	474	526	0.056	$\frac{56}{\frac{582+526}{2}} = 0.101$	0.526
7	51	525	475	0.051	$\frac{51}{\frac{526+475}{2}} = 0.101$	0.475
8	46	571	429	0.046	$\frac{46}{\frac{475+429}{2}} = 0.102$	0.429
9	41	612	388	0.041	$\frac{41}{\frac{429+388}{2}} = 0.100$	0.388
10	37	659	341	0.037	$\frac{37}{\frac{388+341}{2}} = 0.101$	0.341
11	34	683	317	0.034	$\frac{34}{\frac{341+317}{2}} = 0.103$	0.317
12	31	714	286	0.031	$\frac{31}{\frac{317+286}{2}} = 0.102$	0.286
13	28	742	258	0.028	$\frac{28}{\frac{286+258}{2}} = 0.102$	0.258
14	64	806	194	0.064	$\frac{64}{\frac{258+194}{2}} = 0.283$	0.194
15	76	882	118	0.076	$\frac{76}{\frac{194+118}{2}} = 0.487$	0.118
16	62	944	56	0.062	$\frac{62}{\frac{118+56}{2}} = 0.713$	0.056
17	40	984	16	0.040	$\frac{40}{\frac{56+16}{2}} = 1.111$	0.016
18	12	996	4	0.012	$\frac{12}{\frac{16+4}{2}} = 1.2$	0.004
19	4	1000	0	0.004	$\frac{4}{\frac{4+0}{2}} = 2$	0.000

Table 1.1

If n_1 is the number of specimens that failed during first unit of time, n_2 be that during second unit of time, ..., n_l be that during the last (lth) unit of time, then the MTTF for the N specimens will be

$$MTTF = \frac{n_1 + 2n_2 + \cdots + ln_l}{N},$$

where $N = n_1 + n_2 + \cdots + n_l$. If the time interval is δt unit instead of 1 unit, then

$$\begin{aligned} MTTF &= \frac{n_1 + 2n_2 + \cdots + ln_l}{N}\delta t \\ &= \frac{\sum_{k=1}^{l} kn_k}{N}\delta t. \end{aligned}$$

Example 1.2.7. In the life testing of 100 specimens of a particular device, the number of failures during each time interval of 20 hours is shown in the following table:

Time Interval (T)(in hours)	Number of failures during the interval
$T \leq 100$	0
$1000 < T \leq 1020$	25
$1020 < T \leq 1040$	40
$1040 < T \leq 1060$	20
$1060 < T \leq 1080$	10

Estimate the MTTF for these specimens.

Solution. As the number of specimens tested is large, it is tedious to record the time of failure for each specimen. So we note the number of specimen that fail during each 20 hours interval. Thus

$$MTTF = \frac{(0 \times 1000) + (25 \times 1020) + (40 \times 1040) + (20 \times 1060) + (10 \times 1080)}{100} = 1040 \text{ hrs.}$$

■

Example 1.2.8. The following table gives the results of tests conducted under severe adverse conditions on 1000 safety valves. Calculate the failure density $f_d(t)$ and the hazard rates $Z(t)$ where the time interval is 4 hours instead of 1 hour.

Time Interval (in hours)	Number of failures (h)
$t = 0$	0
$0 < t \leq 4$	267
$4 < t \leq 8$	59
$8 < t \leq 12$	36
$12 < t \leq 16$	24
$16 < t \leq 20$	23
$20 < t \leq 24$	11

Solution.

Time interval	Number of failures	Cumulative frequency	Number of Survivals(S)	Failure density(f_d)	Failure/ Hazard rate($Z(t)$)	Reliability (R)
$t = 0$	0	0	1000	0	0	1
$0 < t \leq 4$	267	267	733	0.067	$\frac{267}{\frac{4(1000+733)}{2}} = 0.077$	$1 - 0.067 = 0.933$
$4 < t \leq 8$	59	326	674	0.0148	$\frac{56}{\frac{4(733+674)}{2}} = 0.021$	$1 - (0.067 + 0.0148)$ $= 0.9182$
$8 < t \leq 12$	36	362	638	0.009	$\frac{36}{\frac{4(674+638)}{2}} = 0.014$	0.9092
$12 < t \leq 16$	24	386	614	0.006	$\frac{24}{\frac{4(638+614)}{2}} = 0.009$	0.9032
$16 < t \leq 20$	23	409	591	0.0057	$\frac{23}{\frac{4(614+591)}{2}} = 0.009$	0.8975
$20 < t \leq 24$	11	420	580	0.0027	$\frac{11}{\frac{4(591+580)}{2}} = 0.0047$	0.8948

■

Four Important Points

(i) Sum of the failure densities is 1, that is,

$$f_{d_1} + f_{d_2} + \cdots + f_{d_l} = \sum_{i=1}^{l} f_{d_i} = 1 \text{ (For discrete case)}$$

$$\int_0^T f_d(\xi) d\xi = 1,$$

where the limits of the integration are taken from the beginning of the first at $t = 0$ till the end where all the specimens failed at time $t = T$.

(ii) The reliability $R(i)$ for the ith hour is given by

$$\begin{aligned} R(i) &= 1 - (f_{d_1} + f_{d_2} + \cdots + f_{d_i}) \\ &= 1 - \sum_{k=1}^{i} f_{d_k} \ \text{[For discrete case]} \end{aligned}$$

Hence the reliability $R(t)$, for the tth hour for continuous case is given by

$$\begin{aligned} R(t) &= 1 - \int_0^t f_d(\xi)d\xi \\ &= \int_t^T f_d(\xi)d\xi \text{[For continuous case]}. \end{aligned}$$

(iii) The probability of failure in hours, $F(i)$ is given by

$$F(i) = f_{d_1} + f_{d_2} + \cdots + f_{d_i} = \sum_{k=1}^{i} f_{d_k},$$

that is, $R(i) + F(i) = 1$.

Since the reliability and probability of failure are complementary so, $R(t) + F(t) = 1$. Thus for continuous case,

$$F(t) = 1 - R(t) = \int_0^t f_d(\xi)d\xi.$$

(iv) The failure rate or hazard rate for the ith hour is

$$Z(i) = \frac{n_i}{\frac{1}{2}\left(N - \sum_{k=1}^{i-1} n_k\right)} = \frac{2[R(i-1) - R(i)]}{R(i-1) + R(i)}.$$

(For one hour interval between $t = (i-1)$ hr to $t = i$ hr)

If the interval is δt, instead of 1 hour, then for continuous case,

$$\begin{aligned} Z(t) &= \frac{2[R(t-\delta t) - R(t)]}{[R(t-\delta t) + R(t)]\delta t} \\ \text{i.e, } Z(t+\delta t) &= \frac{2[R(t) - R(t+\delta t)]}{[R(t) + R(t+\delta t)]\delta t} \end{aligned}$$

For continuous case, when $\delta t \to 0$, we have

$$\begin{aligned} \lim_{\delta t \to 0} Z(t+\delta t) &= \lim_{\delta t \to 0} \frac{2[R(t) - R(t+\delta t)]}{[R(t) + R(t+\delta t)]\delta t} \\ \Rightarrow Z(t) &= \lim_{\delta t \to 0} \frac{R(t) - R(t+\delta t)}{R(t)\delta t} \\ &= -\frac{1}{R(t)} \lim_{\delta t \to 0} \frac{R(t+\delta t) - R(t)}{\delta t} \\ &= -\frac{1}{R(t)} \frac{d}{dt}(R(t)) \\ &= -\frac{R'(t)}{R(t)} \end{aligned} \tag{1.2.1}$$

Thus,

$$\begin{aligned} \int_0^t Z(t)dt &= -[\log R(t)]_0^t \\ \Rightarrow \log R(t) &= \log R(0) - \int_0^t Z(t)dt \end{aligned}$$

Since at $t = 0$, $R(0) = 1$, that is, $\log R(0) = 0$, thus,

$$R(t) = \mathrm{e}^{-\int_0^t Z(\xi)d\xi} \tag{1.2.2}$$

Finally we shall get an expression for $f_d(t)$ for continuous case.

By definition, we have

$$\begin{aligned} f_d(t+\delta t) &= \frac{(\text{no. of survivals at time } t=t) - (\text{no. of survivals at time } t=t+\delta t)}{\delta t.(\text{total number of survivals})} \\ \text{or, } f_d(t+\delta t) &= \left[\left(\frac{\text{no. of survivals at } t=t}{\text{total no. of survivals}}\right) - \left(\frac{\text{no. of survivals at } t=t+\delta t}{\text{total no. of survivals}}\right)\right]\frac{1}{\delta t} \\ &= \frac{1}{\delta t}[R(t) - R(t+\delta t)] \end{aligned}$$

Letting $\delta t \to 0$, we get for continuous case,

$$f_d(t) = -\lim_{\delta t \to 0} \frac{R(t+\delta t) - R(t)}{\delta t} = -R'(t) \tag{1.2.3}$$

From equations (1.2.1) and (1.2.3), we get

$$\begin{aligned} Z(t) &= \frac{f_d(t)}{R(t)} \\ \Rightarrow f_d(t) &= Z(t)R(t) \\ &= Z(t)\,\mathrm{e}^{-\int_0^t Z(\xi)d\xi} \end{aligned} \tag{1.2.4}$$

1.3 MTTF in terms of failure density

The mean time to failure is given by

$$\text{MTTF} = \frac{\left(\sum_{k=1}^{l} kn_k\right)\delta t}{N},$$

where N is the initial total survivals; n_1 is the total no. of specimens that failed during the first δt time interval, n_2 is the total no. of specimens that failed during the second δt time interval, ... , n_k is the total no. of specimens failed during the kth δt interval. Now, by definition

$$\begin{aligned} f_{d_k} &= \frac{n_k}{N.\delta t} \\ \Rightarrow \frac{n_k}{N} &= f_{d_k}\delta t. \end{aligned}$$

Further, $f\delta t$ is the elapsed time t. Hence the expression for MTTF can be written as

$$\text{MTTF} = \sum_{k=1}^{l}(k.f_{d_k}.\delta t)\delta t = \sum_{k=1}^{l} f_{d_k}(k\delta t)\delta t \tag{1.3.1}$$

where the summation is for the period from the first δt time interval to lth δt interval.

For continuous case, when $\delta t \to 0$, and $f\delta t$ is the elapsed time t and f_{d_k} will be the failure density $f_d(t)$ at time t, then

$$\text{MTTF} = \int_0^T t f_d(t) dt, \tag{1.3.2}$$

where T is the number of hours after which there are no survivals.

Now we have, $F(t) + R(t) = 1$. Thus,

$$F(t) = 1 - R(t) = \int_0^t f_d(\xi) d\xi$$

Thus,

$$\frac{d}{dt}(F(t)) = -\frac{d}{dt}(R(t)) = f_d(t).$$

Thus,

$$\begin{aligned}
\text{MTTF} &= \int_0^\infty t f_d(t) dt \quad [\text{For } t > T, \text{ there are no survivals, so the values of the integration is } 0, \text{ for } t > T] \\
&= \int_0^\infty -t \frac{d}{dt}(R(t)) dt \\
&= -[t.R(t)]_0^\infty + \int_0^\infty 1.R(t) dt \\
&= \int_0^\infty R(t) dt \quad [\text{Since } R(0) = 1 \text{ and } R(\infty) = 0 \text{ as } t \to \infty, \text{ there are no survivals}]
\end{aligned} \tag{1.3.3}$$

Also, when $t_1 \leq t \leq t_2$, we have,

$$F(t_2) - F(t_1) = \int_{t_1}^{t_2} f_d(\xi) d\xi.$$

For continuous case, when the hazard rate is constant, that is, $Z(t) = \lambda$, a constant, say, then

$$\int_0^t Z(\xi) d\xi = \int_0^t \lambda d\xi = \lambda t.$$

Thus,

$$R(t) = \mathrm{e}^{-\int_0^t Z(\xi) d\xi} = \mathrm{e}^{-\lambda t}$$

and $F(t) = 1 - \mathrm{e}^{-\lambda t}$. Similarly,

$$f_d(t) = Z(t) \times R(t) = \lambda \, \mathrm{e}^{-\lambda t}.$$

Thus,

$$\begin{aligned}
\text{MTTF} &= \int_0^\infty R(t) dt \\
&= \int_0^\infty \mathrm{e}^{-\lambda t} \, dt \\
&= -\left[\frac{\mathrm{e}^{-\lambda t}}{\lambda}\right]_0^\infty = \frac{1}{\lambda}.
\end{aligned}$$

Thus, for a constant hazard model, the MTTF is simply the reciprocal of the hazard rate.

The constant hazard rate is also known as the exponential reliability rate.

$$\begin{aligned}\text{MTTF} = \int_0^\infty t f_d(t)dt &= \int_0^\infty t\,\mathrm{e}^{-\lambda t}\,dt \\ &= \lambda\left[\frac{t\,\mathrm{e}^{-\lambda t}}{-\lambda}\right]_0^\infty + \int_0^\infty \frac{\lambda}{\lambda}\,\mathrm{e}^{-\lambda t}\,dt \\ &= 0 + \int_0^\infty \mathrm{e}^{-\lambda t}\,dt = \frac{1}{\lambda}.\end{aligned}$$

The mean of $\lambda\,\mathrm{e}^{-\lambda t}$ is

$$\int_0^\infty \lambda t\,\mathrm{e}^{-\lambda t}\,dt = \frac{1}{\lambda}.$$

Example 1.3.1. It is found that the random variations with respect to time in the output voltage of a particular system are exponentially distributed with a mean value 100V. What is the probability that the output voltage will be found at any time to lie in the range $90 - 100$V?

Solution. For an exponential distribution, the MTTF is the reciprocal of the hazard rate λ(say), where λ is a constant, that is, MTTF$= \frac{1}{\lambda}$.

Here, we identify the MTTF with a mean value 100V. Thus,

$$\frac{1}{\lambda} = 100 \Rightarrow \lambda = 0.01.$$

Hence the p.d.f $f_d(t)$ for the voltage distribution is $= \lambda\,\mathrm{e}^{-\lambda t} = 0.01 \times \mathrm{e}^{-0.01t}$.

Now, the probability that the voltage lies between V_1 and V_2 is given by

$$\begin{aligned}F(V_2) - F(V_1) &= \int_{V_1}^{V_2} f_d(t)dt \\ &= \int_{V_1}^{V_2} \lambda\,\mathrm{e}^{-\lambda t}\,dt \\ &= 1 - \mathrm{e}^{-\lambda(V_2 - V_1)}.\end{aligned}$$

Here, $V_2 = 100$V, $V_1 = 90$V.

Hence, $F(100) - F(90) = 1 - \mathrm{e}^{-0.01(100-90)} = 1 - \mathrm{e}^{-0.1} \simeq 0.095$. ■

Example 1.3.2. It is observed that the failure pattern of an electronic system follows an exponential distribution with mean time to failure of 100 hours. What is the probability that the system failure occurs within 750 hours?

Solution. MTTF $= \frac{1}{\lambda} = 1000$, where λ is the constant hazard rate. Thus,

$$\lambda = \frac{1}{1000}.$$

Thus,

$$f_d(t) = \lambda\,\mathrm{e}^{-\lambda t}$$

Hence the probability that the system failure occurs within a period V is

$$F(V) = \int_0^V f_d(t)dt = \int_0^V \lambda\,\mathrm{e}^{-\lambda t}\,dt = 1 - \mathrm{e}^{-\lambda V}.$$

Here, $V = 750$ hrs. and $\lambda = 0.001$. Thus,

$$F(750) = 1 - \mathrm{e}^{-0.750} \simeq 0.528.$$

■

Unit 2

Course Structure

- Linearly Increasing Hazard
- System Reliability
- Redundancy

2.1 Linearly Increasing Hazard

Here the hazard increases linearly with time, that is, $Z(t) = kt$, where k is a constant. Thus the time integral of $Z(t)$ is given by

$$\int_0^t ktdt = \frac{k}{2}t^2$$

Therefore,

$$R(t) = \mathrm{e}^{-\int_0^t Z(\xi)d\xi} = \mathrm{e}^{-\frac{k}{2}\,t^2}$$

And thus,

$$f_d(t) = Z(t) \times R(t) = kt\,\mathrm{e}^{-\frac{k}{2}t^2}$$

This function $f_d(t) = kt\,\mathrm{e}^{-\frac{k}{2}t^2}$ is known as the **Rayleigh density function**.

Now,

$$\begin{aligned}\frac{d}{dt}(f_d(t)) &= k\,\mathrm{e}^{-\frac{k}{2}t^2} + kt\left(-\frac{k}{2}.2t\,\mathrm{e}^{-\frac{k}{2}t^2}\right)\\ &= k\,\mathrm{e}^{-\frac{k}{2}t^2}[1-kt^2]\end{aligned}$$

and

$$\begin{aligned}\frac{d^2}{dt^2}(f_d(t)) &= k\left(-\frac{k}{2}.2t\right)\mathrm{e}^{-\frac{k}{2}t^2}[1-kt^2] + k\,\mathrm{e}^{-\frac{k}{2}t^2}[-2kt]\\ &= -k^2t\,\mathrm{e}^{-\frac{k}{2}t^2}[1-kt^2] - 2k^2t\,\mathrm{e}^{-\frac{k}{2}t^2}\\ &= -3k^2t\,\mathrm{e}^{-\frac{k}{2}t^2} + k^3t^3\,\mathrm{e}^{-\frac{k}{2}t^2}\end{aligned}$$

Now,

$$\frac{d}{dt}(f_d(t)) = 0 \Rightarrow t = \frac{1}{\sqrt{k}} \text{ [since } t > 0]$$

At $t = \frac{1}{\sqrt{k}}$,

$$\begin{aligned}
\frac{d^2}{dt^2}(f_d(t)) &= \frac{-3k^2}{\sqrt{k}}\, e^{-\frac{k}{2}t^2} + \frac{k^3}{k\sqrt{k}}\, e^{-\frac{k}{2}t^2} \\
&= -\frac{2k^2}{\sqrt{k}}\, e^{-\frac{k}{2}t^2} \\
&= -2k\sqrt{k}\, e^{-\frac{k}{2}t^2} \\
&= = -2k\sqrt{k}\, e^{-\frac{k}{2}\frac{1}{k}} \\
&= -\frac{2k\sqrt{k}}{\sqrt{e}} < 0
\end{aligned}$$

Thus, $f_d(t)$ is maximum at $t = \frac{1}{\sqrt{k}}$. Thus

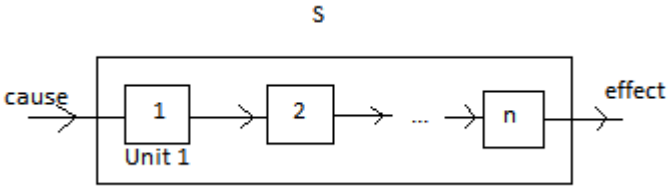
$$f_d(t)\big|_{t=\frac{1}{\sqrt{k}}} = k.\frac{1}{\sqrt{k}}\, e^{-1/2} = \sqrt{k}\, e^{-1/2} = \sqrt{\frac{k}{e}}.$$

Hence $f_d(t)$ reaches a maximum value $\sqrt{\frac{k}{e}}$ at $t = \frac{1}{\sqrt{k}}$ and tends to zero as t becomes larger. Now, we calculate the MTTF when the hazard rate increases linearly.

$$\begin{aligned}
\text{MTTF} = \int_0^\infty R(t)dt &= \int_0^\infty e^{-k/2t^2}\, dt \\
&= \sqrt{\frac{2}{k}} \int_0^\infty e^{-z^2}\, dz \qquad [\text{Put } \sqrt{\frac{k}{2}}t = z] \\
&= \sqrt{\frac{2}{k}}.\frac{\sqrt{\pi}}{2} \\
&= \sqrt{\frac{\pi}{2k}}.
\end{aligned}$$

2.2 System Reliability

A. **Series Configuration:** The simplest combination of units that form a system is a series combination. This is one of the most commonly used structures and is shown in the following figure: The system S



consists of n units which are connected in series as shown. Let the successful operation of these individual units be represented by $X_1, X_2, \ldots, X_n$ and their respective probabilities by $P(X_1), P(X_2), \ldots, P(X_n)$.

For the successful operation of the system, it is necessary that all n units function satisfactorily. Hence the probability of the successful operation of all the units is $P(X_1 \text{ and } X_2 \text{ and } \cdots \text{and } X_n)$.

We shall assume that these units are not independent of one another, that is, the successful operation of unit 1 might affect the successful operation of all other units and so on.

The system reliability is given by

$$\begin{aligned} P(S) &= P(X_1 \text{ and } X_2 \text{ and } \cdots \text{and } X_n) \\ &= P(X_1).P(X_2|X_1).P(X_3|X_1X_2)\ldots P(X_n|X_1X_2\ldots X_{n-1}). \end{aligned}$$

If they are independent, then

$$P(S) = P(X_1)P(X_2)\ldots P(X_n).$$

Example 2.2.1. In a hydraulic control system, the connecting linkage has a reliability factor 0.98 and the valve which has a reliability factor 0.92. Also the pressure sensor which activates the linkage, has a reliability factor 0.90. Assume that all the three elements namely the activator, the linkage and the hydraulic valve are connected in series with independent reliability factors. What is the reliability of the control system?

Solution. Let the successful operation of the elements namely the activator, the linkage and the hydraulic valve be denoted by X_1, X_2 and X_3 respectively. Thus,

$$P(X_1) = 0.98, \;\; P(X_2) = 0.91, \;\; P(X_3) = 0.90 \;\; \text{(given)}$$

Since these elements are connected in series with independent reliability factors, hence the reliability of the control system, S(say) is

$$P(S) = P(X_1)P(X_2)P(X_3) = 0.98 \times 0.92 \times 0.90 = 0.81144.$$

■

Note 2.2.2. There is an important point that the reliability of a series system is always worse than the poorest component of the system.

Example 2.2.3. If the system consists of n identical units in series and if each unit has a reliability factor p, determine the system reliability under the assumption that all units function independently.

Solution. $P(S) = p.p\ldots p(\text{n times}) = p^n$. Now, if q is the probability of failure of each unit, then $p = 1 - q$.

Hence the system reliability

$$P(S) = p^n = (1-q)^n = 1 - nq + \cdots$$

If q is very small, this expression can be approximated to $1 - nq$. Thus,

$$P(S) \simeq 1 - nq.$$

■

Example 2.2.4. A system has 10 identical equipments. It is desired that the system reliability be 0.95. Determine how good each component should be?

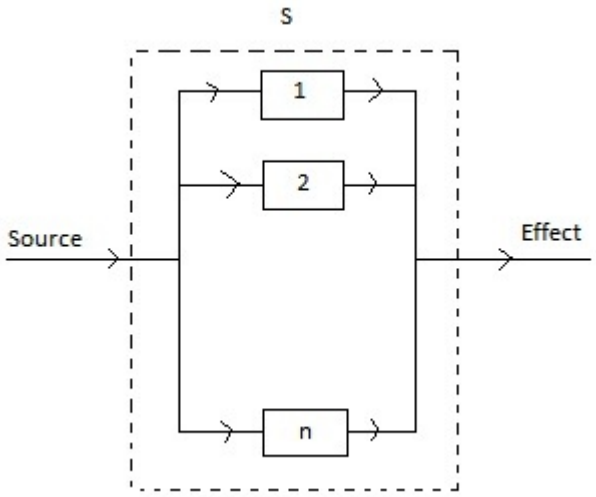
Solution. Let p be the reliability factor of each equipment. Then

$$P(S) = 0.95 = p^{10} \Rightarrow p = \sqrt[10]{0.95} = 0.99488.$$

■

B. **Parallel Configuration:** Several systems exist in which successful operation depends on the satisfactory functioning of any one of their n subsystems or elements. These are said to be connected in parallel. We can also ass a system in which several signal paths perform the same operation and the satisfactory performance of any one of these paths is sufficient to ensure the successful operation of the system. The elements of such a system are said to be connected in parallel.

A block diagram representing a parallel configuration is shown in the figure below The reliability of the



system can be calculated very easily by considering the conditions for system failure.

Let $X_1, X_2, \ldots, X_n$ represent successful operation of units $1, 2, \ldots, n$ respectively. Similarly, let $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_n$ respectively represent their successful operation, that is, the failure of the units.

If $P(X_1)$ is the probability of successful operation of unit 1, then $P(\overline{X}_1) = 1 - P(X_1)$, and so on. For the complete failure of the system S, all the n units have to fail simultaneously. If $P(\overline{S})$ is the probability of failure of the system, then

$$\begin{aligned} P(\overline{S}) &= P(\overline{X}_1 \text{ and } \overline{X}_2 \text{ and } \cdots \text{ and } \overline{X}_n) \\ &= P(\overline{X}_1)P(\overline{X}_2|\overline{X}_1)P(\overline{X}_3|\overline{X}_1\overline{X}_2)\ldots P(\overline{X}_n|\overline{X}_1\overline{X}_2\ldots\overline{X}_{n-1}) \end{aligned}$$

The expression $P(\overline{X}_3|\overline{X}_1\overline{X}_2)$ represents the probability of failure of unit 3 under the condition that units 1 and 2 have failed.

The other terms can also be interpreted in the same manner. If the unit failures are independent of one another, then

$$\begin{aligned} P(\overline{S}) &= P(\overline{X}_1)P(\overline{X}_2)\ldots P(\overline{X}_n) \\ &= [1 - P(X_1)][1 - P(X_2)]\ldots[1 - P(X_n)]. \end{aligned}$$

Since if any one of them does not fail, then the problem of successful configuration of the system is

$$P(S) = 1 - P(\overline{S}).$$

For independent cases, $P(S) = 1 - [1 - P(X_1)][1 - P(X_2)] \dots [1 - P(X_n)]$. If the n elements are identical and the unit failures are independent of one another, then

$$P(S) = 1 - (1 - P(X))^n$$

where, $P(X) = P(X_1) = P(X_2) = \cdots = P(X_n)$.

Example 2.2.5. Consider a system consisting of three identical units connected in parallel. The unit reliability factor is 0.10. If the unit failures are independent of one another and if the successful operation of the system depends on the satisfactory performance of any one unit, then determine the system reliability.

Solution. $P(S) = 1 - (1 - 0.10)^3 = 1 - 0.729 = 0.271$. This reveals the important fact that a parallel configuration can greatly increase system reliability with just three elements connected in parallel. ■

Example 2.2.6. A parallel system is composed of 10 independent identical components. If the system reliability $P(S)$, is to be 0.95, how poor can the components be?

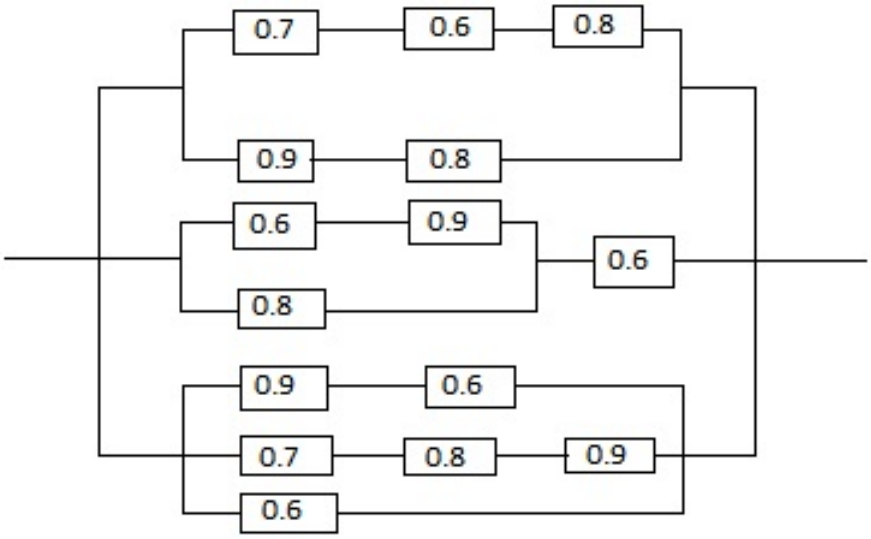
Solution. Let $P(X)$ be the probability of successful operation of each component. Thus,

$$\begin{aligned} P(S) &= 1 - (1 - P(X))^1 0 = 0.95 \\ \Rightarrow (1 - P(X))^1 0 &= 1 - 0.95 \\ &= 0.05 \\ \Rightarrow 1 - P(X) &= \sqrt[10]{0.05} = 0.74113 \\ \Rightarrow P(X) &= 1 - 0.74113 = 0.25887. \end{aligned}$$

Each component can have a very low reliability factor of 0.2589 but still gives the system a reliability factor as high as 0.95. ■

C. **Mixed Configuration:** Consider the following example:

Example 2.2.7. Find the reliability of the above system:



(KU 2011)

Solution. The complete system is composed of the following subsystems:

S_1: $\boxed{0.7} \longrightarrow \boxed{0.6} \longrightarrow \boxed{0.8}$
S_2: $\boxed{0.9} \longrightarrow \boxed{0.8}$
S_3: $\boxed{0.6} \longrightarrow \boxed{0.9}$
S_4: $\boxed{0.9} \longrightarrow \boxed{0.6}$
S_5: $\boxed{0.7} \longrightarrow \boxed{0.8} \longrightarrow \boxed{0.9}$
S_6: $S_1||S_2$
S_7: $S_3||\boxed{0.8}$
S_8: $S_4||S_5||\boxed{0.6}$
S_9: $S_7 \longrightarrow \boxed{0.6}$
S_{10}: $S_6||S_9||S_8$

Now,

$$\begin{aligned}
P(S_1) &= 0.7 \times 0.6 \times 0.8 = 0.336 \\
P(S_2) &= 0.9 \times 0.8 = 0.72 \\
P(S_3) &= 0.6 \times 0.9 = 0.54 \\
P(S_4) &= 0.9 \times 0.6 = 0.54 \\
P(S_5) &= 0.7 \times 0.8 \times 0.9 = 0.504 \\
P(S_6) &= 1 - [(1 - P(S_1))(1 - P(S_2))] \\
&= 1 - [(1 - 0.336)(1 - 0.72)] = 0.81408 \\
P(S_7) &= 1 - [(1 - P(S_3))(1 - 0.8)] \\
&= 1 - [(1 - 0.54)(1 - 0.8)] = 0.908 \\
P(S_8) &= 1 - [(1 - P(S_4))(1 - P(S_5))(1 - 0.6)] \\
&= 1 - [(1 - 0.54)(1 - 0.504)(1 - 0.6)] = 0.908736 \\
P(S_9) &= P(S_7) \times 0.6 = 0.908 \times 0.6 = 0.5448 \\
P(S_{10}) &= 1 - [(1 - P(S_6))(1 - P(S_9))(1 - P(S_8))] \\
&= 1 - [(1 - 0.81408)(1 - 0.5448)(1 - 0.908736)] \simeq 0.99228.
\end{aligned}$$

Hence the system reliability is 0.99228. ■

2.3 Redundancy

If the state of art is such that either it is not possible to produce highly reliable components or the cost of producing such components is very high, then we can improve the system reliability by the technique of introducing redundancies.

This involves the deliberate creation of new parallel path in a system. If two elements A, B with probability of success $P(A)$ and $P(B)$ are connected in parallel, then the probability of the successful operation of the system,

$$\begin{aligned}
P(A \text{ or } B) &= P(A) + P(B) - P(A \text{ and } B) \\
&= P(A) + P(B) - P(A)P(B),
\end{aligned}$$

assuming that the elements are independent.

Since both $P(A)$ and $P(B)$ are less than 1, excluding the condition where $P(A) = P(B) = 1$, then their product is always less than both $P(A)$ and $P(B)$.

This illustrates a simple method of improving the reliability of a system when the element reliability cannot be increased. Although either one of the elements is sufficient for the successful operation of the system, we deliberately use both elements so as to increase the reliability causing the system to become redundant.

Unit 3

Course Structure

- Information Theory: Fundamentals of Information theory
- Measures of information and characterisation

3.1 Introduction

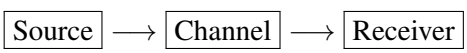
In everyday life we observe that there are numerous means for the transmission of information. For example, the information is usually transmitted by means of a human voice, i.e., as in telephone, radio, television etc., by means of letters, newspapers, books etc. We often come across sentences like

- We have received a lot of information about the postponement of examination.
- We have a bit of information that he will be appointed as a professor.

But few people have suspected that it is really possible to measure information quantitatively. An amount of information has a useful numeric value just like an amount of sugar or an amount of bank balance. For example, suppose a man goes to a new community to rent a house and asks an unreliable agent "is this house cool in summer season?" If the agent answers 'yes', the man has received very little information, because more than likely that agent would have answered 'yes' regardless of the facts. If on the other hand, the man has a friend who lives in a neighbouring house, he can get more information by asking his friend the same question because the answer will be more reliable.

In general way it would appear that the amount of information in the message should be measured by extent of the change in probability produced by the message. There will be atleast three essential parts of simplest communication system:

- Transmitter or Source,
- Communication channel or transmission network which carries the message from the transmitter to the receiver,
- Receiver or Sink



3.2 Fundamental theorem of information theory

It is possible to transmit information through a noisy channel at any rate less than the channel capacity with an arbitrarily small probability of error.

3.2.1 Origination

The information theory is an appealing name assigned to a scientific discipline which deals with the mathematical theory of communication. The origin of information theory dates back to the work of R.V. Hartley ("Transmission of informations", Bellsys technical journal vol. 7, 1928), who tried to develop a quantitative measure of information in the telecommunication system. The field of information theory grown considerably often the publication of C.E. Shannon's ("A mathematical theory of communication", Bellsys technical journal, vol. 27, 1948). Information theory answers two fundamental questions in communication system.

a) What is the ultimate data compression?

b) What is the ultimate data transmission rate?

For this reason, some consider information theory as a subset of communication theory. Indeed it has fundamental contribution in statistical physics, computer science, probability and statistics, Biology, Economics etc. We see information only when we are in doubt which arises when there are number of alternatives and we are uncertain about the outcome of the event. On the other hand, if the event can occur in just one way, there is no uncertainty about it and no information is called for we get some information by the occurrence of the event when there was some uncertainty before its occurrence. Therefore, the amount of information received must be equal to the amount of uncertainty may be before the occurrence of the event.

3.3 Measure of information and characterisation

Let E be an event and p be its probability of occurrence. If we are told that the event E has occurred, then the question is "what is the amount of information conveyed by this message?" If p is close to 1, then it is nearly certain to occur and hence it conveys very little information. On the other hand, if p is close 0, then it is almost certain that E will not occur and consequently the message starting with its occurrence is quite unexpected. In general, let E_1 and E_2 are two events with p_1 and p_2 as their probability of occurrence respectively and let $p_1 < p_2$.

Then the event E_2 is more likely to occur and so the message conveying the occurrence of E_2 contains low information (bit information) than that conveying the occurrence of E_1. Further if p_2 continually decreased to p_1, the uncertainty associated with the occurrence of E_2 increases continually corresponding to the event E_1.

The above intuitive idea suggested that the measure of information conveyed by the message stating the occurrence of event with the probability p must be a function of p only, say $h(p)$, which is non-negative, strictly decreasing, continuous and $h(1) = 0$. Also $h(p)$ is very large when p is nearly equal to 0.

Next consider two events E_1 and E_2 with probability of occurrence p_1 and p_2 respectively. If we are told that the event E_1 has occurred, then we have received an amount of information $h(p_1)$. Giving this message, the probability that E_2 will occur is

$$p_{21} = p(E_2|E_1).$$

Suppose now we are told that the event E_2 has also occurred. Then the additional amount of information received is $h(p_{21})$.

Therefore the total amount of information received form their two successive messages is

$$h(p_1) + h(p_{21}).$$

Assume that the events E_1 and E_2 are independent. Then

$$p_{21} = p_2.$$

So the total amount of information received in this case is $h(p_1) + h(p_2)$.

Again, the probability of both the events E_1 and E_2 is p_1p_2 and the amount of information conveyed by the message stating that both the events E_1 and E_2 have occurred is $h(p_1p_2)$.

So from the above considerations we have

$$h(p_1p_2) = h(p_1) + h(p_2).$$

Thus from the above discussion we see that the amount of information received from the message stating that the event E with probability p has occurred is a function of p only, say $h(p)$ and has the following characterisations.

(i) $h(p)$ is non-negative, continuous and strictly decreasing function in p in $(0, 1]$.

(ii) $h(1) = 0$ and $h(p)$ is very large when p is very close to 0, i.e., $h(p) \to \infty$ as $p \to 0$.

(iii) if E_1 and E_2 are independent events with probability of occurrence p_1 and p_2 respectively, then the amount of information conveyed by the message stating that the occurrence of both events E_1 and E_2 is equal to the amount of information conveyed by the massage dealing with the event E_1 plus the amount of information dealing with the event E_2, i.e.,

$$h(p_1p_2) = h(p_1) + h(p_2).$$

Theorem 3.3.1. Let $h(p)$ denote the amount of information received form the message stating the event E with probability p has occurred. Then

$$h(p) = -k \log p,$$

where, k is a positive constant.

Proof. The function $h(p)$ has the following properties:

(i) $h(p)$ is non-negative, continuous, strictly decreasing in $(0, 1]$.

(ii) $h(1) = 0$ and $h(p) \to \infty$ as $p \to 0$.

(iii) $h(p_1p_2) = h(p_1) + h(p_2)$.

Take any $p \in (0, 1]$ and let n be a positive integer. We first show that

$$h(p^n) = nh(p) \tag{3.3.1}$$

Clearly, (3.3.1) holds for $n = 1$.

Assume that (3.3.1) holds for the positive integer n. Then

$$\begin{aligned} h(p^{n+1}) &= h(p^n \cdot p) \\ &= h(p^n) + h(p) \quad \text{[using property (iii)]} \\ &= n\,h(p) + h(p) \\ &= (n+1)\,h(p) \end{aligned}$$

Therefore, (3.3.1) holds for the positive integer $(n+1)$.

Hence by the principle of finite induction, (3.3.1) holds for all $n \in \mathbb{N}$.

Let $p \in (0,1]$ and $n \in \mathbb{N}$. Consider $q = p^{1/n}$ and $q \in (0,1]$.

$$\begin{aligned} &\therefore p = q^n \text{ and } h(p) = h(q^n) = n\,h(q) \text{ (By (3.3.1))}. \\ &\Rightarrow h(q) = \frac{1}{n}h(p) \\ &\Rightarrow h(p^{1/n}) = \frac{1}{n}h(p) \end{aligned} \tag{3.3.2}$$

Let r be a positive rational number and $r = \dfrac{m}{n}$, where $m, n \in \mathbb{N}$.

$$\begin{aligned} \text{Then } h(p^r) &= h(p^{m/n}) \\ &= h\Big(\big(p^{1/n}\big)^m\Big) \\ &= m\,h(p^{1/n}) \\ &= \frac{m}{n}h(p) \\ &= r\,h(p) \end{aligned}$$

Let r be any positive number. Then choose any sequence $\{r_n\}$ of positive rational numbers such that $r_n \to r$ as $n \to \infty$. For such n, we get

$$h(p^{r_n}) = r_n\,h(p).$$

Since h is a constant function, letting $n \to \infty$, we get,

$$h(p^r) = r\,h(p) \tag{3.3.3}$$

Putting $p = \frac{1}{2}$ in (3.3.3) we get

$$h\left(\left(\frac{1}{2}\right)^r\right) = r\,h\left(\frac{1}{2}\right) \tag{3.3.4}$$

Let $p \in (0,1]$. We write $r = \dfrac{\log p}{\log 1/2}$ so that $r > 0$ and $\left(\dfrac{1}{2}\right)^r = p$. Substituting in (3.3.4), we get

$$h(p) = -\frac{h(1/2)}{\log 2}\log p = -k \log p \quad \text{where} \quad k = \frac{h(1/2)}{\log 2}$$

Since h is strictly decreasing and $h(1) = 0$, therefore

$$h(1/2) > 0 \quad \text{and so} \quad k > 0.$$

□

3.3.1 Units of information

Taking $k = 1$, we have $h(p) = -\log p$. The choice of the base of the logarithmic amounts to the choice of the units of information,

(i) when base 2, i.e., $h(p) = -\log_2 p$, the unit is 'bits'.

(ii) when base is natural 'e', unit is 'nats'

(iii) when base is 10, unit is 'Hartley'

Note 3.3.2. 1 Har =3.32 bits and 1 nat =1.44 bits

Unit 4

Course Structure

- Entropy and its properties

4.1 Entropy (Shannon's Definition)

Let X be the random variable with range $\{x_1, x_2, \ldots, x_n\}$ and probability mass function (p.m.f)

$$\rho_X(x) = \begin{cases} p_i & \text{for } x = x_i \ (i = 1, 2, \ldots, n) \\ 0 & \text{otherwise.} \end{cases}$$

Then the quantity $-\sum_{i=1}^{n} p_i \log p_i$ is called the *entropy* of the random variable X and is denoted by $H(x)$ or $H_n(p_1, p_2, \ldots, p_n)$.

$$\therefore \text{ We have } H(X) = H_n(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i.$$

Clearly, $H(X) \geq 0$.

Note 4.1.1. $x \log x \to 0$ as $x \to 0$ and we have used the conversion that $0 \log 0 = 0$.

4.1.1 Units of entropy

(i) when base is 2, unit of entropy is bits

(ii) when base is e, unit of entropy is nats

(iii) when base is 10, unit of entropy is Hartley

4.1.2 Properties of entropy function

(Shannon's characterization of entropy function)

1. For a fixed n, $H_n(p_1, p_2, \ldots, p_n)$ is a continuous function of $p_1, p_2, \ldots, p_n$ $(0 \leq p_i \leq 1,\ i = 1, 2, \ldots, n)$. It is obvious from the definition of H_n.

2. If $p_i = \frac{1}{n},\ i = 1, 2, \ldots, n$, then

$$\begin{aligned} H_n(p_1, p_2, \ldots, p_n) &= H_n\left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right) \\ &= -\sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n} \\ &= \log n. \end{aligned}$$

   So, $H_n\left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right)$ is strictly increasing function of n.

3.

$$\begin{aligned} \text{Let } s_1 &= p_1 + p_2 + \ldots + p_{n_1} \\ s_2 &= p_{n_1+1} + p_{n_1+2} + \ldots + p_{n_2} \\ \ldots &\quad \ldots \quad \ldots\ldots\ldots\ldots \\ s_k &= p_{n_{k-1}+1} + p_{n_{k-1}+2} + \ldots + p_{n_k}, \quad (n_k = n) \\ \text{and } m_1 &= n_1,\ m_2 = n_2 - n_1, \ldots, m_k = n_k - n_{k-1} \end{aligned}$$

Then,

$$\begin{aligned} H_n(p_1, p_2, \ldots, p_n) &= H_k(s_1, s_2, \ldots, s_k) + s_1 H_{m_1}\left(\frac{p_1}{s_1}, \frac{p_2}{s_1}, \cdots, \frac{p_{n_1}}{s_1}\right) \\ &+ s_2 H_{m_2}\left(\frac{p_{n_1+1}}{s_2}, \cdots, \frac{p_{n_2}}{s_2}\right) + \cdots + s_k H_{m_k}\left(\frac{p_{n_{k-1}+1}}{s_k}, \cdots, \frac{p_{n_k}}{s_k}\right). \end{aligned}$$

The above relation may be expressed as follows:

If a random experiment is decomposed into several successive ones, then the original value of H is equal to the weighted sum of the corresponding values of H with weights $1, s_1, s_2, \ldots, s_k$.

Now, we have $H_k(s_1, s_2, \ldots, s_k) = -\sum_{i=1}^{k} s_i \log s_i$.

$$\begin{aligned} \therefore s_1 H_{m_1}\left(\frac{p_1}{s_1}, \frac{p_2}{s_1}, \ldots, \frac{p_{n_1}}{s_1}\right) &= -s_1 \sum_{i=1}^{n_1} \frac{p_i}{s_1} \log \frac{p_i}{s_1} \\ &= -\sum_{i=1}^{n_1} p_i \log p_i + \sum_{i=1}^{n_1} p_i \log s_1 \\ &= -\sum_{i=1}^{n_1} p_i \log p_i + s_1 \log s_1. \end{aligned}$$

Similarly, we find

$$s_2 H_{m_2}\left(\frac{p_{n_1+1}}{s_2},\ldots,\frac{p_{n_2}}{s_2}\right) = -\sum_{i=n_1+1}^{n_2} p_i \log p_i + s_2 \log s_2$$

$$\vdots$$

$$s_k H_{m_k}\left(\frac{p_{n_{k-1}}+1}{s_k},\ldots,\frac{p_{n_k}}{s_k}\right) = -\sum_{i=n_{k-1}+1}^{n_k} p_i \log p_i + s_k \log s_k$$

Adding the above expressions, we get

$$\begin{aligned} H_n(p_1,p_2,\ldots,p_n) &= -\sum_{i=1}^{n_k} p_i \log p_i \\ &= -\sum_{i=1}^{n} p_i \log p_i, \quad \text{where } n_k = n. \end{aligned}$$

Theorem 4.1.2. For a fixed n, the entropy function $H_n(p_1,p_2,\ldots,p_n)$ is maximum when $p_1 = p_2 = \ldots = p_n = \frac{1}{n}$ and $H_n(\max) = \log n$.

Proof. We first show that $\log x \leq x-1$ for all $x>0$ and the equality holds for $x=1$.

Let $\phi(x) = x - 1 - \log x$ for all $x > 0$.

$\therefore \phi'(x) = 1 - \frac{1}{x}$.

If $x > 1$, then $\phi'(x) > 0$ and if $0 < x < 1$, then $\phi'(x) < 0$.

So $\phi(x)$ is a strictly increasing function in $(1,\infty)$ and strictly decreasing in $(0,1)$.

Therefore, $\phi(x) \geq \phi(1) = 0$ for all $x > 0$.

$$\therefore \log x \leq x - 1 \text{ for all } x > 0 \tag{4.1.1}$$

Let us take $x = \dfrac{1}{np_i}$ in (4.1.1) and we get

$$\begin{aligned} &\log\frac{1}{np_i} \leq \frac{1}{np_i} - 1 \\ \Rightarrow\ & p_i \log\frac{1}{np_i} \leq \frac{1}{n} - p_i \\ \Rightarrow\ & -\sum_{i=1}^{n} p_i \log p_i - \sum_{i=1}^{n} p_i \log n \leq 1 - \sum_{i=1}^{n} p_i \\ \Rightarrow\ & -\sum_{i=1}^{n} p_i \log p_i \leq \log n \quad \left(\because \sum_{i=1}^{n} p_i = 1\right) \\ \Rightarrow\ & H_n(p_1,p_2,\ldots,p_n) \leq \log n \end{aligned} \tag{4.1.2}$$

When $p_1 = p_2 = \ldots = p_n = \frac{1}{n}$, then

$$\begin{aligned} H_n(p_1, p_2, \ldots, p_n) &= H_n\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right) \\ &= -\sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n} \\ &= \log n \end{aligned} \tag{4.1.3}$$

From (4.1.2) and (4.1.3) we see that when the events are equally likely, H_n is maximum and its maximum value is $\log n$ i.e.,

$$H_n(\max) = \log n$$

□

Note 4.1.3. In this case units are taken as 'nats', since

$$\log_e x = \log_D x \log_e D \quad \text{for any} \quad D \geq 2.$$

Note 4.1.4. The entropy of X may be interpreted as the expected value of the function $\log \frac{1}{p_i}$ where p_i is the p.m.f of X. Thus

$$E\left[\log \frac{1}{p_i}\right] = \sum_{i=1}^{n} p_i \log \frac{1}{p_i} = -\sum_{i=1}^{n} p_i \log p_i = H(X).$$

Unit 5

Course Structure

- Bivariate Information Theory
- Joint, conditional and relative entropies
- Mutual Information

5.1 Joint, conditional and relative entropies

Let X, Y be two discrete random variables with ranges $\{x_1, x_2, \ldots, x_m\}$ and $\{y_1, y_2, \ldots, y_n\}$ respectively and probability mass functions $p(x)$ and $q(y)$ and joint p.m.f $p(x, y) = P(X = x; Y = y)$.

i) The joint entropy, $H(X, Y)$ of the pair of random variables X, Y is defined as

$$\begin{aligned} H(X,Y) &= -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log p(x_i, y_i) \\ &= E\left[\log \frac{1}{p(x,y)}\right] \end{aligned} \tag{5.1.1}$$

(ii) The conditional entropy, $H(X|Y)$ is defined by

$$\begin{aligned} H(X|Y) &= \sum_{j=1}^{n} q(y_j) H(X|Y = y_j) \\ &= -\sum_{j=1}^{n} q(y_j) \sum_{i=1}^{m} p(x_i|y_j) \log p(x_i|y_j) \\ &= -\sum_{i=1}^{m}\sum_{j=1}^{n} q(y_j) p(x_i|y_j) \log p(x_i|y_j) \\ &= -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log p(x_i|y_j) \\ &= E_{p(x,y)}\left[\log \frac{1}{p(x|y)}\right] \end{aligned}$$

Similarly, we can show that $H(Y|X) = E_{p(x,y)}\left[\log \frac{1}{p(y|x)}\right]$.

(iii) The relative entropy or Kullback-leibler distance between two probability mass functions $p(x)$ and $q(x)$ with $X = \{x_1, x_2, \ldots, x_m\}$ is defined as

$$\begin{aligned} D(p||q) &= \sum_{i=1}^{m} p(x_i) \log \frac{p(x_i)}{q(x_i)} \\ &= E_{p(x)}\left[\log \frac{p(x)}{q(x)}\right] \end{aligned}$$

5.2 Mutual information

Let X and Y be two discrete random variables with ranges $X = \{x_1, x_2, \ldots, x_m\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$ respectively and probability mass functions $p(x)$ and $q(y)$ with joint p.m.f $p(x, y) = p(X = x; Y = y)$. Then the mutual information of the random variables X and Y is denoted by $I(X; Y)$ and is defined by

$$\begin{aligned} I(X, Y) &= \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)q(y_j)} \\ &= D\big(p(x, y)||p(x)q(y)\big) \\ &= E_{p(x,y)}\left[\log \frac{p(x, y)}{p(x)q(y)}\right]. \end{aligned}$$

Theorem 5.2.1. Let $p_1, p_2, \ldots, p_n$ and $q_1, q_2, \ldots, q_n$ be two sets of non-negative numbers and $\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i$, then

$$\sum_{i=1}^{n} p_i \log_D q_i \leq \sum_{i=1}^{n} p_i \log_D p_i,$$

where D is any positive number greater than 1. Equality holds if and only if $p_i = q_i$ for all i.

Proof. We use the convention $0 \log 0 = 0$. First consider the case when $D = e$ and $p_i > 0,\ q_i > 0$ for all $i = 1, 2, \ldots, n$.

For any positive number x, we have

$$\log x \leq (x - 1) \tag{5.2.1}$$

equality holds if and only if $x = 1$. Taking $x = \frac{q_i}{p_i}$ in (5.2.1), we get

$$\log \frac{q_i}{p_i} \leq \frac{q_i}{p_i} - 1$$

Multiplying by p_i and taking summation we get

$$\begin{aligned} &\sum_{i=1}^{n} p_i \log \frac{q_i}{p_i} \leq \sum_{i=1}^{n} (q_i - p_i) = 0 \\ \Rightarrow &\sum_{i=1}^{n} p_i \log q_i \leq \sum_{i=1}^{n} p_i \log p_i \end{aligned} \tag{5.2.2}$$

Now, let $p_k = 0$ for some k and $q_k \neq 0$, but $p_i > 0,\ q_i > 0$ for $i \neq k$. Then clearly (5.2.2) holds because $p_k \log p_k = 0$ and $p_k \log q_k = 0$ if $q_k = 0$ for some k but $q_k \neq 0$. $p_k \log q_k = -\infty$ and so (5.2.2) holds.

Suppose that the equality holds in (5.2.2). Also assume that $p_k \neq q_k$ for some k. Then $\frac{q_k}{p_k} \neq 1$ and so

$$\log \frac{q_k}{p_k} < \frac{q_k}{p_k} - 1.$$

This gives

$$\begin{aligned} \sum_{i=1}^{n} p_i \log \frac{q_i}{p_i} &< \sum_{i=1}^{n} (q_i - p_i) = 0 \\ \Rightarrow \sum_{i=1}^{n} p_i \log q_i &< \sum_{i=1}^{n} p_i \log p_i \end{aligned}$$

which contradicts (5.2.2) since here equality does not hold because $\frac{q_k}{p_k} \neq 1$.

$$\therefore p_i = q_i \quad \text{for all } i.$$

Now, let $D \neq e$ for any $x > 0$. Then $\log_D x = \log_D e \cdot \log_e x$ and $\log_D e > 0$. So, multiplying (5.2.2) by $\log_D e$ we get

$$\sum_{i=1}^{n} p_i \log_D q_i \leq \sum_{i=1}^{n} p_i \log_D p_i.$$

□

Theorem 5.2.2. For any two discrete random variables X and Y

$$H(X,Y) \leq H(X) + H(Y)$$

Equality holds if and only if X, Y are independent.

Proof. Let X, Y be two discrete random variables with ranges $X = \{x_1, x_2, \ldots, x_m\}$, $Y = \{y_1, y_2, \ldots, y_n\}$ and probability mass functions (p.m.f) $p(x)$ and $q(y)$ with the joint p.m.f $p(x,y) = p(X = x; Y = y)$. We have

$$\begin{aligned} H(X) + H(Y) &= -\sum_{i=1}^{m} p(x_i) \log p(x_i) - \sum_{j=1}^{n} q(y_j) \log q(y_j) \\ &= -\sum_{i=1}^{m} \sum_{j=1}^{n} p(x_i, y_j) \log p(x_i) - \sum_{i=1}^{m} \sum_{j=1}^{n} p(x_i, y_j) \log q(y_j) \\ &= -\sum_{i=1}^{m} \sum_{j=1}^{n} p(x_i, y_j) \log \Big(p(x_i) q(y_j) \Big) \\ \text{Also, } H(X,Y) &= -\sum_{i=1}^{m} \sum_{j=1}^{n} p(x_i, y_j) \log p(x_i, y_j) \\ \text{Now, } \sum_{i=1}^{m} \sum_{j=1}^{n} p(x_i) q(y_j) &= \sum_{i=1}^{m} p(x_i) \sum_{j=1}^{n} q(y_j) = 1 \quad \text{and} \quad \sum_{i=1}^{m} \sum_{j=1}^{n} p(x_i, y_j) = 1 \end{aligned}$$

By Theorem 5.2.1, we have

$$\begin{aligned}
\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log p(x_i,y_j) &\geq \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log\Big(p(x_i)q(y_j)\Big)\\
\Rightarrow -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log p(x_i,y_j) &\leq -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log\Big(p(x_i)q(y_j)\Big)\\
\Rightarrow H(X,Y) &\leq H(X)+H(Y)
\end{aligned}$$

Equality holds if and only if $p(x_i,y_j)=p(x_i)q(y_j)$, i.e., if and if $X,\ Y$ are independent random variables. □

Theorem 5.2.3. For any two discrete random variables X and Y

$$H(X,Y)=H(X)+H(Y|X)=H(Y)+H(X|Y)$$

Proof. Let $X,\ Y$ be two discrete random variables with ranges $X=\{x_1,x_2,\ldots,x_m\}$ and $Y=\{y_1,y_2,\ldots,y_n\}$ and p.m.f $p(x)$ and $q(y)$ with joint p.m.f $p(x,y)=p(X=x;Y=y)$. Then

$$\begin{aligned}
H(X)+H(Y|X) &= -\sum_{i=1}^{m} p(x_i)\log p(x_i) - \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log p(y_j|x_i)\\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log p(x_i) - -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log p(y_j|x_i)\\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log\Big(p(x_i)\cdot p(y_j|x_i)\Big)\\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log p(x_i,y_j) \quad [\because p(x_i)p(y_j|x_i)=p(x_i,y_j)]\\
&= H(X,Y)
\end{aligned}$$

In a similar way, we can show that

$$H(Y)+H(X|Y)=H(X,Y)$$

□

Theorem 5.2.4. For any two discrete random variables X and Y,

$$I(X,Y)=H(X)-H(X|Y)=H(Y)-H(Y|X)$$

Proof. We have

$$\begin{aligned}
H(X)-H(X|Y) &= -\sum_{i=1}^{m} p(x_i)\log p(x_i) + \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log p(x_i|y_j)\\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log p(x_i) + \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log p(x_i|y_j)\\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log\frac{p(x_i|y_j)}{p(x_i)}\\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log\frac{p(x_i,y_j)}{p(x_i)q(y_j)}\\
&= I(X,Y)
\end{aligned}$$

Similarly, we can show that

$$H(Y) - H(Y|X) = I(X;Y)$$

□

Note 5.2.5. I(X;Y)=I(Y,X)

Theorem 5.2.6. For any three discrete random variables $X,\ Y,\ Z$,

$$H\big((X,Y)|Z\big) = H(X|Z) + H\big(Y|(X,Z)\big)$$

Proof. Let X, Y, Z be three discrete random variables with ranges $X = \{x_1, x_2, \ldots, x_m\}$, $Y = \{y_1, y_2, \ldots, y_n\}$ and $Z = \{z_1, z_2, \ldots, z_k\}$ respectively and probability mass functions are $p(x),\ q(y)$ and $r(z)$ with joint p.m.f $p(x,y,z) = p(X = x; Y = y; Z = z)$. Then

$$\begin{aligned}
H(X|Z) + H\big(Y|(X,Z)\big) &= -\sum_{i=1}^{m}\sum_{l=1}^{k} p(x_i,z_l)\log p(x_i|z_l) - \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k} p(x_i,y_j,z_l)\log p(y_j|(x_i,z_l)) \\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k} p(x_i,y_j,z_l)\log p(x_j|z_l) - \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k} p(x_i,y_j,z_l)\log p(y_j|(x_i,z_l)) \\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k} p(x_i,y_j,z_l)\log p(x_i|z_l)p(y_j|(x_i,z_l)) \\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k} p(x_i,y_j,z_l)\log\left(\frac{p(x_i,z_l)}{p(z_l)}\cdot\frac{p(x_i,y_j,z_l)}{p(x_i,z_l)}\right) \\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k} p(x_i,y_j,z_l)\log\left(\frac{p(x_i,y_j,z_l)}{p(z_l)}\right) \\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k} p(x_i,y_j,z_l)\log p\big((x_i,y_j)|z_l\big) \\
&= H\Big((X,Y)|Z\Big)
\end{aligned}$$

□

Note 5.2.7. For n random variables

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, X_{i-2}, \ldots, X_1)$$

5.2.1 Conditional mutual information

i) The conditional mutual information of random variables $X,\ Y$ given Z is defined by

$$\begin{aligned}
I(X;Y|Z) &= H(X|Z) - H(X|(Y,Z)) \\
&= E_{p(x,y,z)}\left[\log\frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}\right]
\end{aligned}$$

ii) The conditional mutual information of random variables X and Y given $Z_1, Z_2, \ldots, Z_n$ is defined by

$$\begin{aligned} I(X;Y|Z_1,Z_2,\ldots,Z_n) &= H(X|Z_1,Z_2,\ldots,Z_n) - H(X|(Y,Z_1,Z_2,\ldots,Z_n)) \\ &= E_{p(x,y,z_1,\ldots,z_n)}\left[\log \frac{p(X,Y|Z_1,Z_2,\ldots,Z_n)}{p(X|Z_1,Z_2,\ldots,Z_n)p(Y|Z_1,Z_2,\ldots,Z_n)}\right] \end{aligned}$$

Theorem 5.2.8. (i) For the random variables X, Y, Z

$$I(X;Y,Z) = I(X;Y) + I(X;Z|Y) = I(X;Z) + I(X;Y|Z)$$

Proof.

$$\begin{aligned} I(X;Y) + I(X;Z|Y) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} + \sum_x \sum_y \sum_z p(x,y,z) \log \frac{p(x,z|y)}{p(x|y)p(z|y)} \\ &= \sum_x \sum_y \sum_z p(x,y,z) \log \frac{p(x,y)}{p(x)p(y)} + \sum_x \sum_y \sum_z p(x,y,z) \log \frac{p(x,z|y)}{p(x|y)p(z|y)} \\ &= \sum_x \sum_y \sum_z p(x,y,z) \log \left\{ \frac{p(x,y)}{p(x)p(y)} \cdot \frac{p(x,y,z)}{p(y)} \cdot \frac{p(y)}{p(x,y)} \cdot \frac{p(y)}{p(y,z)} \right\} \\ &= \sum_x \sum_y \sum_z p(x,y,z) \log \frac{p(x,y,z)}{p(x)p(y,z)} \\ &= I(X;Y,Z) \end{aligned}$$

Similarly, we can show that

$$I(X;Z) + I(X;Y|Z) = I(X;Y,Z)$$

□

Theorem 5.2.9. (ii) For the random variables $X_1, X_2, \ldots, X_n, Y$

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, X_{i-2}, \ldots, X_1)$$

Proof. Follows from induction on n. □

Theorem 5.2.10. (Information inequality): Let $p(x)$ and $q(x)$ for $x \in X$ be two probability mass functions. Then

$$D(p||q) \geq 0$$

Proof. Let $X = \{x_1, x_2, \ldots, x_n\}$ and let $p_i = p(x_i),\ q_i = q(x_i),\ i = 1, 2, \ldots, n$. Then by definition,

$$\begin{aligned} D(p||q) &= \sum_{i=1}^{n} p(x_i) \log \frac{p(x_i)}{q(x_i)} \\ &= \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \end{aligned}$$

Also we have $\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i = 1$. So, by Theorem 5.2.1,

$$\begin{aligned} &\sum_{i=1}^{n} p_i \log q_i \leq \sum_{i=1}^{n} p_i \log p_i \\ \Rightarrow &\sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \geq 0 \\ \Rightarrow &D(p||q) \geq 0. \end{aligned}$$

□

Theorem 5.2.11. (Non-negativity of mutual information) For any two random variables X and Y, $I(X;Y) \geq 0$.

Proof. Let X and Y be two discrete random variables with range $\{x_1, x_2, \ldots, x_m\}$ and $\{y_1, y_2, \ldots, y_n\}$ respectively and the p.m.f $p(x)$ and $q(y)$, joint p.m.f $p(x,y) = P(X = x, Y = y)$. Then the mutual information $I(X;Y)$ between X and Y is given by

$$I(X;Y) = \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)q(y_j)}.$$

Now, we have

$$\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) = 1 \quad \text{and} \quad \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i)q(y_j) = \sum_{i=1}^{m} p(x_i) \sum_{j=1}^{n} q(y_j) = 1$$

So by Theorem 5.2.1

$$\begin{aligned} &\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log p(x_i, y_j) \geq \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log \Big(p(x_i)q(y_j)\Big) \\ i.e., \quad &\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)q(y_j)} \geq 0 \\ i.e., \quad &I(X;Y) \geq 0. \end{aligned}$$

□

Theorem 5.2.12. (Non-negativity of conditional mutual information) For any two random variables X and Y given Z, the conditional mutual information $I(X;Y|Z) \geq 0$.

Proof. Let X, Y, Z be three discrete random variables with ranges $\{x_1, x_2, \ldots, x_m\}$, $\{y_1, y_2, \ldots, y_n\}$, $\{z_1, z_2, \ldots, z_k\}$ respectively and probability mass functions $p(x)$, $p(y)$, $p(z)$ with joint p.m.f $p(x,y,z) = P(X = x, Y = y, Z = z)$.

Then by definition,

$$I(X;Y|Z) = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k} p(x_i, y_j, z_l) \log \frac{p(x_i, y_j|z_l)}{p(x_i|z_l)p(y_j|z_l)}$$

$$\text{Now, } \frac{p(x_i, y_j|z_l)}{p(x_i|z_l)p(y_j|z_l)} = \frac{p(x_i, y_j, z_l)}{p(z_l)} \frac{p(z_l)}{p(x_i, z_l)} \frac{p(z_l)}{p(y_j, z_l)}$$

$$= \frac{p(x_i, y_j, z_l)}{\frac{p(x_i, z_l)p(y_j, z_l)}{p(z_l)}}$$

$$\therefore I(X, Y|Z) = \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{l=1}^{k} p(x_i, y_j, z_l) \log \frac{p(x_i, y_j, z_l)}{\frac{p(x_i, z_l)p(y_j, z_l)}{p(z_l)}}$$

$$\text{Now, } \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{l=1}^{k} p(x_i, y_j, z_l) = 1$$

$$\text{and, } \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{l=1}^{k} \frac{p(x_i, z_l)p(y_j, z_l)}{p(z_l)} = \sum_{i=1}^{m} \sum_{l=1}^{k} p(x_i, z_l) \cdot \sum_{j=1}^{n} \frac{p(y_j, z_l)}{p(z_l)}$$

$$= \sum_{i=1}^{m} \sum_{l=1}^{k} p(x_i, z_l) \frac{p(z_l)}{p(z_l)} \quad \left(\because \sum_{j=1}^{n} p(y_j, z_l) = p(z_l) \sum_{j=1}^{n} p(y_j) = p(z_l) \right)$$

$$= \sum_{i=1}^{m} \sum_{l=1}^{k} p(x_i, z_l)$$

$$= 1$$

Therefore, by Theorem 5.2.1, $I(X; Y|Z) \geq 0$. □

Unit 6

Course Structure

- Conditional Relative Entropy
- Channel Capacity
- Redundancy

6.1 Conditional relative entropy

The conditional relative entropy $D(p(y|x)||q(y|x))$ is the average of the relative entropies between the conditional probability mass functions $p(y|x)$ and $q(y|x)$ averaged over the p.m.f $p(x, y)$.

$$\begin{aligned} \therefore D(p(y|x)||q(y|x)) &= \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} \\ &= \sum_x \sum_y p(x, y) \log \frac{p(y|x)}{q(y|x)} \\ &= E_{p(x,y)} \left[\log \frac{p(Y|X)}{q(Y|X)} \right] \end{aligned}$$

6.1.1 Convex and Concave functions

Let I be an interval and $f : I \to \mathbb{R}$ be a function. The function f is said to be convex if for any two points $x_1,\ x_2\ (x_1 \neq x_2)$ in I and $\lambda,\ \mu \geq 0$ with $\lambda + \mu = 1$, the relation

$$f(\lambda x_1 + \mu x_2) \leq \lambda f(x_1) + \mu f(x_2)$$

holds. A function $g : I \to \mathbb{R}$ is said to be concave if $-g$ is convex.

6.1.2 Jensen's Inequality

If $f : \mathbb{R} \to \mathbb{R}$ is a convex function and X is a random variable, then $f(E \cdot X) \leq Ef(X)$, where E is a constant. Moreover, if f strictly convex, then the equality implies that X is constant.

Theorem 6.1.1. (Log-Sum Inequality) Let $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ be two sets of n non-negative numbers. Then

$$\sum_{i=1}^{n} a_i \log_D \frac{a_i}{b_i} \geq \sum_{i=1}^{n} a_i \log_D \left(\frac{\sum a_i}{\sum b_i} \right)$$

where D is any positive number and $D > 1$. Equality holds if and only if $\frac{a_i}{b_i}$ is constant.

Proof. We use the conventions $0 \log 0 = 0,\ a \log \frac{a}{0} = +\infty,\ 0 \log \frac{0}{0} = 0$. Without loss of generality we may assume that $a_i > 0,\ b_i > 0,\ i = 1, 2, \ldots, n$.

Consider the function $f(t) = t \log_D t,\ t > 0$. Therefore, we have

$$f'(t) = (1 + \log_e t) \log_D e$$
$$f''(t) = \frac{1}{t} \log_D e > 0 \text{ for all } t > 0$$

So, $f(t)$ is strictly convex for $t > 0$.

Now consider

$$\lambda = \sum_{i=1}^{n} b_i, \quad \alpha_i = \frac{b_i}{\lambda}, \quad t_i = \frac{a_i}{b_i}$$

Then $\sum_{i=1}^{n} \alpha_i = 1$ and $\alpha_i > 0$ for all i.

So, by Jensen's inequality, we have

$$\sum_{i=1}^{n} \alpha_i f(t_i) \geq f\left(\sum_{i=1}^{n} \alpha_i t_i \right) \tag{6.1.1}$$
$$\Rightarrow \sum_{i=1}^{n} \frac{b_i}{\lambda} \frac{a_i}{b_i} \log_D \left(\frac{a_i}{b_i} \right) \geq \left(\sum_{i=1}^{n} \frac{a_i}{\lambda} \right) \log_D \left(\sum_{i=1}^{n} \frac{b_i}{\lambda} \frac{a_i}{b_i} \right)$$
$$\Rightarrow \sum_{i=1}^{n} a_i \log_D \left(\frac{a_i}{b_i} \right) \geq \sum_{i=1}^{n} a_i \log_D \left(\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \right) \tag{6.1.2}$$

If $\frac{a_i}{b_i}$ = constant = k (say), for $i = 1, 2, \ldots, n$.
Then clearly equality in (6.1.2) holds.

Suppose that equality holds in (6.1.2) i.e., in (6.1.1). Then,

$$t_1 = t_2 = \ldots = t_n$$
$$\Rightarrow \frac{a_1}{b_1} = \frac{a_2}{b_2} = \ldots = \frac{a_n}{b_n}$$
$$i.e., \quad \frac{a_i}{b_i} = \text{ constant}; \quad i = 1, 2, \ldots, n$$

□

Theorem 6.1.2. $D(p||q)$ is convex in pair (p, q) i.e., if $(p_1, q_1),\ (p_2, q_2)$ be two pairs of probability mass functions and $\lambda > 0,\ \mu > 0$ with $\lambda + \mu = 1$, then

$$D((\lambda p_1 + \mu p_2)||(\lambda q_1 + \mu q_2)) \leq \lambda D(p_1||q_1) + \mu D(p_2||q_2).$$

Proof. Let (p_1, q_1) and (p_2, q_2) be two pairs of probability mass functions and $\lambda > 0,\ \mu > 0$ with $\lambda + \mu = 1$. Then by Log-Sum inequality, we have

$$\begin{aligned}(\lambda p_1(x) + \mu p_2(x)) \log \frac{\lambda p_1(x) + \mu p_2(x)}{\lambda q_1(x) + \mu q_2(x)} &\leq \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + \mu p_2(x) \log \frac{\mu p_2(x)}{\mu q_2(x)} \\ &= \lambda p_1(x) \log \frac{p_1(x)}{q_1(x)} + \mu p_2(x) \log \frac{p_2(x)}{q_2(x)}\end{aligned}$$

Now, taking summation we get

$$\begin{aligned}&\sum_x \{\lambda p_1(x) + \mu p_2(x)\} \log \frac{\lambda p_1(x) + \mu p_2(x)}{\lambda q_1(x) + \mu q_2(x)} \leq \sum_x \lambda p_1(x) \log \frac{p_1(x)}{q_1(x)} + \sum_x \mu p_2(x) \log \frac{p_2(x)}{q_2(x)} \\ &\Rightarrow D\Big((\lambda p_1 + \mu p_2)||(\lambda q_1 + \mu q_2)\Big) \leq \lambda D(p_1||q_1) + \mu D(p_2||q_2) \\ &i.e.,\ \ D(p||q) \ \text{ is convex in } \ (p, q).\end{aligned}$$

□

Theorem 6.1.3. The entropy function $H(p)$ is a concave function of p.

Proof. The entropy function $H(p)$ is defined by

$$H(p) = -\sum_{i=1}^{n} p_i \log_D p_i$$

$$\begin{aligned}\text{Now, } \frac{\partial H}{\partial p_i} &= -\{1 + \log_e p_i\} \log_D e \\ \frac{\partial^2 H}{\partial p_i^2} &= -\frac{1}{p_i} \log_D e \\ \frac{\partial^2 H}{\partial p_i \partial p_j} &= 0, \quad i \neq j\end{aligned}$$

The Hessian matrix is given by

$$\nabla^2 H(p) = \begin{bmatrix} \frac{\partial^2 H}{\partial p_1^2} & \frac{\partial^2 H}{\partial p_1 \partial p_2} & \cdots & \frac{\partial^2 H}{\partial p_1 \partial p_n} \\ \frac{\partial^2 H}{\partial p_2 \partial p_1} & \frac{\partial^2 H}{\partial p_2^2} & \cdots & \frac{\partial^2 H}{\partial p_2 \partial p_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 H}{\partial p_n \partial p_1} & \frac{\partial^2 H}{\partial p_n \partial p_2} & \cdots & \frac{\partial^2 H}{\partial p_n^2} \end{bmatrix} = \begin{bmatrix} -\frac{1}{p_1} & 0 & \cdots & 0 \\ 0 & -\frac{1}{p_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -\frac{1}{p_n} \end{bmatrix} \log_D e$$

Clearly, $\nabla^2 H(p)$ is negative definite for $p_i > 0,\ \ (\because \log_D e > 0)$.

Hence, $H(p)$ is a concave function of p. □

Theorem 6.1.4. Non-negativity of conditional relative entropy

$$D\Big(p(y|x)||q(y|x)\Big) \geq 0.$$

Proof.

$$\begin{aligned}
\text{We have, } D(p(y|x)||q(y|x)) &= \sum_x \sum_y p(x,y) \log \frac{p(y|x)}{q(y|x)} \\
&= \sum_x \sum_y p(x,y) \log \frac{p(x,y)\ q(x)}{q(x,y)\ p(x)} \\
\text{Now, } \sum_x \sum_y p(x,y) \cdot q(x) &= \sum_x \sum_y q(x,y)\ p(x) = 1
\end{aligned}$$

$$\therefore \text{By Theorem 5.2.1,} \quad D(p(y|x)||q(y|x)) \geq 0.$$

□

Example 6.1.5. In a certain community, 25% of all girls are blondes, and 75% of all blondes are blue eyed. Also, 50% of all girls in the community have blue eyes. If you know that a girl has blue eyes, how much additional information do you being informed that she is blond?

Solution. Let p_1 = probability of a girl being blonde $= 0.25$.

p_2 = probability of a girl to have blue eyes if she is blonde = p_{blonde} (blue eyes) = 0.75

$p_3 = p$ (blue eyes) = 0.50

$p_4 = p$ (blonde, blue eyes) = probability that a girl is blonde and has blue eyes

and $p_x = p_{\text{ blue eyes}}$ (blonde) = probability that a blue eyed girl is blonde = ?

Then

$$p_4 = p_1 p_2 = p_3 p_x \Rightarrow p_x = \frac{p_1 p_2}{p_3} = \frac{0.25 \times 0.75}{0.50}$$

If a girl has blue eyes, the additional information obtained by being informed that she is blonde is

$$\begin{aligned}
\log_2 \frac{1}{p_x} &= \log_2 \frac{p_3}{p_1 p_2} \\
&= \log_2 p_3 - \log_2 p_1 - \log_2 p_2 \\
&= \log_2 \frac{1}{2} - \log_2 \frac{1}{4} - \log_2 \frac{3}{4} \\
&= \log_2 4 + \log_2 \frac{4}{3} - \log_2 2 \\
&= 1.41503 \\
&\approx 1.42 \text{ bits}
\end{aligned}$$

■

Example 6.1.6. Evaluate the average uncertainty associated with the probability of events A, B, C, D with probability of events $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}$ respectively.

Solution.

$$\begin{aligned}\text{We have, } H\left(\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{8}\right) &= -\frac{1}{2}\log\frac{1}{2}-\frac{1}{4}\log\frac{1}{4}-\frac{1}{8}\log\frac{1}{8}-\frac{1}{8}\log\frac{1}{8}\\ &= \frac{1}{2}\log_2 2+\frac{1}{2}\log_2 2+\frac{3}{4}\log_2 2\\ &= \left(\frac{1}{2}+\frac{1}{2}+\frac{3}{4}\right)\log_2 2\\ &= \frac{7}{4}\text{ bits}\end{aligned}$$

which is the averaged uncertainty associated with the probability of events A, B, C, D. ■

Example 6.1.7. A transmitter has an alphabet consisting of 5 letters $\{x_1, x_2, x_3, x_4, x_5\}$ and the receiver has an alphabet consisting of 4 letters $\{y_1, y_2, y_3, y_4\}$. The joint probabilities for communication are given below

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array}\begin{array}{c} \begin{array}{cccc} y_1 & y_2 & y_3 & y_4 \end{array} \\ \begin{pmatrix} 0.25 & 0.00 & 0.00 & 0.00 \\ 0.10 & 0.30 & 0.00 & 0.00 \\ 0.00 & 0.05 & 0.10 & 0.00 \\ 0.00 & 0.00 & 0.05 & 0.10 \\ 0.00 & 0.00 & 0.05 & 0.00 \end{pmatrix} \end{array}$$

Determine the marginal, conditional and joint entropies for this channel. (Assume $0\log 0 \equiv 0$)

Solution. The channel is described here by joint probabilities $p_{ij},\ i = 1,2,3,4,5;\ j = 1,2,3,4$. Then the conditional and marginal probabilities are easily obtained from p_{ij}'s as follows:

$$\begin{aligned} p_{10} &= 0.25+0.00+0.00+0.00 = 0.25\\ p_{20} &= 0.10+0.30+0.00+0.00 = 0.40\\ p_{30} &= 0.00+0.05+0.10+0.00 = 0.15\\ p_{40} &= 0.00+0.00+0.05+0.10 = 0.15\\ p_{50} &= 0.00+0.00+0.05+0.00 = 0.05\\ p_{01} &= 0.25+0.10+0.00+0.00+0.00 = 0.35\\ p_{02} &= 0.00+0.30+0.05+0.00+0.00 = 0.35\\ p_{03} &= 0.00+0.00+0.10+0.05+0.05 = 0.20\\ p_{04} &= 0.00+0.00+0.00+0.10+0.00 = 0.10\end{aligned}$$

By using the result, $p_{j|i} = \dfrac{p_{ij}}{p_{i0}}$, the conditional probabilities are given in the following channel matrix

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array}\begin{array}{c} \begin{array}{cccc} y_1 & y_2 & y_3 & y_4 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{3}{4} & 0 & 0 \\ 0 & \frac{1}{3} & \frac{2}{3} & 0 \\ 0 & 0 & \frac{1}{3} & \frac{2}{3} \\ 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

Marginal entropies:

$$\begin{aligned}
\therefore H(X) &= -\sum_{i=1}^{5} p_{i0} \log_2 p_{i0} \\
&= -(0.25 \log_2 0.25 + 0.40 \log_2 0.40 + \ldots + 0.05 \log_2 0.05) \\
&= 1.326 \text{ bits} \\
\therefore H(Y) &= -\sum_{j=1}^{4} p_{0j} \log_2 p_{0j} \\
&= 1.8556 \text{ bits}
\end{aligned}$$

Conditional entropies

$$\begin{aligned}
H(Y|X) &= -\sum_{i=1}^{5} \sum_{j=1}^{4} p_{ij} \log_2 p_{j|i} = 0.6 \text{ bits} \\
\text{Similarly, } H(X|Y) &= H(X) + H(Y|X) - H(Y) \\
&= 1.3260 + 0.6 - 1.8336 = 0.0704 \text{ bits}
\end{aligned}$$

Joint Entropy

$$\begin{aligned}
H(X,Y) &= H(X) + H(Y|X) \\
&= 1.3260 + 0.6 \\
&= 1.9260 \text{ bits}
\end{aligned}$$

■

6.2 Channel Capacity

Definition 6.2.1. Mutual information $I(X;Y)$ indicates a measure of the average information per symbol transmitted in the system. According to Shannon, in a discrete communication system, the channel capacity is the maximum of the mutual information, i.e.,

$$C = \max I(X;Y) = \max\{H(X) - H(X|Y)\}$$

For noise free channel, $I(X;Y) = H(X) = H(Y) = H(X,Y)$. Thus

$$C = \max I(X;Y) = \max\{H(X)\} = \max\left\{-\sum_{i=1}^{n} p_i \log p_i\right\}$$

Since $\max\{H(X)\}$ occurs when all symbols have equal probabilities, hence the channel capacity for a noise free channel is

$$C = -\log\left(\frac{1}{n}\right) = \log_2 n \text{ bits/symbol.}$$

6.3 Redundancy

i)

$$\begin{aligned}
\text{Absolute redundancy} &= C - I(X;Y) \\
&= C - H(X) \\
&= \log n - H(X) \quad \text{(For noise free channel)}
\end{aligned}$$

ii)

$$\begin{aligned}\text{Relative redundancy} &= \frac{C - I(X;Y)}{C} \\ &= \frac{\log n - H(X)}{\log n} = 1 - \frac{H(X)}{\log n}\end{aligned}$$

iii)

$$\begin{aligned}\text{Efficiency of a noise free system} &= \frac{H(X)}{\log n} \\ &= 1 - \text{Relative redundancy.}\end{aligned}$$

Example 6.3.1. Find the capacity of the memory less channel specified by the channel matrix

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 1 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \end{bmatrix}$$

Solution. The capacity of the memoryless channel is given by

$$\begin{aligned} C &= \max I(X,Y) \\ &= \max\{H(X) + H(Y) - H(X,Y)\} \\ &= -\sum_{i=1}^{4} p_{ij} \log p_{ij}, \quad j = 1,2,3,4 \\ &= -\sum_{i=1}^{4} p_{i1} \log p_{i1} - \sum_{i=1}^{4} p_{i2} \log p_{i2} - \sum_{i=1}^{4} p_{i3} \log p_{i3} - \sum_{i=1}^{4} p_{i4} \log p_{i4} \end{aligned}$$

where

$$\begin{aligned} p_{i1} &= \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}, 0\right) \\ p_{i2} &= \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) \\ p_{i3} &= (0,0,1,0) \\ p_{i1} &= \left(\frac{1}{2}, 0, 0, \frac{1}{2}\right) \end{aligned}$$

$$\begin{aligned} \text{Thus, } C &= \frac{1}{2}\log_2 \frac{1}{2} + 2\left(\frac{1}{4}\log_2 \frac{1}{4}\right) + 4\left(\frac{1}{4}\log_2 \frac{1}{4}\right) + 1\log_2 1 + 2\left(\frac{1}{2}\log_2 \frac{1}{2}\right) \\ &= \frac{3}{2}\log_2 2 + 3\log_2 2 \\ &= \frac{9}{2} \text{ bits/symbol} \end{aligned}$$

■

Example 6.3.2. Show that the entropy of the following probability distribution is $2 - \left(\frac{1}{2}\right)^{n-2}$.

Events	x_1	x_2	$\ldots$	x_i	$\ldots$	x_{n-1}	x_n	x_{n+1}
Probabilities	$\frac{1}{2}$	$\frac{1}{2^2}$	$\ldots$	$\frac{1}{2^i}$	$\ldots$	$\frac{1}{2^{n-1}}$	$\frac{1}{2^{n-1}}$	$\frac{1}{2^n}$

Solution. From the given data of the problem, we have

$$\begin{aligned} p_i &= \frac{1}{2^i},\ \ i = 1, 2, \ldots, n-1 \ \text{ and } \ p_n = \frac{1}{2^{n-1}} \\ \text{and} \quad \sum_{i=1}^{n} p_i &= \left[\frac{1}{2} + \frac{1}{2^2} + \ldots + \frac{1}{2^{n-1}}\right] + \frac{1}{2^{n-1}} \\ &= \frac{1}{2}\frac{1 - \frac{1}{2^{n-1}}}{1 - \frac{1}{2}} + \frac{1}{2^{n-1}} \\ &= 1 - \frac{1}{2^{n-1}} + \frac{1}{2^{n-1}} \\ &= 1 \end{aligned}$$

The entropy function H is defined as

$$\begin{aligned} H(p_1, p_2, \ldots, p_n) &= -\sum_{i=1}^{n} p_i \log p_i \\ \Rightarrow H(p_1, p_2, \ldots, p_n) &= -\sum_{i=1}^{n-1} p_i \log p_i - p_n \log p_n \\ \Rightarrow H(p_1, p_2, \ldots, p_n) &= -\sum_{i=1}^{n-1} \left(\frac{1}{2^i}\right) \log_2 \left(\frac{1}{2^i}\right) - \frac{1}{2^{n-1}} \log_2 \left(\frac{1}{2^{n-1}}\right) \\ \Rightarrow H(p_1, p_2, \ldots, p_n) &= \sum_{i=1}^{n-1} \left(\frac{1}{2^i}\right) \log_2 (2^i) + \frac{1}{2^{n-1}} \log_2 (2^{n-1}) \\ \Rightarrow H(p_1, p_2, \ldots, p_n) &= \sum_{i=1}^{n-1} i \cdot \frac{1}{2^i} + (n-1)\frac{1}{2^{n-1}} \\ \Rightarrow H(p_1, p_2, \ldots, p_n) &= \left\{\frac{1}{2} + \frac{2}{2^2} + \frac{3}{2^3} + \ldots + \frac{n-1}{2^{n-1}}\right\} + \frac{n-1}{2^{n-1}} \qquad (6.3.1) \\ \Rightarrow \frac{1}{2} H(p_1, p_2, \ldots, p_n) &= \left\{\frac{1}{2^2} + \frac{2}{2^3} + \frac{3}{2^4} + \ldots + \frac{n-1}{2^n}\right\} + \frac{n-1}{2^n} \qquad (6.3.2) \end{aligned}$$

Subtracting (6.3.2) from (6.3.1) we get,

$$\begin{aligned}
\frac{1}{2}H(p_1,p_2,\ldots,p_n) &= \left(\frac{1}{2}-\frac{1}{2^2}\right)+\left(\frac{2}{2^2}-\frac{2}{2^3}\right)+\left(\frac{3}{2^3}-\frac{3}{2^4}\right)+\ldots \\
&+ \left(\frac{n-1}{2^{n-1}}-\frac{n-1}{2^n}\right)+\left(\frac{n-1}{2^{n-1}}-\frac{n-1}{2^n}\right) \\
&= \frac{1}{2}+\left(\frac{2}{2^2}-\frac{1}{2^2}\right)+\left(\frac{3}{2^3}-\frac{2}{2^3}\right)+\left(\frac{4}{2^4}-\frac{3}{2^4}\right)+\ldots \\
&+ \left(\frac{n-1}{2^{n-1}}-\frac{n-2}{2^{n-1}}\right)-\frac{n-1}{2^n}+\frac{n-1}{2^n} \\
&= \frac{1}{2}+\frac{1}{2^2}+\frac{1}{2^3}+\ldots++\frac{1}{2^{n-1}} \\
&= 1-\left(\frac{1}{2}\right)^{n-1}
\end{aligned}$$

$$\therefore H(p_1,p_2,\ldots,p_n) = 2-\left(\frac{1}{2}\right)^{n-2}$$

■

Example 6.3.3. If the probability distribution $P=\{p_1,p_2,\ldots\},\ p_i\geq 0, \sum_{i=1}^{\infty} p_i = 1$ is such that the entropy function, $H(P) = -\sum_{i=1}^{\infty} p_i \log p_i < \infty$, then show that $\sum_{i=1}^{\infty} p_i \log i < \infty$.

Solution. Let us assume that $\{p_i\}$ are decreasing in i, which is quite possible because reordering of the $\{p_i\}$ does not affect the value of entropy. Then

$$1 = \sum_{j=1}^{\infty} p_j \geq \sum_{j=1}^{i} p_j \geq i p_i$$

Thus we have $-\log p_i > \log i$ and consequently

$$\sum_{i=1}^{\infty} p_i \log i \leq -\sum_{i=1}^{\infty} p_i \log p_i = H(P) < \infty.$$

Hence, $\sum_{i=1}^{\infty} p_i \log i < \infty$.

■

The following example is similar.

Example 6.3.4. If the probability distribution $\Phi=(p_1,p_2,\ldots),\ p_i\geq 0,\ \sum_{i=1}^{\infty} p_i = 1$ is such that $\sum_{i=1}^{\infty} p_i \log i < \infty$, then show that $H(\Phi) = -\sum_{i=1}^{\infty} p_i \log p_i < \infty$.

Example 6.3.5. Let H be the entropy of the probability distribution $p_1, p_2, \ldots, p_n$. If H_1 be the entropy of the probability distribution $p_1 + p_2, p_3, \ldots, p_n$, then show that

$$H - H_1 = P_s H_s \text{ where } P_s = p_1 + p_2 \text{ and } H_s = \left[\frac{p_1}{P_s}\log\frac{P_s}{p_1} + \frac{p_2}{P_s}\log\frac{P_s}{p_2}\right]$$

Solution. We have

$$\begin{aligned} H &= -p_1\log p_1 - p_2\log p_2 - p_3\log p_3 \ldots - p_n\log p_n \qquad (6.3.3)\\ H_1 &= -(p_1+p_2)\log(p_1+p_2) - p_3\log p_3 - \ldots - p_n\log p_n \\ &= -P_s\log P_s - p_3\log p_3 - \ldots - p_n\log p_n \qquad (6.3.4)\end{aligned}$$

Subtracting (6.3.4) from (6.3.3), we get

$$\begin{aligned} H - H_1 &= -p_1\log p_1 - p_2\log p_2 + P_s\log P_s \\ &= P_s\cdot\frac{1}{P_s}\Big[-p_1\log p_1 - p_2\log p_2 + P_s\log P_s\Big] \\ &= P_s\left[-\frac{p_1}{P_s}\log p_1 - \frac{p_2}{P_s}\log p_2 + \frac{p_1+p_2}{P_s}\log P_s\right] \\ &= P_s\left[\frac{p_1}{P_s}\log P_s - \frac{p_1}{P_s}\log p_1 + \frac{p_2}{P_s}\log P_s - \frac{p_2}{P_s}\log p_2\right] \\ &= P_s\left[\frac{p_1}{P_s}\log\frac{P_s}{p_1} + \frac{p_2}{P_s}\log\frac{P_s}{p_2}\right] \\ &= H_s\,P_s\end{aligned}$$

where $P_s = p_1 + p_2$, $H_s = \left[\frac{p_1}{P_s}\log\frac{P_s}{p_1} + \frac{p_2}{P_s}\log\frac{P_s}{p_2}\right]$. ■

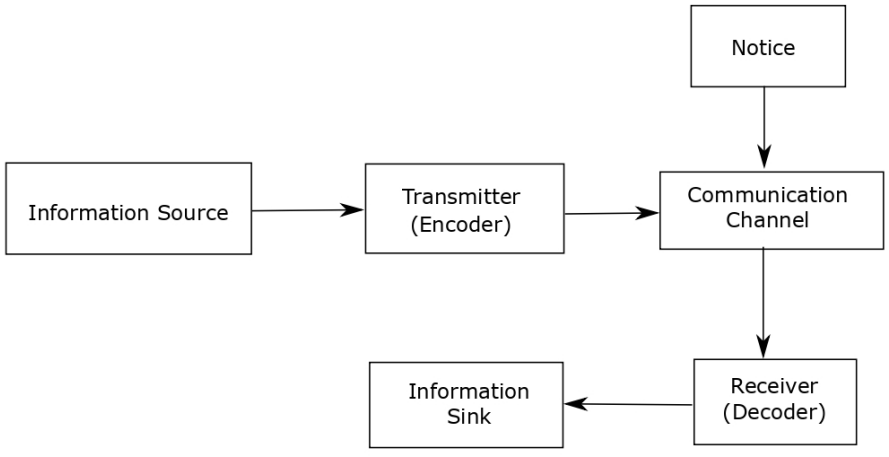
Unit 7

Course Structure

- Coding Theory
- Expected or average length of a code
- Uniquely decodable code

7.1 Introduction

Coding theory is the study of the method for efficient transfer of information from source; the physical medium through which the information transmitted for the channel, the telephone line and atmosphere are examples of channel. The undesirable disturbances are called noises. The following diagram provides a rough idea of the general information system:



Definition 7.1.1. Code: Let X be a random variable with range $S = \{x_1, x_2, \ldots, x_q\}$ and let $\mathfrak{D}$ be the D-ary alphabet, i.e., the set of all finite strings of symbols $\{0, 1, 2, \ldots, D-1\}$. A mapping $C : S \to \mathfrak{D}$ will be

called a code for the random variable X and S is called the source alphabet and $\mathfrak{D}$ is called the code alphabet.

If $x_i \in S$, then $C(x_i)$ is called codeword. Corresponding to x_i, the number of symbols in codeword $C(x_i)$ is called the length of the codeword and it is denoted by $l(x_i)$.

Example 7.1.2. Let X be a random variable with range $S = \{x_1, x_2, x_3, x_4\}$, $\mathfrak{D} = \{0, 1\}$ be the code alphabet. Define $C : S \to \mathfrak{D}$ as follows

$$x_1 \to 0, \quad x_2 \to 00, \quad x_3 \to 01, \quad x_4 \to 11$$

Then C is a code for the random variable X.

Definition 7.1.3. A code with code alphabet $\mathfrak{D} = \{0, 1\}$ is called a binary code. A code with code alphabet $\mathfrak{D} = \{0, 1, 2\}$ is called a ternary code.

Definition 7.1.4. A code C is said to be non-singular code if the mapping C is one-to-one, i.e., if $C(x_i) \neq C(x_j)$ for $x_i \neq x_j$. Clearly the code C in Example 7.1.2 is a non-singular code.

Definition 7.1.5. Extension of code: Let X be a random variable with range $S = \{x_1, x_2, \ldots, x_q\}$ and $\mathfrak{D} = \{0, 1, 2, \ldots, D-1\}$ as the code alphabet and C be a code for the random variable X. The n-th extension of C is a mapping $C^* : S^n (= S \times S \times \ldots \times S(n \text{ times})) \to \mathfrak{D}$ defined by

$$C^*(x_{i1}, x_{i2}, \ldots, x_{in}) = C(x_{i1})\, C(x_{i2})\, \ldots C(x_{in})$$

Example 7.1.6. Let X be a random variable with range $S = \{x_1, x_2, x_3, x_4\}$, $\mathfrak{D} = \{0, 1\}$ as the code alphabet and $C : S \to \mathfrak{D}$ be a code defined by

$$x_1 \to 0, \quad x_2 \to 00, \quad x_3 \to 01, \quad x_4 \to 11$$

Then the 2^{nd} extension of the above code C is given by

$$\begin{array}{llll} x_1x_1 \to 00, & x_1x_2 \to 000, & x_1x_3 \to 001, & x_1x_4 \to 001, \\ x_2x_1 \to 000, & x_2x_2 \to 0000, & x_2x_3 \to 0001, & x_2x_4 \to 0011, \\ x_3x_1 \to 010, & x_3x_2 \to 0100, & x_3x_3 \to 0101, & x_3x_4 \to 0111, \\ x_4x_1 \to 110, & x_4x_2 \to 1100, & x_4x_3 \to 1101, & x_4x_4 \to 1111. \end{array}$$

The 3^{rd} extension is

$$x_1x_2x_3 \to 000001; \quad x_1x_2x_4 \to 00011, \quad \ldots \quad \text{so on.}$$

7.1.1 Expected or average length of a code

Let X be a random variable with range $S = \{x_1, x_2, \ldots, x_q\}$ and p.m.f $p(x)$. Let $\mathfrak{D} = \{0, 1, 2, \ldots, D-1\}$ be the code alphabet. Then the expected length of the code C for the random variable X is denoted by $L(C)$ and is defined by

$$L(C) = \sum_{i=1}^{q} p(x_i) l(x_i) = \sum_{i=1}^{q} p_i l_i$$

7.1.2 Uniquely decodable (separable) code

A code is said to be uniquely decodable if all its extensions including itself are non-singular. For example, the code C in Example 7.1.6 is non-singular but its second extension is not singular. So it is not uniquely decodable ($\because x_1x_2 \neq x_2x_1$ but $C(x_1, x_2) = C(x_2, x_1) = 000$).

Examples of uniquely decodable codes are given below:

$$(a)\ \ x_1 \to 0, \quad x_2 \to 10, \quad x_3 \to 110, \quad x_4 \to 111$$
$$(b)\ \ x_1 \to 0, \quad x_2 \to 01, \quad x_3 \to 011, \quad x_4 \to 0111$$

Example 7.1.7. Let X be a random variable with range $S = \{x_1, x_2, x_3, x_4\}$ and code alphabet $\mathfrak{D} = \{0, 1\}$ with p.m.f $p(x)$ defined by

$$p(x_1) = \frac{1}{2}, \ \ p(x_2) = \frac{1}{4}, \ \ p(x_3) = \frac{1}{8} = p(x_4)$$

Let the code C be defined as follows:

$$x_1 \to 0, \quad x_2 \to 10, \quad x_3 \to 110, \quad x_4 \to 111$$
$$\therefore l(x_1) = 0, \quad l(x_2) = 2, \quad l(x_3) = 3, \quad l(x_4) = 3$$
$$\therefore \text{Expected length of } C, \ \ L(C) = 0 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} + 3 \cdot \frac{1}{8} = 1.25$$

Definition 7.1.8. Prefix: Let $i_1i_2 \ldots i_m$ be a codeword for some code C. Then $i_1i_2 \ldots i_\nu, \ \ \nu \leq m$ is called the prefix of the codeword $i_1i_2 \ldots i_m$. From definition it follows that every codeword is a prefix of itself.

Definition 7.1.9. Prefix code or instantaneous code: This is a code in which no codeword is a prefix of any other codeword. For example, the code in Example 7.1.7 is an instantaneous code whereas the code in Example 7.1.2 is not an instantaneous code. Another example of instantaneous code is the code defined by

$$x_1 \to 00, \quad x_2 \to 01, \quad x_3 \to 10, \quad x_4 \to 110$$

Theorem 7.1.10. An instantaneous code is uniquely decodable.

Proof. Let $S = \{x_1, x_2, \ldots, x_q\}$ be the source alphabet and $\mathfrak{D} = \{0, 1, 2, \ldots, D-1\}$ be the code alphabet for a random variable X.

Let $C : S \to \mathfrak{D}$ be an instantaneous code of the random variable X. The codewords are $C(x_1), \ C(x_2), \ \ldots, \ C(x_q)$. Since no codeword is a prefix of any other codeword, we have $C(x_i) \neq C(x_j)$ for $x_i \neq x_j$.

So C is one-to-one. Assuming C is not uniquely decodable, then there is a positive integer $n > 1$ such that $2^{nd}, \ 3^{rd}, \ \ldots, \ (n+1)^{th}$ extension of C are one to one. But the n^{th} extension is not one-to-one.

So, there are two elements

$$x = x_{i_1}x_{i_2} \ldots x_{i_n} \ \text{ and } \ y = y_{\nu_1}y_{\nu_2} \ldots y_{\nu_n} \ \text{ in } \ S \ \text{ such that } \ x \neq y \tag{7.1.1}$$

But

$$C^n(x) = C^n(y) \tag{7.1.2}$$

Write $x' = x_{i_2}x_{i_3} \ldots x_{i_n}$ and $y' = y_{\nu_2}y_{\nu_3} \ldots y_{\nu_n}$, then

$$x = x_{i_1}x', \quad y = y_{\nu_1}y'$$

$$\therefore \text{ We have } C^n(x) = C(x_{i_1})C(x_{i_2})\ldots C(x_{i_n}) = C(x_{i_1})C^{n-1}(x') \tag{7.1.3}$$

$$\text{Similarly, } C^n(y) = C(y_{\nu_1})C^{n-1}(y')$$

Without loss of generality, we may suppose that

$$l(x_{i_1}) \leq l(y_{\nu_1}) \tag{7.1.4}$$

where $l(x_{i_1})$ is the length of the codeword $C(x_{i_1})$ and $l(y_{\nu_1})$ be that of $C(y_{\mu_1})$. From (7.1.2), and (7.1.3) (7.1.4), it follows that the codeword $C(x_{i_1})$ is a prefix of the codeword $C(y_{\nu_1})$. Since C is an instantaneous code, it follows that

$$x_{i_1} = y_{\nu_1} \Rightarrow C^{n-1}(x') = C^{n-1}(y')$$

Since C^{n-1} is one-to-one, we have $x' = y'$. So, we have $x = y$ $[\because x_{i_1} = y_{\nu_1}]$ which contradicts (7.1.1).

Hence C is uniquely decodable code. □

Theorem 7.1.11. Kraft inequality for instantaneous code: Let $S = \{x_1, x_2, \ldots, x_q\}$ be the source alphabet and $\mathfrak{D} = \{0, 1, 2, \ldots, D-1\}$ be a code alphabet for a random variable X. Then a necessary and sufficient condition for the existence of an instantaneous code for the random variable X with codeword lengths $l_1, l_2, \ldots, l_q$ formed by the elements of $\mathfrak{D}$ is that

$$\sum_{i=1}^{q} D^{-l_i} \leq 1.$$

Proof. We first show that the condition is sufficient assuming that we have given codeword lengths $l_1, l_2, \ldots, l_q$ satisfying the condition

$$\sum_{i=1}^{q} D^{-l_i} \leq 1. \tag{7.1.5}$$

We show that there exists an instantaneous code for the random variable X with these codeword lengths. The lengths $l_1, l_2, \ldots, l_q$ may or may not be distinct. We shall find it useful to consider all codewords of the same length at a time.

Let $l = \max\{l_1, l_2, \ldots, l_q\}$. We denote by n_1, the number of codewords of length 1; by n_2, the number of codewords of length 2, and so on.

$$\therefore n_1 + n_2 + \ldots + n_l = q. \tag{7.1.6}$$

The inequality (7.1.5) may be written as

$$\sum_{i=1}^{l} n_i D^{-i} \leq 1 \tag{7.1.7}$$

Multiplying (7.1.7) by D^l, we have

$$\sum_{i=1}^{l} n_i D^{l-i} \leq D^l$$

$$n_l \leq D^l - n_1 D^{l-1} - n_2 D^{l-2} - \ldots - n_{l-1} D \tag{7.1.8}$$

From (7.1.8), we have,

$$n_{l-1} \leq D^{l-1} - n_1 D^{l-2} - n_2 D^{l-3} - \ldots - n_{l-2} D \tag{7.1.9}$$

Proceeding in this way we obtain,

$$\begin{aligned} n_{l-2} &\leq D^{l-2} - n_1 D^{l-3} - n_2 D^{l-4} - \ldots - n_{l-3} D \\ \cdots &\quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \\ n_3 &\leq D^3 - n_1 D^2 - n_2 D \\ n_2 &\leq D^2 - n_1 D \end{aligned} \tag{7.1.10}$$

We form n_1 codewords of length 1. Then there are $(D - n_1)$ unused codewords of length 1 which may be used as prefixes. By adding one symbol to the end of these permissible prefixes we may form as many as $(D - n_1)D = D^2 - n_1 D$ codewords of length 2. The inequalities (7.1.10) assures that we need no more than these number of (i.e., $D^2 - n_1 D$) codewords of length 2. As before, we chose n_2 codewords arbitrarily from $(D^2 - n_1 D)$ choices and we are left with $(D^2 - n_1 D - n_2)$ unused prefixes of length 2 with which we may form $(D^2 - n_1 D - n_2)D = D^3 - n_1 D^2 - n_2 D$ codewords of length 3. We select arbitrarily n_3 codewords from them and left with $D^3 - n_1 D^2 - n_2 D - n_3$ unused prefixes of length 3.

Continuing this process we obtain a code in which no codeword is prefix of any other codeword. So the code constructed is an instantaneous code.

We now show that the condition is necessary. Suppose that the codewords $C(x_1),\ C(x_2), \ldots, C(x_q)$ of lengths $l_1, l_2, \ldots, l_q$ for an instantaneous code for a random variable X.

Let $l = \max\{l_1, l_2, \ldots, l_q\}$ and let $n_i (i = 1, 2, \ldots, l)$ denote the number of codewords of length i.

There are all together D codewords of length 1 of which only n_1 codewords have been used. So $(D - n_1)$ codewords of length 1 are left unused. By adding one symbol to the end of these $(D - n_1)$ permissible prefixes we may form as $(D - n_1)D = D^2 - n_1 D$ codewords of length 2. Of these $(D^2 - n_1 D)$ codewords of length 2, n_2 are used.

$$\therefore n_1 \leq D, \quad n_2 \leq D^2 - n_1 D$$

Similarly,

$$\begin{aligned} & n_3 \leq D^3 - n_1 D^2 - n_2 D \\ & n_4 \leq D^4 - n_1 D^3 - n_2 D^2 - n_3 D \\ & \cdots \quad \cdots \quad \cdots \quad \cdots \\ & n_l \leq D^l - n_1 D^{l-1} - n_2 D^{l-2} - \ldots - n_{l-1} D \\ \Rightarrow\ & n_l + n_{l-1} D + n_{l-2} D^2 + \ldots + n_1 D^{l-1} \leq D^l \\ \Rightarrow\ & \sum_{i=1}^{l} n_i D^{-i} \leq 1 \\ \Rightarrow\ & \sum_{i=1}^{q} D^{-l_i} \leq 1 \end{aligned}$$

□

Definition 7.1.12. Optimal code: An instantaneous code is said to be optimal if the expected length of the code is less than or equal to the expected length of all other instantaneous codes for the same source alphabet and the same code alphabet.

Theorem 7.1.13. Let $S = \{x_1, x_2, \ldots, x_q\}$ be the source alphabet and $\mathfrak{D} = \{0, 1, 2, \ldots, D-1\}$ be the code alphabet for a random variable X. Then the expected length L^* of an optimal instantaneous code for the random variable X is given by

$$L^* = \frac{H(X)}{\log D},$$

where $H(X)$ is the entropy of the random variable X.

Theorem 7.1.14. Let $S = \{x_1, x_2, \ldots, x_q\}$ be the source alphabet and $\mathfrak{D} = \{0, 1, 2, \ldots, D-1\}$ be the code alphabet for the random variable X with p.m.f $p(X)$. Then the expected length $L(C)$ of any instantaneous code C for X satisfies the inequality

$$L(C) \geq \frac{H(X)}{\log D}.$$

Proof. Let $p_i = p(x_i) = P(X = x_i)$ and $l_i = l(x_i)$. Since C is an instantaneous code, by Kraft inequality

$$\sum_{i=1}^{q} D^{-l_i} \leq 1. \tag{7.1.11}$$

For any $x > 0$, we have

$$\log x \leq x - 1. \tag{7.1.12}$$

Write $\mu = \sum_{i=1}^{q} D^{-l_i}$, $0 < \mu \leq 1$. Taking $x = \frac{D^{-l_i}}{\mu p_i}$ in inequality (7.1.12), we get

$$\log \frac{D^{-l_i}}{\mu p_i} \leq \frac{D^{-l_i}}{\mu p_i} - 1$$

$$-l_i \log D - \log \mu - \log p_i \leq \frac{D^{-l_i}}{\mu p_i} - 1$$

Multiplying by p_i and taking sum we get

$$-\sum_{i=1}^{q} p_i l_i \log D - \sum_{i=1}^{q} p_i \log \mu - \sum_{i=1}^{q} p_i \log p_i \leq \sum_{i=1}^{q} \frac{D^{-l_i}}{\mu} - \sum_{i=1}^{q} p_i$$

$$\Rightarrow -L(C) \log D - \log \mu + H(X) \leq \frac{1}{\mu} \cdot \mu - 1$$

$$\Rightarrow H(X) - L(C) \log D \leq \log \mu \qquad \left[\because \mu = \sum_{i=1}^{q} D^{-l_i},\ \sum_{i=1}^{q} p_i = 1,\ L(C) = \sum_{i=1}^{q} p_i l_i\right]$$

$$\Rightarrow H(X) - L(C) \log D \leq 0 \quad [\because 0 < \mu \leq 1,\ \log \mu \leq 0]$$

$$\Rightarrow L(C) \geq \frac{H(X)}{\log D}$$

□

Theorem 7.1.15. Let L^* be the expected length of an instantaneous optimal code for the random variable X with code alphabet $\mathfrak{D} = \{0, 1, 2, \ldots, D-1\}$. Then

$$\frac{H(X)}{\log D} \leq L^* \leq \frac{H(X)}{\log D} + 1,$$

where $H(X)$ is the entropy function of the random variable X.

Proof. Let $S = \{x_1, x_2, \ldots, x_q\}$ be the source alphabet and $\mathfrak{D} = \{0, 1, \ldots, D-1\}$ be the code alphabet of the random variable X with p.m.f $p(x)$.

Let us define $p_i = p(x_i) = P(X = x_i),\ l_i = l(x_i)$. Now, L^* be the minimum value of $\sum_{i=1}^{q} p_i l_i$ subject to the constraint

$$\sum_{i=1}^{q} D^{-l_i} \leq 1 \tag{7.1.13}$$

We neglect the integer constraint on $l_1, l_2, \ldots, l_q$ and assume the inequality (7.1.13) hold.

The choice of the codeword length $l_i = -\dfrac{\log p_i}{\log D},\ \ (i = 1, 2, \ldots, q)$ gives

$$L = \sum_{i=1}^{q} p_i l_i = \sum_{i=1}^{q} \frac{-p_i \log p_i}{\log D} = \frac{H(X)}{\log D}.$$

$\because -\dfrac{\log p_i}{\log D}$ may not equal to an integer

Therefore, we round it upto the even integer. So we take $l_i = \left[-\frac{\log p_i}{\log D}\right]$, where for any real $x > 0$, $[x]$ denote the greatest positive integer not greater than x. Then

$$-\frac{\log p_i}{\log D} \leq l_i \leq -\frac{\log p_i}{\log D} + 1 \tag{7.1.14}$$

From (7.1.14), we have

$$\begin{aligned} &-\log p_i \leq l_i \log D \\ &\Rightarrow \log p_i \geq \log D^{-l_i} \\ &\Rightarrow p_i \geq D^{-l_i} \end{aligned}$$

Therefore,

$$\sum_{i=1}^{q} D^{-l_i} \leq \sum_{i=1}^{q} p_i = 1$$

Thus the codeword lengths $l_1, l_2, \ldots, l_q$ satisfies the Kraft inequality.
Hence the code with word lengths $l_1, l_2, \ldots, l_q$ as chosen is an instantaneous code.

Multiplying (7.1.14) by p_i and taking sum, we get

$$\begin{aligned} &-\sum_{i=1}^{q} \frac{p_i \log p_i}{\log D} \leq \sum_{i=1}^{q} p_i l_i \leq -\sum_{i=1}^{q} \frac{p_i \log p_i}{\log D} + \sum_{i=1}^{q} p_i \\ &\Rightarrow \frac{H(X)}{\log D} \leq L \leq \frac{H(X)}{\log D} + 1 \end{aligned} \tag{7.1.15}$$

Since L^* is the expected length of the optimal code, hence we have

$$\frac{H(X)}{\log D} \leq L^* \leq L \tag{7.1.16}$$

From (7.1.15) and (7.1.16) the result follows.

$$\therefore \frac{H(X)}{\log D} \leq L^* \leq \frac{H(X)}{\log D} + 1.$$

□

Example 7.1.16. Let $S = \{x_1, x_2, \ldots, x_q\}$ be the source alphabet and $\mathfrak{D} = \{0, 1, \ldots, D-1\}$ be the code alphabet of the random variable X with p.m.f $p(X_i) = D^{-\alpha_i}$, where $\alpha_1, \alpha_2, \ldots, \alpha_q$ are positive integers. Show that any code $C : S \to \mathfrak{D}$ for X with codeword lengths $\alpha_1, \alpha_2, \ldots, \alpha_q$ is an instantaneous optimal code.

Solution. Let C be any code for the random variable X with codeword lengths $l_i = l(x_i) = \alpha_i, \quad i = 1, 2, \ldots, q$. Then

$$\sum_{i=1}^{q} D^{-l_i} = \sum_{i=1}^{q} D^{-\alpha_i} = \sum_{i=1}^{q} p_i = 1$$

Thus the codeword lengths $l_1, l_2, \ldots, l_q$ of the code C satisfy Kraft inequality. Hence C is an instantaneous code. Again

$$\begin{aligned}
&p_i = D^{-\alpha_i} = D^{-l_i} \\
&\therefore \log p_i = -l_i \log D \\
&\Rightarrow -\sum_{i=1}^{q} p_i \log p_i = \sum_{i=1}^{q} l_i p_i \log D \\
&\Rightarrow H(X) = L(C) \log D \\
&\Rightarrow L(C) = \frac{H(X)}{\log D}
\end{aligned}$$

Therefore, the expected length $L(C)$ of the code C is minimum. Hence C is an instantaneous optimal code. ■

Definition 7.1.17. Efficiency of a code: Let C be a uniquely decodable D-ary code for the random variable X and $L(C)$ be its expected length. Then the efficiency η of the code C is defined by

$$\eta = \frac{H(X)}{L(C) \log D}$$

Redundancy of a code = $\beta = 1 - \eta$.

Theorem 7.1.18. Let C^* be a code of the random variable X of the following distribution

X :	x_1	x_2	...	x_n
p_i :	p_1	p_1	...	p_n

where $p_1 \geq p_2 \geq \ldots \geq p_n$. If $L(C^*) \leq L(C)$ for any code C of X, then $l_1^* \leq l_2^* \leq \ldots \leq l_n^*$ where l_i^* is the length of the code $C^*(x_i)$.

If $p_i = p_{i+1}$, it is assumed that $l_i^* \leq l_i^* + 1$.

Proof. Let $E = \{1, 2, \ldots, n\}$. We take any two elements i and j of E with $i < j$. Denote by α, the permutation of the set E such that $\alpha(i) = j$ and $\alpha(j) = i$ but all other elements of E remain unchanged.

Let C be a code of the random variable X such that $l_k = l^*_{\alpha(k)}$, where l_k is the length of the codeword $C(x_k)$. Then

$$\begin{aligned} l_i &= l^*_{\alpha(i)} = l^*_j \quad [\because \alpha(i) = j] \\ \text{and } l_j &= l^*_{\alpha(j)} = l^*_i \quad [\because \alpha(j) = i] \end{aligned}$$

and $l_k = l^*_k$ for all other elements k of E.

$$\begin{aligned} L(C) - L^*(C) &= p_i l_i + p_j l_j - p_i l^*_i - p_j l^*_j \\ &= p_i l^*_j + p_j l^*_i - p_i l^*_i - p_j l^*_j \\ &= (p_i - p_j)(l^*_j - l^*_i) \quad [\because l_i = l^*_j;\ l_j = l^*_i] \end{aligned}$$

Since $p_i \geq p_j$, we must have

$$\begin{aligned} l^*_i &\leq l^*_j. \\ \therefore l^*_1 &\leq l^*_2 \leq \ldots \leq l^*_n \end{aligned}$$

Hence the result follows. □

Unit 8

Course Structure

- Shannon-Fano Encoding Procedure for Binary code

8.1 Shannon-Fano Encoding Procedure for Binary code:

Let $S = \{x_1, x_2, \ldots, x_q\}$ be the source alphabet and $\mathfrak{D} = \{0, 1\}$ be the code alphabet of a random variable X with p.m.f $p(x)$. We shall give here an encoding procedure of assigning an efficient uniquely decodable binary code for the random variable X. This is known as Shannon-Fano encoding procedure.

Let $p_i = p(x_i) = P(X = x_i),\ i = 1, 2, \ldots, q.$

The two necessary requirements are

(i) No complete codeword can be prefix of some other codeword.

(ii) The binary digit in each codeword appeared independent with equal probabilities.

The encoding procedure follows the following steps.

Step 1: We arrange source symbols in descending order of their probabilities.

Step 2: Partition the set S of source symbols into two equiprobable groups S_0 and S_1 as

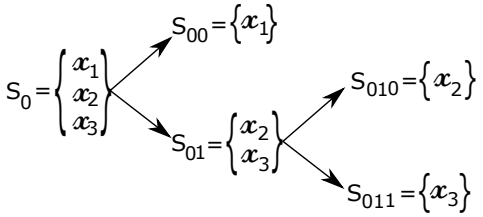
$$S_0 = \{x_1, x_2, \ldots, x_r\}, \quad S_1 = \{x_{r+1}, \ldots, x_q\}$$
$$\text{i.e., } P(S_0) \equiv P(S_1)$$

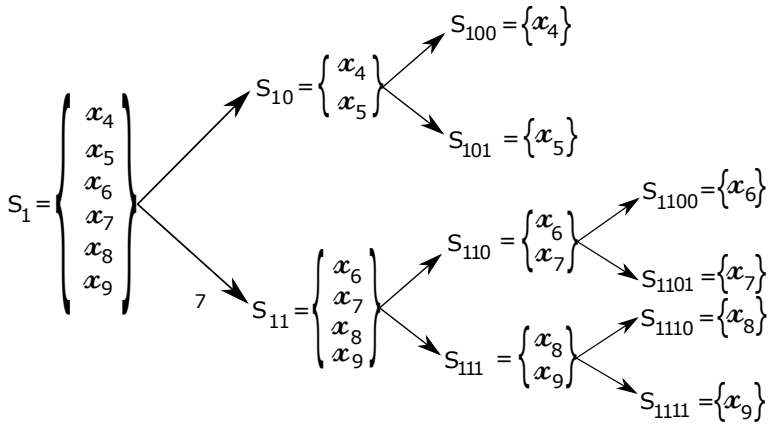
where $P(S_0) = p_1 + p_2 + \cdots + p_r$ and $P(S_1) = p_{r+1} + p_{r+2} + \cdots + p_q$.

Step 3: We further partition each of the subgroups S_0 and S_1 into two most equiprobable subgroups $S_{00},\ S_{01}$ and $S_{10},\ S_{11}$ respectively.

Step 4: We continue partitioning each of the resulting subgroups into two most equiprobable subgroups till each subgroup contain only one source symbol.

For example, let $S = \{x_1, x_2, \ldots, x_9\}$.





Therefore, the codes are

$$x_1 \to 00, \quad x_2 \to 010, \quad x_3 \to 011, \quad x_4 \to 100, \quad x_5 \to 101$$
$$x_6 \to 1100, \quad x_7 \to 1101, \quad x_8 \to 1110, \quad x_9 \to 1111$$

Clearly no codeword is a prefix of any other codeword. So it is an instantaneous code and hence it is uniquely decodable.

Advantages:

(1) Efficiency is nearly 100%.

(2) Expected length of the code is minimum.

(3) Entropy per digit of the encoded message is maximum.

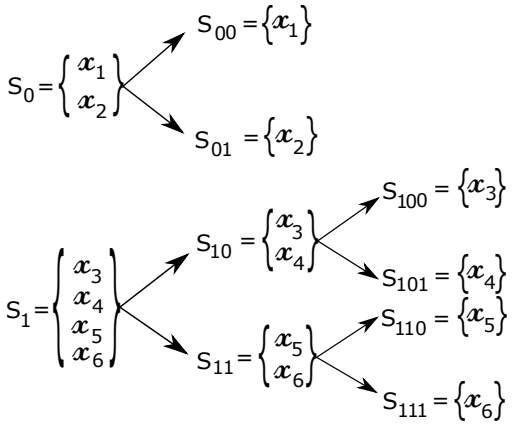
Example 8.1.1. Construct Shannon Fanno binary code for the random variable X with the following distribution.

Source symbols :	x_1	x_2	x_3	x_4	x_5	x_6
Probability :	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{12}$

Calculate the expected length and the efficiency of the code.

Solution. We have

$$p_1 + p_2 = \frac{1}{3} + \frac{1}{4} = \frac{7}{12}$$
$$p_3 + p_4 + p_5 + p_6 = \frac{1}{4} + \frac{1}{6} = \frac{5}{12}$$
$\therefore \frac{7}{12}$ and $\frac{5}{12}$ are close to each other
$\therefore$ We consider the equiprobable groups as follows.



So the Shannon-Fano binary code will be as follows:

$$x_1 \to 00, \quad x_2 \to 01, \quad x_3 \to 100, \quad x_4 \to 101, \quad x_5 \to 110, \quad x_6 \to 111$$

$$\begin{aligned} L(C) = \text{Expected length} &= 2 \cdot \frac{1}{3} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} + 3 \cdot \frac{1}{8} + 3 \cdot \frac{1}{12} + 3 \cdot \frac{1}{12} \\ &= \frac{2}{3} + \frac{1}{2} + \frac{3}{4} + \frac{1}{2} \\ &= 1 + \frac{2}{3} + \frac{3}{4} \\ &= \frac{12+8+9}{12} = \frac{29}{12} \text{ bits/symbol} \\ \text{Entropy, } H(X) &= -\sum_{i=1}^{6} p_i \log_2 p_i = 2.3758 \text{ bits} \\ \text{Efficiency, } \eta &= \frac{H(X)}{L(C)\log_2 2} = \frac{2.3758}{29/12} = 98.30\% \end{aligned}$$

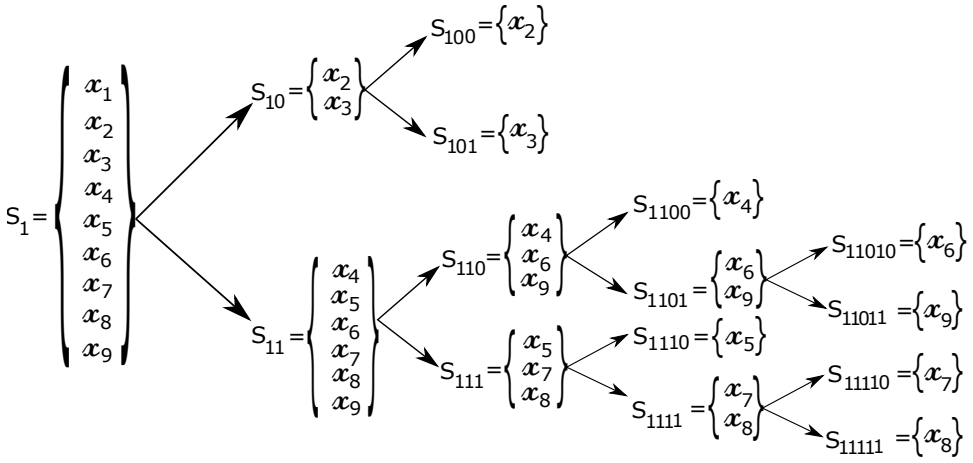
■

Similar Problems:

Example 8.1.2.

Source symbols :	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
Probability :	0.49	0.14	0.14	0.07	0.07	0.04	0.02	0.02	0.01

Solution. Here $p_1 = 0.49$ and $p_2 + \ldots + p_9 = 0.51$. Therefore, we take $S_0 = \{x_1\}$ and



So the code is

$$\begin{aligned} x_1 &\rightarrow 0 \\ x_2 &\rightarrow 100 \\ x_3 &\rightarrow 101 \\ x_4 &\rightarrow 1100 \\ x_5 &\rightarrow 1110 \\ x_6 &\rightarrow 11010 \\ x_7 &\rightarrow 11110 \\ x_8 &\rightarrow 11111 \\ x_9 &\rightarrow 11011 \end{aligned}$$

Therefore,

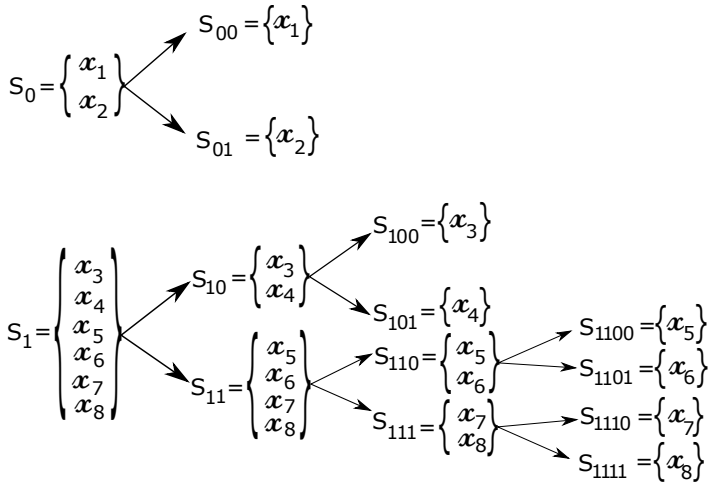
$$\begin{aligned} L(C) &= (1 \times 0.49) + (3 \times 0.14) + (3 \times 0.14) + (4 \times 0.07) + (4 \times 0.07) \\ &\quad +(5 \times 0.04) + (5 \times 0.02) + (5 \times 0.02) + (5 \times 0.01) \\ &= 2.34 \text{ bits/symbol.} \\ H(X) &= 2.3136 \text{ bits} \\ \therefore \eta &= \frac{2.3136}{2.34} = 38.87\% \end{aligned}$$

■

Example 8.1.3.

Source symbols :	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Probability :	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$

Solution. Here $p_1 + p_2 = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ and $p_3 + p_4 + p_5 + p_6 + p_7 + p_8 = \frac{1}{8} + \frac{1}{8} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16} = \frac{1}{2}$. Therefore, we take two equiprobable groups as:



So the code is

$$\begin{aligned} x_1 &\rightarrow 00 \\ x_2 &\rightarrow 01 \\ x_3 &\rightarrow 100 \\ x_4 &\rightarrow 101 \\ x_5 &\rightarrow 1100 \\ x_6 &\rightarrow 1101 \\ x_7 &\rightarrow 1110 \\ x_8 &\rightarrow 1111 \end{aligned}$$

Therefore,

$$\begin{aligned} L(C) &= \left(\frac{1}{4}\cdot 2\right) + \left(\frac{1}{4}\cdot 2\right) + \left(\frac{1}{8}\cdot 3\right) + \left(\frac{1}{16}\cdot 4\right) + \left(\frac{1}{16}\cdot 4\right) + \left(\frac{1}{16}\cdot 4\right) + \left(\frac{1}{16}\cdot 4\right) \\ &= 1 + \frac{3}{4} + 1 \\ &= \frac{11}{4} = 2.75 \text{ bits/symbol.} \\ H(X) &= -\sum_{i=1}^{8} p_i \log p_i = \frac{11}{4} = 2.75 \text{ bits} \end{aligned}$$

$$\therefore \text{Efficiency of the code} = \eta = \frac{H(X)}{L(C)\log_2 2} = \frac{2.75}{2.75} = 100\%$$

■

Unit 9

Course Structure

- Construction of Haffman binary code
- Construction of Haffman D-ary code

9.1 Construction of Haffman binary code

Let X be a random variable with the following distribution

$X:$	x_1	x_2	$\ldots$	x_n
Probability :	p_1	p_2	$\cdots$	p_n

Step 1: We arrange the source symbols $x_i's$ in decending order of their probabilities. Without loss of generality we may assume that $p_1 \geq p_2 \geq \ldots \geq p_n$. We thus have

$X:$	x_1	x_2	$\ldots$	x_n
Probability :	p_1	p_2	$\cdots$	p_n

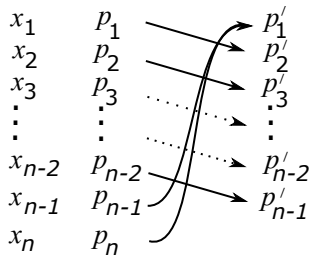
Step 2: We combine the last two symbols to form a new symbol. Then we arrange the source symbols in descending order of their probabilities. Let us suppose that

$$p_{n-1} + p_n \geq p_1.$$

We take

$$x'_1 = x_{n-1} + x_n, \quad x'_2 = x_1, \quad x'_3 = x_2, \quad x'_4 = x_3, \quad \ldots, \quad x'_{n-1} = x_{n-2}$$
$$p'_1 = p_{n-1} + p_n, \quad p'_2 = p_1, \quad p'_3 = p_2, \quad \ldots, \quad p'_{n-1} = p_{n-2}$$

This may be shown as follows:



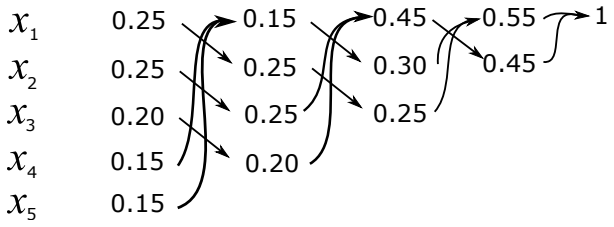
Step 3: Again we combine the last two symbols to form a new symbol and proceed as in Step 2.

Step 4: The process is continued until we reach a stage where we get only one symbol.

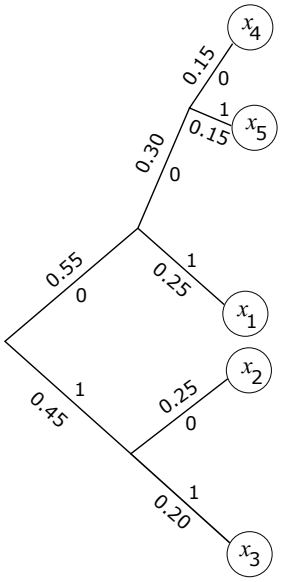
Example 9.1.1. Construct Haffman binary code for the random variable X whose distribution is given by

X :	x_1	x_2	x_3	x_4	x_5
Probability :	0.25	0.25	0.2	0.15	0.15

Solution. Consider the following scheme.



We arrange the above scheme as a tree in reverse order from which we can write down the corresponding Haffman binary code.



So the Haffman binary code is

$$\begin{aligned} x_1 &\rightarrow 01 \\ x_2 &\rightarrow 10 \\ x_3 &\rightarrow 11 \\ x_4 &\rightarrow 000 \\ x_5 &\rightarrow 001 \end{aligned}$$

■

9.2 Construction of Haffman D ary code (D>2)

Let the random variable X has the following distribution

$X:$	x_1	x_2	$\ldots$	x_q
Probability :	p_1	p_2	$\cdots$	p_q

Case 1: Let $(q-D)$ is divisible by $(D-1)$.

Step 1: Arrange the symbols in descending order of their probabilities.

Step 2: We consider last D symbols to a single composite symbol whose probability is equal to the sum of the probabilities of the last D symbols.

Step 3: Repeat Step 1 and Step 2 on the resulting set of symbols until we reach a stage where we get composite symbol only.

Step 4: Following above stage carefully we construct a tree diagram from which codes are assigned for the symbols.

Case 2: If $(q-D)$ is not divisible by $(D-1)$, then we add new *dummy symbols with zero probability* to make (q^*-D) divisible by $(D-1)$ where q^* is the number of symbols after addition of dummy symbols.

Now, we proceed as in Case 1. The codes for the dummy symbols are discarded.

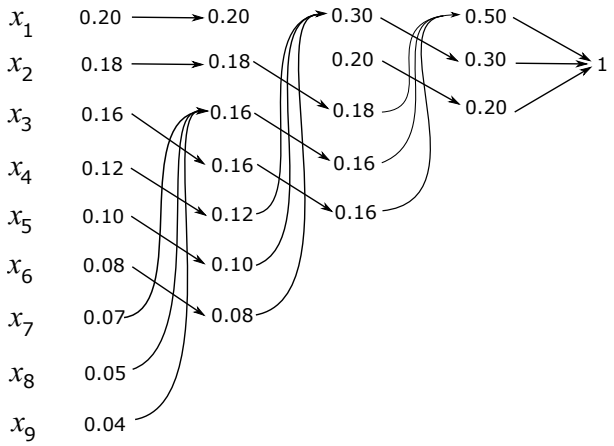
Example 9.2.1.

Source symbols :	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
Probability :	0.20	0.18	0.16	0.12	0.10	0.08	0.07	0.05	0.04

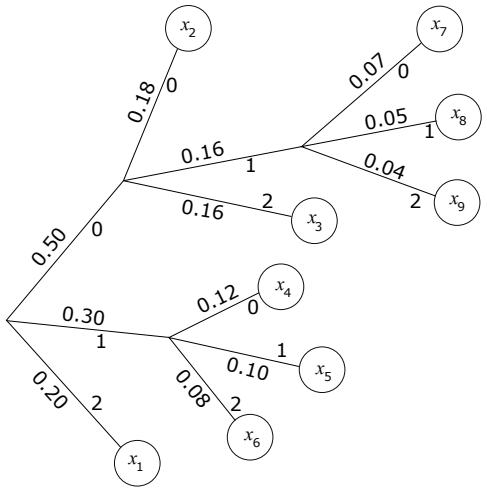
Construct a Haffman ternary code for X. Calculate the expected length and efficiency of the code.

Solution. Here $q=9$, $\mathfrak{D}=\{0,1,2\}$, $D=3$.

$$\therefore q-D=9-3=6 \text{ is divisible by } 2=3-1=D-1$$



We arrange the above scheme as a tree in reverse order from which we can write down the corresponding Haffman binary code.



So the code is

$$\begin{aligned} x_1 &\rightarrow 2 \\ x_2 &\rightarrow 00 \\ x_3 &\rightarrow 02 \\ x_4 &\rightarrow 10 \\ x_5 &\rightarrow 11 \\ x_6 &\rightarrow 12 \\ x_7 &\rightarrow 010 \\ x_8 &\rightarrow 011 \\ x_9 &\rightarrow 012 \end{aligned}$$

Therefore,

$$\begin{aligned} L(C) &= (1 \times 0.20) + (2 \times 0.18) + (2 \times 0.16) + (2 \times 0.12) + (2 \times 0.10) \\ &\quad +(2 \times 0.08) + (3 \times 0.07) + (3 \times 0.05) + (3 \times 0.04) \\ &= 1.96 \text{ bits/symbol.} \\ H(X) &= -\sum_{i=1}^{9} p_i \log p_i = 2.99388 \text{ bits} \end{aligned}$$

$$\therefore \text{Efficiency of the code} = \eta = \frac{H(X)}{L(C)\log_2 3} = 0.9637 = 96.37\%$$

■

Example 9.2.2. Construct Haffman ternary code with the following distribution

Source symbols :	x_1	x_2	x_3	x_4	x_5	x_6
Probability :	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{12}$

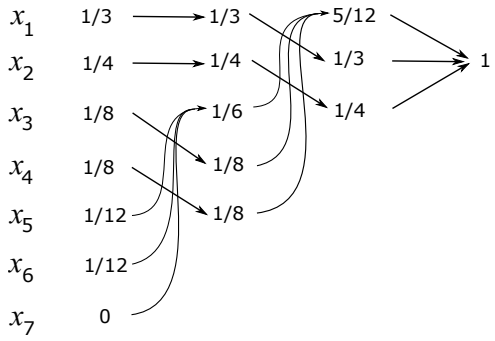
Calculate the expected length and its efficiency.

Solution. Here $q = 6,\ \mathfrak{D} = \{0, 1, 2\},\ D = 3.$

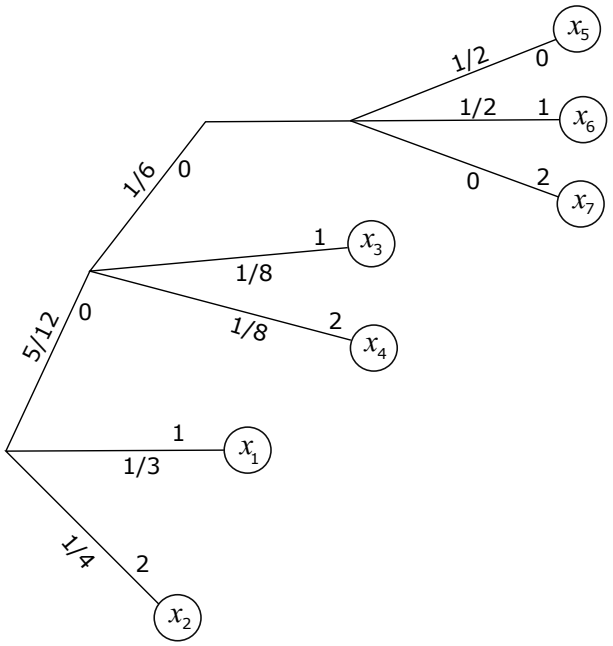
$$\therefore q - D = 6 - 3 = 3 \text{ which is not divisible by } 2. \tag{9.2.1}$$

Therefore, we introduce a dummy source alphabet x_7 with probability zero.

Now, $q^* = 7$, so $q^* - D = 4$ which is divisible by 2



We arrange the above scheme as a tree in reverse order from which we write down the Haffman ternary code.



So the code is

$$\begin{aligned}
x_1 &\rightarrow 1\\
x_2 &\rightarrow 2\\
x_3 &\rightarrow 01\\
x_4 &\rightarrow 02\\
x_5 &\rightarrow 000\\
x_6 &\rightarrow 001\\
x_7 &\rightarrow 002 \text{ (discarded)}.
\end{aligned}$$

Therefore,

$$\begin{aligned} L(C) &= \left(1\times\frac{1}{3}\right)+\left(1\times\frac{1}{4}\right)+\left(2\times\frac{1}{8}\right)+\left(2\times\frac{1}{8}\right)+\left(3\times\frac{1}{12}\right)+\left(3\times\frac{1}{12}\right) \\ &= \frac{19}{12} \text{ bits/symbol.} \\ H(X) &= -\sum_{i=1}^{6} p_i \log p_i = 1.4990 \text{ bits} \end{aligned}$$

$$\therefore \text{Efficiency of the code} = \eta = \frac{H(X)}{L(C)\log_2 3} = 0.9467 = 94.67\%$$

■

Example 9.2.3. Construct Shannon Fanno ternary code for the following distribution of the random variable X

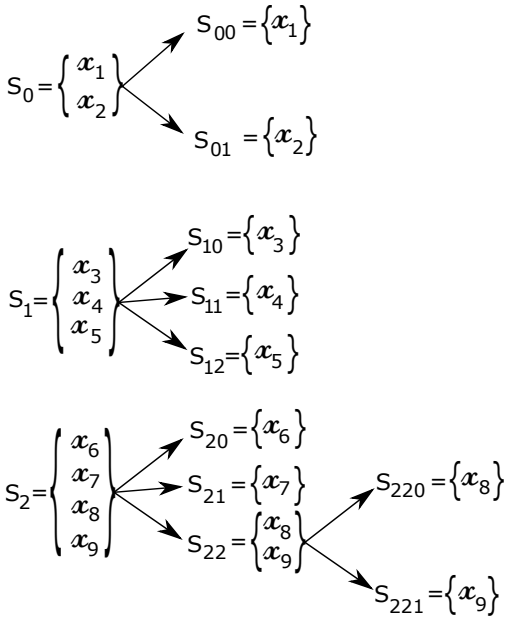
Source symbols :	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
Probability :	0.20	0.18	0.16	0.12	0.10	0.08	0.07	0.05	0.04

Hence calculate the expected length and efficiency of the code.

Solution: Here we see that

$$\begin{aligned} p_1+p_2 &= 0.38 \\ p_3+p_4+p_5 &= 0.38 \\ p_6+p_7+p_8+p_9 &= 0.24 \end{aligned}$$

So we take the three equiprobable groups as



Therefore, the Shannon Fanno ternary codes are obtained as

$$\begin{aligned} x_1 &\rightarrow 00 \\ x_2 &\rightarrow 01 \\ x_3 &\rightarrow 10 \\ x_4 &\rightarrow 11 \\ x_5 &\rightarrow 12 \\ x_6 &\rightarrow 20 \\ x_7 &\rightarrow 21 \\ x_8 &\rightarrow 220 \\ x_9 &\rightarrow 221 \end{aligned}$$

Therefore,

$$\begin{aligned} L(C) &= (2 \times 0.20) + (2 \times 0.18) + (2 \times 0.16) + (2 \times 0.12) + (2 \times 0.10) \\ &\quad +(2 \times 0.08) + (2 \times 0.07) + (3 \times 0.05) + (3 \times 0.04) \\ &= 2.09 \text{ bits/symbol.} \\ H(X) &= -\sum_{i=1}^{9} p_i \log p_i = 2.99388 \text{ bits} \end{aligned}$$

$$\therefore \text{Efficiency of the code} = \eta = \frac{H(X)}{L(C) \log_2 3} = 0.9038 = 90.38\%$$

Unit 10

Course Structure

- Error correcting codes
- Construction of linear codes
- Standard form of parity check matrix
- Hamming code, Cyclic code

10.1 Error correcting codes

Let F_q be a finite field with q elements and let $n(>1)$ be a given positive integer. We denote by $V_n(F_q)$, the set of all n-tuples $x = (x_1, x_2, \ldots, x_n)$ with $x_i \in F_q,\ i = 1, 2, \ldots, n$. For any $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n) \in V_n(F_q)$ and $\lambda \in F_q$, define

$$\begin{aligned} x + y &= (x_1 + y_1,\ x_2 + y_2,\ \ldots,\ x_n + y_n) \\ \lambda x &= (\lambda x_1, \lambda x_2, \ldots, \lambda x_n) \end{aligned}$$

Then

$$x + y,\ \ \lambda x \in V_n(F_q).$$

It is easy to see that $V_n(F_q)$ is a vector space over the field F_q.

Theorem 10.1.1. For any $x, y \in V_n(F_q)$, if we define $d(x, y) =$ number of i's with $x_i \neq y_i$, then d is a metric on $V_n(F_q)$.

Proof. From definition it is clear that for $x, y \in V_n(F_q)$

$$\begin{aligned} &(i)\ \ d(x, y) \geq 0 \ \text{ and } \ d(x, y) = 0 \ \text{ iff } \ x = y, \\ &(ii)\ \ d(x, y) = d(y, x). \end{aligned}$$

Now, let $x, y, z \in V_n(F_q)$. Then we show that

$$d(x, y) \leq d(x, z) + d(z, y) \tag{10.1.1}$$

If $x = y$, then $d(x, y) = 0$ and so (10.1.1) holds.
If $x = z$, then $d(x, z) = 0$ and $d(z, y) = d(x, y)$.
Hence (10.1.1) holds.
Similarly if $y = z$, then (10.1.1) also holds.
Suppose $x \neq y,\ x \neq z,\ z \neq y$.
Let $E = \{i :\ x_i \neq y_i\}$, $A = \{i : x_i \neq z_i\}$ and $B = \{i : y_i \neq z_i\}$.
$|E|$ denote the number of elements in E. Then $d(x, y) = |E|,\ d(x, z) = |A|,\ d(y, z) = |B|$.
Let $i \in E$. If $x_i \neq z_i$, then $i \in A$ also. Suppose that $x_i = z_i$. Since $x_i \neq y_i$, we have $z_i \neq y_i$. So $i \in B$.

$$\therefore i \in A \cup B.$$

This gives $E \subset A \cup B$. Therefore, $|E| \leq |A| + |B|$.

$$\therefore d(x, y) \leq d(x, z) + d(z, y)$$

Thus d is a metric on $V_n(F_q)$. □

Definition 10.1.2. q-ary code of length n: A non empty subset C of $V_n(F_q)$ is called a q-ary code of length n and members of C are called codeword. If $q = 2$, the corresponding code is called binary code and so on.

Definition 10.1.3. Weight of a codeword: An element x in $V_n(F_q)$ is a codeword. The weight of the codeword x, denoted by $w(x)$ and is defined by

$$w(x) = \text{number of } i's \text{ with } x_i \neq 0.$$
$$\text{e.g., } x = 1\,2\,0\,1\,0\,0\,\ldots\,0. \text{ Then, } w(x) = 3.$$

Definition 10.1.4. Linear code: A linear subspace C of $V_n(F_q)$ is called a linear code of length l over the field F_q and the dimension k of the subspace C is called the dimension of the code C. It is also called an (n, k) linear code over the field F_q.

Definition 10.1.5. Minimum distance of the code: Let C be a code in $V_n(F_q)$. The minimum distance $\delta(C)$ of the code C is defined by

$$\delta(C) = \min\{d(x, y) :\ x, y \in C \text{ and } x \neq y\}$$

Definition 10.1.6. Generator matrix: Let C be an (n, k) linear code over the field F_q with q elements. A $k \times n$ matrix G with entries from the field F_q is said to be the generator matrix of code C if the row space of the matrix G is the same as the subspace C. We also say that the matrix G generates the code C. Since the dimension of C is k, the dimension of the rowspace of G is k which implies that the row vectors of G are linearly independent and so they form a basis of C.

Definition 10.1.7. Parity check matrix: Let C be an (n, k) linear code over the field F_q with q elements. An $(n - k) \times n$ matrix H with entries from the field F_q is called a parity check matrix of code C iff $Hx = 0$ for all $x \in C$.

The matrix H also generates an $(n, n - k)$ linear code over F_q which is denoted by $C^{\perp}$ and is called the dual space of C.

$$\therefore \dim(C) + \dim(C^{\perp}) = n \quad \text{and} \quad \text{rank}(H) = n - k.$$

10.2 Construction of linear codes

• **By using generator matrix:** Let G be a $k \times n$ $(k < n)$ generator matrix with entries from F_q with q elements and rank$(G) = k$.

Let C denote the row space of the matrix G. Then C is and (n, k) linear code denoted by $\alpha_1\alpha_2 \ldots \alpha_k$, the row vectors of G.

Let $a = (a_1 \;\; a_2 \;\; \ldots \;\; a_k) \in V_k(F_q)$. Then

$$u = aG = a_1\alpha_1 + a_2\alpha_2 + \ldots + a_k\alpha_k \in C$$

Thus every u in C is of the form $u = aG$ where $a \in V_k(F_q)$.

Example 10.2.1. Find the codewords determined by the binary generator matrix

$$G = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

Solution. G is a binary generation matrix with 5 columns. Also, it is clear that rank(G)=3. The linear code C generated by G is given by

$$C = \{x : x = aG \text{ and } a \in V_3(F_2)\}$$

The vector $a = (a_1 \;\; a_2 \;\; a_3)$ may be considered in $2^3 = 8$ ways , namely, $(0\;\;0\;\;0)$, $(0\;\;0\;\;1)$, $(0\;\;1\;\;0)$, $(1\;\;0\;\;0)$, $(0\;\;1\;\;1)$, $(1\;\;0\;\;1)$, $(1\;\;1\;\;0)$, $(1\;\;1\;\;1)$.

$$\therefore \begin{pmatrix} 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\begin{pmatrix}0 & 0 & 1\end{pmatrix}\begin{pmatrix}1 & 0 & 0 & 1 & 1\\ 0 & 1 & 0 & 0 & 1\\ 0 & 0 & 1 & 1 & 1\end{pmatrix} = \begin{pmatrix}0 & 0 & 1 & 1 & 1\end{pmatrix}$$

$$\begin{pmatrix}0 & 1 & 0\end{pmatrix}\begin{pmatrix}1 & 0 & 0 & 1 & 1\\ 0 & 1 & 0 & 0 & 1\\ 0 & 0 & 1 & 1 & 1\end{pmatrix} = \begin{pmatrix}0 & 1 & 0 & 0 & 1\end{pmatrix}$$

$$\begin{pmatrix}1 & 0 & 0\end{pmatrix}\begin{pmatrix}1 & 0 & 0 & 1 & 1\\ 0 & 1 & 0 & 0 & 1\\ 0 & 0 & 1 & 1 & 1\end{pmatrix} = \begin{pmatrix}1 & 0 & 0 & 1 & 1\end{pmatrix}$$

$$\begin{pmatrix}0 & 1 & 1\end{pmatrix}\begin{pmatrix}1 & 0 & 0 & 1 & 1\\ 0 & 1 & 0 & 0 & 1\\ 0 & 0 & 1 & 1 & 1\end{pmatrix} = \begin{pmatrix}0 & 1 & 1 & 1 & 0\end{pmatrix} \quad [\because 1+1=0]$$

$$\begin{pmatrix}1 & 0 & 1\end{pmatrix}\begin{pmatrix}1 & 0 & 0 & 1 & 1\\ 0 & 1 & 0 & 0 & 1\\ 0 & 0 & 1 & 1 & 1\end{pmatrix} = \begin{pmatrix}1 & 0 & 1 & 0 & 0\end{pmatrix} \quad [\because 1+1=0]$$

$$\begin{pmatrix}1 & 1 & 0\end{pmatrix}\begin{pmatrix}1 & 0 & 0 & 1 & 1\\ 0 & 1 & 0 & 0 & 1\\ 0 & 0 & 1 & 1 & 1\end{pmatrix} = \begin{pmatrix}1 & 1 & 0 & 1 & 0\end{pmatrix} \quad [\because 1+1=0]$$

$$\begin{pmatrix}1 & 1 & 1\end{pmatrix}\begin{pmatrix}1 & 0 & 0 & 1 & 1\\ 0 & 1 & 0 & 0 & 1\\ 0 & 0 & 1 & 1 & 1\end{pmatrix} = \begin{pmatrix}1 & 1 & 1 & 0 & 1\end{pmatrix} \quad [\because 1+1=0;\ 1+1+1=0+1=1]$$

■

• **By using Parity check matrix:** Let H be an $r \times n (r < n)$ parity check matrix with entries from F_q with q elements and rank$(H) = r$. Let

$$C = \{x : x \in V_n(F_q) \text{ and } Hx = 0\}$$

Take $x, y \in C$ and any $\alpha \in F_q$, then we have

$$\begin{aligned} &Hx = 0, \quad Hy = 0.\\ &H(x+y) = H(x) + H(y) = 0\\ \text{and } &H(\alpha x) = \alpha H(x) = 0. \end{aligned}$$

Therefore, C is a linear subspace of $V_n(F_q)$ and so a linear code over the field F_q.

Clearly H is a parity check matrix for the code C. The dimension of code C is $n - r$.

Example 10.2.2. Find a codeword determined by the binary parity check matrix

$$H = \begin{pmatrix}1 & 0 & 1 & 0\\ 0 & 1 & 1 & 1\end{pmatrix}$$

Solution. Here H is a binary parity check matrix with 4 columns and rank$(H) = 2$. Therefore, the linear

code C determined by the parity check matrix H consists of binary codewords $(x_1\ x_2\ x_3\ x_4)$ satisfies

$$\begin{pmatrix}1&0&1&0\\0&1&1&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}=0$$

$$\Rightarrow x_1+x_3=0 \quad \text{and} \quad x_2+x_3+x_4=0$$

$$\Rightarrow x_1=x_3 \quad \text{and} \quad x_2=x_3+x_4 \ [\because 2\cdot 1=0;\ 1+1=0;\ -1=1]$$

If the values of x_3 and x_4 are assigned then x_1 and x_2 are determined. There are four ways of choosing x_3 and x_4 i.e., 00, 01, 10, 11, leading to the codewords 0000, 0101, 1110, 1011. ■

The following is a similar problem.

Example 10.2.3. Find the codewords determined by the P.C.M

$$H=\begin{pmatrix}1&0&0&0&1\\0&1&0&1&1\\0&0&1&1&1\end{pmatrix}$$

10.3 Standard form of parity check matrix:

The standard $r\times n$ parity check matrix H is given by

$$H=\begin{bmatrix}1&0&\dots&0&b_{11}&b_{12}&\dots&b_{1\overline{n-r}}\\0&1&\dots&0&b_{21}&b_{22}&\dots&b_{2\overline{n-r}}\\0&0&\dots&0&b_{31}&b_{32}&\dots&b_{3\overline{n-r}}\\\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots\\0&0&\dots&1&b_{r1}&b_{r2}&\dots&b_{r\overline{n-r}}\end{bmatrix}_{r\times n}$$

with entries from the field F_q with q elements.

$$\begin{aligned}x_1 &= b_{11}x_{r+1}+b_{12}x_{r+2}+\dots+b_{1\overline{n-r}}x_n\\x_2 &= b_{21}x_{r+1}+b_{22}x_{r+2}+\dots+b_{2\overline{n-r}}x_n\\ &\dots \quad\quad \dots \quad\quad \dots \quad\quad \dots\\x_r &= b_{r1}x_{r+1}+b_{r2}x_{r+2}+\dots+b_{r\overline{n-r}}x_n\end{aligned}$$

These equations determine $x_1,x_2,\dots,x_r$ when the values of $x_{r+1},\ x_{r+2},\dots,x_n$ are assigned, since there are q^{n-r} ways of choosing the values to obtain the linear code C of dimension $n-r$.

10.4 Hamming Code:

Let r be a given positive integer. We determine a binary matrix H with r rows and with maximum number of columns such that no column of H consist entirely of $0's$ and no two columns of H are same. Then linear code C determined by the parity check matrix H, we correct one error. This code C is called a Hamming code. Since each column of H has r entries and each entry is either 0 or 1, then the maximum number of different column is $n=2^r-1$. (The column consisting of entirely $0's$ being excluded.)

Exercise 10.4.1. Determine the Hamming code by the following P.C.M

$$H=\begin{pmatrix}1&0&0&1&1&0&1\\0&1&0&1&0&1&1\\0&0&1&0&1&1&1\end{pmatrix}$$

10.5 Cyclic Code

Here we shall denote the word "a" of length n by $a_0a_1a_2\ldots a_{n-1}$. The word $\hat{a} = a_{n-1}a_0a_1a_2\ldots a_{n-2}$ is called the 1st cyclic shift of the word a. A code C in $V_n(F_q)$ is said to be cyclic if it is linear and $a \in C \Rightarrow \hat{a} \in C$.

Let C be a cyclic code in $V_n(F_q)$ and $a \in C$. Then the words are obtained from a by n number of cyclic shifts. Any number of cyclic shifts such as

$$a_ia_{i+1}\cdots a_{n-1}a_0a_1\cdots a_{i-1}$$

belong to C.

Cyclic codes are useful for two reasons; from the practical point of view, it is possible to implement by simple devices known as shift resister. On the other hand, cyclic code can be constructed and investigated by means of algebraic theory of rings and polynomials.

Construction of a cyclic code

Let C be a cyclic code in $V_n(F_q)$ generated by $g(x)$. Then $g(x)$ is a divisor of $x^n - 1$. So we have

$$x^n - 1 = h(x)g(x) \tag{10.5.1}$$

Let

$$\begin{aligned} h(x) &= h_0 + h_1x + h_2x^2 + \ldots + h_kx^k \\ g(x) &= g_0 + g_1x + g_2x^2 + \ldots g_{n-k-1}x^{n-k-1} + g_{n-k}x^{n-k} \end{aligned}$$

where $g_{n-k} = 1$.

It is easy to see from (10.5.1) that $h_k = 1$ and $h_0g_0 = -1$, which gives that $h_0 \neq 0, g_0 \neq 0$. The polynomial $g(x)$ corresponds to the codeword

$$g = g_0\ g_1\ g_2\ \cdots\ g_{n-k}\ 0\ 0\ \cdots\ 0 \quad \text{in} \quad V_n(F_q)$$

The polynomial $x^ig(x)$ $(1 \leq i \leq k-1)$ corresponds to the codeword

$$g_{(i)} = 0\ 0\ \cdots\ g_0\ g_1\cdots\ g_{n-k}\ 0\ 0\ \cdots 0$$

There are i zeros at the beginning and $k-1-i$ zeros at the end. We denote by $\overline{h}$, the codeword whose 1st $k+1$ bits are $h_kh_{k-1}\cdots h_1h_0$ followed by $n-k-1$ zeros.

$$\therefore \overline{h} = h_k\ h_{k-1}\cdots\ h_1\ h_0\ 0\ 0\ \cdots 0$$

Let H denote the $(n-k)\times n$ matrix whose rows are $\overline{h},\ \overline{h}_{(1)},\ \overline{h}_{(2)}, \ldots, \overline{h}_{(n-k+1)}$, where $\overline{h}_{(i)}$ is the i-th cyclic shift of the codeword $\overline{h}$. Hence

$$H = \begin{bmatrix} h_k & h_{k-1} & \cdots & h_1 & h_0 & 0 & 0 & \cdots & 0 \\ 0 & h_k & h_{k-1} & \cdots & h_1 & h_0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & h_k & h_{k-1} & \cdots & h_1 & h_0 \end{bmatrix}$$

Example 10.5.1. Determine the binary parity check matrix for the cyclic code $C =< g(x) >$ of length 7 where $g(x) = 1 + x^2 + x^3$ and obtain the code C.

Solution. The factorization of $x^7 - 1$ into irreducible polynomials, i.e.,

$$\begin{aligned} x^7 - 1 &= (1+x)(1+x+x^3)(1+x^2+x^3) \\ &= h(x)g(x) \quad [\because \text{In a binary code, -1} = 1] \end{aligned} \tag{10.5.2}$$

$$\begin{aligned} \therefore h(x) &= (1+x)(1+x+x^3) \\ &= (1+x^2+x^3+x^4) \quad [\because 1+1=0] \end{aligned}$$

$$\therefore h_0 = 1, \quad h_1 = 0, \quad h_2 = 1, \quad h_3 = 1, \quad h_4 = 1.$$

$$H = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 \end{pmatrix}$$

No column of H consist entirely $0's$ and no two columns are exactly same. So the code determined by H is a Hamming code of length 7. ■

Unit 11

Course Structure

- Golay code, BCH codes, Reed-Muller code, Perfect code, codes and design.

11.1 Golay Code

Definition 11.1.1. (Binary) Code: A (binary) $[n, k, d]$ code is a k-dimensional subspace C of $\mathbb{F}_2^n$ with the property that any two distinct points in C have (Hamming) distance $\geq d$ (i.e. any two distinct points differ in at least d coordinates). We call elements of C codewords.

Note 11.1.2. If $u, v \in C$ have distance d then $0, v - u$ have distance d or equivalently $v - u$ has weight d (i.e. has d coordinates with value 1). Thus, the minimum distance between two distinct codewords is equal to the minimum weight of a nonzero codeword.

Example 11.1.3. Let V be the points of the Fano plane and let $C \subset \mathbb{F}_2^n$ consist of the vectors $0, 1$, the incidence vector of every line, and the complement of the incidence vector of every line. It is straightforward to check that C is a subspace so this is a $[7, 4, 3]$ code. This code can be generated by the rows of the following matrix.

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Error Correcting: If C has a distance $d \geq 2e + 1$ then the Hamming balls of radius e around each codeword are disjoint, so if a codeword was transmitted over a noisy channel causing at most e bitwise errors to occur, these could be reliably corrected.

Definition 11.1.4. We say that an an $[n, k, 2e + 1]$ code is perfect if the Hamming balls of radius e partition $\mathbb{F}_2^n$. In this case we must have

$$2^n = 2^k \sum_{i=0}^{e} \binom{n}{i}.$$

Note that the code in the example above is perfect as the Hamming ball of radius 1 around each point contains $1 + 7 = 2^3$ points and the code is a 4-dimensional subspace of $\mathbb{F}_2^7$.

11.1.1 The Golay Code

Let N be the generator matrix and define the matrix P as follows.

$$P = \begin{bmatrix} 0 & 1 & \dots & 1 \\ 1 & & & \\ \vdots & & N & \\ 1 & & & \end{bmatrix} \tag{11.1.1}$$

We define the Golay Code, G_{24}, to be the code generated by the rows of the matrix $[IP]$.

Observation 11.1.4.1. G_{24} is a $[24, 12, 8]$-code.

Proof. It is immediate from the properties of N that any two rows of the generator matrix have dot product 0, so $G_{24}^T = G_{24}$. Every row of the generator matrix has weight a multiple of 4 and it then follows from an easy inductive argument that every codeword of G_{24} has weight a multiple of 4. The sum of two rows of N has weight 6 and the sum of three or four rows of N is nonzero. It follows from this that G_{24} has no codeword of weight 4, so it is a $[24, 12, 8]$ code. □

Definition 11.1.5. M_{24}**:** We define the Matthieu Group, M_{24}, to be the subgroup of permutations of the 24 coordinates of G_{24} which map codewords to codewords.

Theorem 11.1.6. M_{24} acts 5-transitively on the coordinates of G_{24}.

Theorem 11.1.7. Let G act faithfully and 3-transitively on the set Ω. Then one of the following holds:

1. G contains all permutations of Ω or all even permutations of Ω;
2. This action is isomorphic to $AGL(n, 2)$ acting on $AG(n, 2)$
3. $|\Omega| = q + 1$ and this action contains the action of $PSL(2, q)$ on $PG(1, q)$
4. This action is the action of M_{12} on a set of size 12, or the actions obtained by fixing one or two points of this set.
5. This action is the action of M_{24} on a set of size 24, or the actions obtained by fixing one or two points of this set.

Note 11.1.8. The codewords of weight 8 form a $5 - (24, 8, 1)$ design.

Definition 11.1.9. G_{23}**:** We let G_{23} be the code obtained from G_{24} by deleting one coordinate. Then G_{23} is a $[23, 12, 7]$ code and since every codeword of G_{24} has even weight, we can recover G_{24} from G_{23} by adding a new bit to each codeword so that it has even weight. Note that the sum of the sizes of the Hamming balls of radius 3 around codewords of G_{23} is

$$2^{12} \sum_{i=0}^{3} = 2^{12}(1 + 23 + 253 + 1771) = 2^{12} \cdot 2^{11} = 2^{23}$$

so G_{23} is a perfect code.

Theorem 11.1.10. The only perfect $[n, k, 2e + 1]$ code with $k > 1$ and $e > 2$ is G_{23}.

Alternate Constructions of the Golay Code:

1. We construct G_{24} by taking M to be the 12×12 matrix which is the complement of the adjacency matrix of an icosahedron and then taking $[IM]$ as our generator matrix.
2. We can construct G_{24} by the following procedure: In the space $\mathbb{F}_2^{24}$ we order the words lexicographically, and at each step choose the smallest word of distance ≥ 8 to any already chosen word.
3. We can construct G_{23} by taking the rowspace of the (11-dimensional) matrix $M = \{m_{ij}\}_{i,j \in \mathbb{F}_{23}}$ given by
$$m_{ij} = \begin{cases} 1 \text{ if } i-j \in \mathbb{F}_{23} \\ 0 \text{ otherwise} \end{cases}.$$

There are many ways of making the Golay code(s). We'll describe just one. Adding an overall parity check to the perfect code gives one of length 24, in which the minimal weight is 8 instead of 7. This is the code I shall construct.

Put the 24 coordinates in a 6×4 array, with the 6 columns labelled by the coordinates $0, 1, 2, 3, 4, 5$ of the hexacode, and the 4 columns labelled by the four elements of $\mathbb{F}_4$. Now the 24 coordinates lie in $\mathbb{F}_2$ and satisfy 12 independent linear conditions, as follows:

- The parity of all the columns equals the parity of the top row. (6 conditions)
- The sums over each column give a hexacode word. Equivalently, these sums give a word which is perpendicular to all hexacode words. Equivalently, perpendicular to six hexacode words forming an $\mathbb{F}_2$-basis. (6 conditions)

In effect, the first column is arbitrary (16 choices), then the second and third columns have to have the same parity (8×8 choices), at which point the hexacode word is uniquely determined. Then the fourth and fifth columns are determined up to complementation (2×2 choices) and the last column is determined by the parity condition. In any case, the Golay code has 212 words.

It is linear because it is defined by linear conditions. It is also self-dual: this follows easily from the fact that the hexacode is self-dual. Or if you doubt this, check it on a basis instead:

- Take six vectors of shape one column plus (i.e. symmetric difference) the top row.
- Take six vectors of shape the top row plus a hexacode word (i.e. 6 such words forming an $\mathbb{F}_2$-basis of the hexacode).

The weight distribution of the Golay code is $0^1 8^{759} 12^{2576} 16^{759} 24^1$. To prove this, first observe that (1^{24}) is in the code. We can find the following words of weight 8:

- Two columns: 15 of these;
- One column plus a hexacode word: $6 \times 64 = 384$ of these;
- The top row plus a hexacode word of weight 4, plus an even number of these four columns: $45 \times 8 = 360$ of these.

The words of weight 16 are the complements of these, and we find the following words of weight 12:

- A hexacode word plus 3 columns: $64 \times 20 = 1280$ of these;
- The top row plus a hexacode word of weight 8, plus an even number of columns: $18 \times 32 = 576$ of these;

- The top row plus a hexacode word of weight 4, plus an even number of columns including one of the other two columns: $45 \times 8 \times 2 = 720$ of these.

Since we have already found 212 codewords, these are all.

The unique linear perfect 3-error-correcting code is obtained by deleting one coordinate from this. It still has dimension 12 of course, and weight distribution $0^1 7^{253} 8^{506} 11^{1288} 12^{1288} 15^{506} 16^{253} 23^1$.

Round each codeword we count

- 1 codeword
- 23 vectors at distance 1
- $23.22/2 = 253$ at distance 2
- $23.22.21/3.2.1 = 1771$ at distance 3

making $2048 = 2^{11}$ altogether, thereby neatly accounting for all $212 \times 211 = 223$ vectors in the space.

The extended code has the following numbers of vectors at various distances:

- 1 codeword
- 24 at distance 1
- $24.23/2 = 276$ at distance 2
- $24.23.22/3.2.1 = 2024$ at distance 3
- $24.23.22.21/4.3.2 = 10626$ at distance 4

In particular, 2325 cosets of the code contain representatives of weight at most 3, so the remaining 1771 each have 6 representatives of weight 4, since $6 \times 1771 = 10626$.

Sextets are the corresponding partitions of the 24 points into six 4s. For example the six columns of our diagram (Curtis's MOG) form such a sextet, since the sum of two columns lies in the code.

The stabiliser of a sextet permutes the six columns as S_6: an A_6 from the automorphism group of the hexacode, together with swapping the last two columns and simultaneously applying the field automorphism.

Fixing all the columns setwise, we still have the additive symmetry of the hexacode. Therefore the full stabiliser has shape $2^6 : 3S_6$.

The full automorphism group of the extended Golay code is transitive on the sextets (needs to be proved!), and so has order 244823040. It is the simple Mathieu group M_{24}.

11.2 BCH Code

11.2.1 Introduction

Multiple error correcting polynomial codes were invented by mathematicians Bose, Ray-Chaudhuri, and Hocquenghem in the 1950's. These codes are called BCH codes in their honor. Although BCH codes can be defined over any field, we will again, for simplicity, restrict to the binary field and study binary BCH codes.

11.2.2 The BCH Code

Denote messages, generators (encoding polynomials), codewords, and received mes- sages by $m(x)$, $p(x)$, $c(x)$, respectively. These are represented as sequences corresponding to the coefficients of a polynomial, where we take the convention of writing the coefficients from lowest to highest degree.

Recall from the last section that polynomial codes are obtained by multiplying message polynomials by encoding polynomials. Thus,

$$c(x) = m(x)p(x) = \sum_{i=1}^{n} c_i x^i \tag{11.2.1}$$

which is also represented by $c = (c_1, c_2, \ldots, c_n)$ for $c_i \in GF(2)$.

A binary BCH code is defined as follows.

Let $p(x)$ be a primitive polynomial of degree r with coefficients in the binary field. If $c(x)$ is a non-zero polynomial such that $c(x) = c(x^3) = c(x^5) = \ldots = c(x^{2t-1}) = 0 \pmod{p(x)}$ for all t such that $1 \leq (2t-1) \leq 2^r - 1$, then $c(x)$ is a t-error correcting code of length $n = 2^r - 1$.

A received sequence can have at most t errors to guarantee correct decoding. Let e_i, $1 \leq e_i \leq n$ and $1 \leq i \leq t$, represent the location of the ith bit in error. Then the error monomials are x^{e_i}, the error polynomial $e(x) = \sum_{i=1}^{k} x^{e_i}$ is their sum. The received polynomial is then

$$r(x) = c(x) + e(x). \tag{11.2.2}$$

Suppose we have a t-error correcting BCH code. Then the remainder of the received ploynomial $r(x) \pmod{p(x)}$ is equal to the sum of the remainders of the error monomials. Furthermore, if we evaluate the received polynomial at a higher power of x and then divide by $p(x)$, then this is equal to the error polynomial evaluated at the higher power and taking its remainder. That is, $\text{Rem}\big[r(x^j)\big] = \text{Rem}\big[e(x^j)\big]$ for $1 \leq j \leq 2t-1$. Recall that if $p(x)$ is primitive, there is a bijective map from $x^{e_i} \rightarrow \text{Rem}\big[e(x^j)\big]$ for all e_i, $1 \leq e_i \leq n$. Thus, if the error monomial or the remainder value after dividing the error monomial x^{e_i} by $p(x)$ is known, then the position of the error is known. For multiple errors, we only have the remainder of the sums of error monomials. To determine each bit position in error, we need to extract each monomial from this information.

In the following section, we show

(i) how to find the encoding polynomial and

(ii) how to determine the bit positions in error from the remainders of the received polynomial evaluated at higher powers.

11.2.3 The Generator Polynomial

For a t-error correcting code, the generator polynomial $Q(x)$ of standard BCH codes has the form

$$Q(x) = p(x)p_3(x)p_5(x)\ldots p_{2t-1}(x) \tag{11.2.3}$$

where $p(x)$ is a primitive polynomial and all the polynomials $p_3(x^3), p_5(x^5), \ldots, p_{2t-1}(x^{2t-1})$ must be divisible by $p(x)$, i.e.,

$$p_3(x^3) = p_5(x^5)\ldots p_{2t-1}(x^{2t-1}) = 0 \pmod{p(x)}. \tag{11.2.4}$$

We have shown that this form is sufficient and demonstrated how to find it in the previous lecture notes on polynomial codes. We show it here again with an example. For primitive $p(x) = 1 + x + x^4$, to find $p_3(x)$, note that $x^6 + x^9 + x^{12} = 1 + x^3 \pmod{p(x)}$. Then $1 + x^3 + x^6 + x^9 + x^{12} = 0 \pmod{p(x)}$. Let $y = x^3$ and let $p_3(y) = 1 + y + y^2 + y^3 + y^4$. It follows that $p_3(x^3) = 0 \pmod{p(x)}$.

Thus $Q(x) = p(x)p_3(x) = (1 + x + x^4)((1 + x + x^2 + x^3 + x^4)$ is a generator polynomial for a 2-error correcting code. Using a similar procedure, we find $p_5(x) = 1+x+x^2$. Therefore, $Q(x) = p(x)p_3(x)p_5(x) = (1 + x + x^4)((1 + x + x^2 + x^3 + x^4)(1 + x + x^2)$ is the generator polynomial for a 3-error correcting BCH code.

11.2.4 The Error Locator Polynomial and the Elementary Symmetric Functions

Define $t_i = \sum_{j=1}^{t} x^{e_j}$ to be the i-th power sums of the error monomials. Since $t_2 = t_1^2$ in the binary field, the even t_i provide no new information. The t_i's are called the power sum symmetric functions, and we will use these to find the bit error locations.

Define the error locator polynomial $E(y)$ of degree t such that it is equal to zero when evaluated at the error monomials and nonzero otherwise.

$$\begin{aligned} E(y) &= (y - x^{e_1})(y - x^{e_2}) \dots (y - x^{e_k}) \\ &= y^k + s_1 y^{k-1} + s_2 y^{k-2} + \dots + s_{k-1} y + s_k \\ &= 0. \end{aligned} \tag{11.2.5}$$

Then, from the fundamental theorem of algebra, for a t-error correcting code, all the roots of the error locator polynomial are the error monomials, x^{e_j}, $1 \le e_i \le n$ and $1 \le i \le t$. We need to find a relationship between the coefficients, s_j's of the error locator polynomial and the odd power sum symmetric functions, t_j's.

The k-th elementary symmetric function of d elements is defined as the sum of the products of k different elements from among the d elements, combined in all possible ways. For example, the second elementary function of a, b, c, d is $ab + ac + ad + bc + bd + cd$.

There is a linear relationship between the elementary symmetric functions and the odd power sums, t_i, of the error monomials. The coefficients, s_j, $1 \le j \le k$ of the error locator polynomial are related to the t_i's in the following for k odd and $s_0 = t_0 = 1$,

$$\sum_{i=1}^{k} s_i t_{k-1} = s_k + s_{k-1} t_1 + \dots + s_1 t_{k-1} + t_k = 0. \tag{11.2.6}$$

To see this for a k-error correcting code, note that the error locator polynomial in equation (11.2.5) evaluated at each error monomial $y = x^{e_i}$ must be satisfied. That is, for each j, $1 \le j \le t$

$$x^{te_j} + s_1 e^{(t-1)e_j} + \dots + s_{t-1} x^{e_j} + s_t = 0. \tag{11.2.7}$$

Sum equation (11.2.5) over all the error monomials. Then, t is the coefficient of s_t and t_{t-j} is the coefficient of s_j, except that the coefficient of s_t is t. For any $t' > t$, multiply equation (11.2.6) by $y^{t'-t}$, and follow the same proof – evaluate at each error monomial and sum. The relationship between s_i and t_i in equation (11.2.6) follows.

For the case of a two error correcting code over a binary field, the standard encoding polynomial is $Q(x) = p(x)p_3(x)$. The error locator polynomial will have degree 2, and its coefficients determined from equation (11.2.6) are given by $s_1 = t_1$ and $s_2 = t_1^2 + \frac{t_3}{t_1}$.

11.2.5 Example: 3 Error Correcting BCH Code

The encoding polynomial is of the form $Q(x) = p(x)p_3(x)p_5(x)$. Suppose we use $p(x) = 1 + x + x^4$, then $Q(x) = (1+x+x^4)((1+x+x^2+x^3+x^4)(1+x+x^2)$. Suppose the received sequence is 101000110110010,

then

$$\begin{aligned} r(x) &= 1 + x^2 + x^6 + x^7 + x^9 + x^{10} + x^{13} \\ r(x^3) &= 1 + x^6 + x^{18} + x^{21} + x^{27} + x^{30} + x^{39} \\ &= 1 + x^6 + x^3 + x^6 + x^{12} + x^0 + x^9 \\ &= x^{13} \\ r(x^5) &= 1 + x^2 + x^6 + x^7 + x^9 + x^{10} + x^{13} \\ &= 1 + x^{10} + x^{30} + x^{35} + x^{45} + x^{50} + x^{65} \\ &= 0 \end{aligned} \tag{11.2.8}$$

We divide by $p(x) = 1 + x + x^4$ for all 3 received sequence above. Note that $r(x^5) = 0$ means there are only 2 errors in our received sequence. Since $r(x^3) = x^{13}$, it is already a monomial. We only need to add the remainders of the monomials of $r(x)$. $1000 + 0010 + 0011 + 1101 + 0101 + 1110 + 1011 = 0100$ means $r(x) = x \pmod{p(x)}$.

$$\text{Rem}[r(x)] = t_1 = x \tag{11.2.9}$$

$$\text{Rem}[r(x^3)] = t_3 = x^{13} \tag{11.2.10}$$

$$\text{Rem}[r(x^5)] = t_5 = 0.$$

From the $s - t$ relations in equation (11.2.6), we have

$$s_1 + t_1 = 0 \tag{11.2.11}$$

$$s_3 + s_2 t_1 + s_1 t_2 + t_3 = 0 \tag{11.2.12}$$

$$s_3 t_2 + s_2 t_3 + s_1 t_4 + t_5 = 0 \tag{11.2.13}$$

Substituting equations (11.2.9) into equations (**??**) and solving, we get we get $s_1 = x,\ s_2 = x^7$ and $s_3 = 0$.

From the elementary symmetric functions,

$$s_1 = x^{e_1} + x^{e_2} + x^{e_3} = x \tag{11.2.14}$$

$$s_2 = x^{e_1}x^{e_2} + x^{e_2}x^{e_3} + x^{e_3}x^{e_1} = x^{e_1+e_2} + x^{e_2+e_3} + x^{e_3+e_1} = x^7 \tag{11.2.15}$$

$$s_3 = x^{e_1}x^{e_2}x^{e_3} = x^{e_1+e_2+e_3} = 0 \tag{11.2.16}$$

the error locator polynomial becomes

$$\begin{aligned} E(y) &= (y - x^{e_1})(y - x^{e_2})(y - x^{e_3}) \\ &= y^3 + s_1 y^2 + s_2 y + s_3 \\ &= y^3 + xy^2 + x^7 y \\ &= y^2 + xy + x^7 = 0. \end{aligned}$$

We now substitute in all the monomials $y = x^j,\ 0 \le j \le 2^r$ to see which monomials are solutions to the error locator polynomial. We find that x^2 and x^5 are roots. The correct sequence is 100001110110010 and the message is 11010.

Unit 12

Course Structure

- Reed-Muller Codes

12.1 Reed-Muller Codes

12.1.1 Introduction to Reed-Muller Codes

Reed-Muller codes were invented in 1954 by the American mathematician David E. Muller and so RM codes are one of the oldest families of codes. The first efficient decoding algorithm was created by Irving S. Reed.

RM codes are linear block codes and initially they belonged to the class of binary codes. It is this traditional form of RM codes that we will study in this chapter. Characteristic for RM codes are their advantageous properties for coding and decoding. They are useful in a wide range of functions, especially in wireless and deep-space communication.

Definition 12.1.1. The rth order Reed-Muller Code $R(r, m)$ has the positive integers r and m for which $0 \leq r \leq m$. The length is $n = 2^m$ and the minimum distance is $d = 2^{m-r}$. The RM code consists of the vectors f where $f(v_1, v_2, \ldots, v_m)$. This function is a Boolean function since it only includes the values 0 and 1. It forms a polynomial of the maximum degree r. An example of this is the first-order RM code of length 16 which is a polynomial of the first degree: $a_0 1 + a_1 v_1 + a_2 v_2 + a_3 v_3 + a_4 v_4$, $a_i = 0$ or 1.

12.1.2 First-Order RM Codes

First-order RM codes have the advantage of working over particularly noisy channels. They can correct many errors and is notably easy to encode and decode. Consequently, first-order RM codes have been found useful in deep-space data transmission where they, for example, have been applied for transmitting pictures from Mars.

Definition 12.1.2. The 1th order Reed-Muller code $R(1, m)$ has the positive integer m and $r = 1$. It can be described as a $\left[2^m, m+1, 2^{m-1}\right]$ code and it is defined for all integers $m \geq 1$. Due to the low rate of first-order RM codes, they have the ability to correct numerous errors. For this reason, they have proved to be suitable for especially noisy channels.

In the case of $R(1,1)$ we find that the codewords are $(00, 01, 10, 11)$. If $m > 1$ then $R(1,m) = \{(u,u), (u, u+1) : u \in R(1, m-1)\}$. This is an example of the $|u|u+v|$ construction which declares the method of forming a new code consisting of two previous codes. We may therefore have a code $C_1[n, M_1, d_1]$ and another code $C_2[n, M_2, d_2]$. As we can see these codes are of the same length. Together they form a new code C_3 which consists of all the vectors $|u|u+v|$ where $u \in C_1$ and $v \in C_2$. As a result, the new code is of double length $2n$.

Theorem 12.1.3. $R(1,m)$ is a $[2^m, m+1, 2^{m-1}]$ code where $m > 0$. The code has minimum distance (Hamming weight) 2^{m-1} since every codeword except 0 and 1 has weight 2^{m-1}.

Proof. We will prove this theorem by induction.

Base case: We have previously described that it is obvious that $R(1,1)$ is a $[2,2,1]$ code. So the theorem holds for $R(1,1)$.

Inductive step: We assume that $R(1, m-1)$ is a $[2^{m-1}, m, 2^{m-2}]$ code. $R(1,m)$ can therefore be created using the $|u|u+v|$ construction. Let $C_1 = R(1, m-1)$ and $C_2 = \{0,1\} = R(0, m-1)$.

Consequently, $R(1,m)$ is a $[2(2^{m-1}), m+1, 2(2^{m-2})]$ code, in other words, a $[2^m, m+1, 2^{m-1}]$ code. Furthermore, we assumed that $R(1, m-1)$ had weight 2^{m-2}. Since a codeword in $R(1,m)$ is constructed as $(u, u+v)$, (u,u) has weight $2(2^{m-2}) = 2^{m-1}$. Additionally, we study the other codewords $(u, u+1)$:

If $u = 0$, then $u+1$ is 1. We see that half of each codeword consists of 1's. Therefore, $wt(u, u+1) = 2^{m-1}$,

If $u = 1$, then $u + 1 = 0$. Again, half of each codeword is 1's and $wt(u, u+1) = 2^{m-1}$.

For every other u in $R(1, m-1) : wt(u) = 2^{m-2}$, which means that half of each vector u consists of 1's. The same applies to $u+1$ and so $wt(u, u+1) = 2(2^{m-2}) = 2^{m-1}$. □

Theorem 12.1.4. $R(r+1, m+1) = \{|u|u+v| : u \in R(r+1, m),\ v \in R(r,m)\}$.

12.1.3 Encoding

In this part of the unit, we will describe some of the methods for encoding first-order RM codes. In the following examples, the reader will find first-order RM codes of length 8 and 16. When it comes to the first-order RM code of length 16 a generator matrix, afterwards included in the first encoding method, is presented. Later, we will describe the clock circuit which makes encoding of first-order RM codes exceedingly simple.

Example 12.1.5. $R(1,3)$ According to the previous definition the first order RM code of length 8 is a polynomial of the first degree:

$$a_0 1 + a_1 v_1 + a_2 v_2 + a_3 v_3, \quad a_i = 0 \text{ or } 1.$$

In this example, $n = 8 = 2^m$. Hence the positive integer $m = 3$ and $R(1,3)$.

$$R(1,3) = \{(u,u),\ (u, u+1) : u \in R(1,2)\},$$

and so, the $8 \times 2 = 16$ codewords are given in figure 12.1

The weight of every codeword except 0 and 1 is 2^{m-1}. In this case the weight is $2^2 = 4$. Furthermore,

$$k = 1 + \binom{m}{1} + \binom{m}{2} + \ldots + \binom{m}{r}$$

where k is the dimension of the code. When the length of the first-order RM code is 8 then

$$k = 1 + \binom{3}{1} = 4.$$

As a result, the dimension of the code is 4 which also indicates the number of basic vectors.

$\mathbf{0}$	0 0 0 0 0 0 0 0	$\mathbf{1}$	1 1 1 1 1 1 1 1
v_3	0 0 0 0 1 1 1 1	$\mathbf{1}+v_3$	1 1 1 1 0 0 0 0
v_2	0 0 1 1 0 0 1 1	$\mathbf{1}+v_2$	1 1 0 0 1 1 0 0
v_1	0 1 0 1 0 1 0 1	$\mathbf{1}+v_1$	1 0 1 0 1 0 1 0
v_2+v_3	0 0 1 1 1 1 0 0	$\mathbf{1}+v_2+v_3$	1 1 0 0 0 0 1 1
v_1+v_3	0 1 0 1 1 0 1 0	$\mathbf{1}+v_1+v_3$	1 0 1 0 0 1 0 1
v_1+v_2	0 1 1 0 0 1 1 0	$\mathbf{1}+v_1+v_2$	1 0 0 1 1 0 0 1
$v_1+v_2+v_3$	0 1 1 0 1 0 0 1	$\mathbf{1}+v_1+v_2+v_3$	1 0 0 1 0 1 1 0

Figure 12.1

Example 12.1.6. The first order RM code of length 16 is a polynomial of the first degree:

$$a_0 1 + a_1 v_1 + a_2 v_2 + a_3 v_3 + a_4 v_4, \;\; a_i = 0 \;\text{ or }\; 1.$$

$n = 16 = 2^4$ and $R(1,4)$. The weight of every codeword, except 0 and 1, is $2^{4-1} = 8$,

$$k = 1 + \binom{4}{1} = 5.$$

The generator matrix for first-order RM codes of length 16 consists of 5 basic vectors. Basic vectors are always linearly independent. The fifth dimension signifies the first five rows in the generator matrix for all $R(r,4)$ up to the 4th order RM code. The generator matrix for first-order RM codes of length 16 consists of the five basic vectors (figure 12.2).

$\mathbf{1}$	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
v_4	0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1
v_3	0 0 0 0 1 1 1 1 0 0 0 0 1 1 1 1
v_2	0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1
v_1	0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1

Figure 12.2: Basic vectors for first-order RM codes of length 16

We will now apply the generator matrix. $R(1,4)$ 4) in order to show the reader how encoding using a generator matrix works. We use the generator matrix, G, consisting of the five basic vectors listed above, and the message symbols:

$$a = a_0 a_4 a_3 a_2 a_1, \;\; a_i = 0 \;\text{ or }\; 1.$$

Together they are encoded into the codeword x:

$$\begin{aligned} x = aG &= \begin{bmatrix} a_0 \\ a_4 \\ a_3 \\ a_2 \\ a_1 \end{bmatrix} \begin{bmatrix} 1111111111111111 \\ 0000000011111111 \\ 0000111100001111 \\ 0011001100110011 \\ 0101010101010101 \end{bmatrix} \\ &= a_0 1 + a_4 v_4 + a_3 v_3 + a_2 v_2 + a_1 v_1 \\ (\quad &= x_0 x_1 \dots x_{15}, \text{ for example}). \end{aligned}$$

Encoding using a generator matrix can be applied for all RM codes, $R(r, m),\ 0 \leq r \leq m$. The minimum distance, $d = 2^{m-r}$, and the code can correct $\frac{1}{2}(d-1)$ errors. And so, $R(1, 4)$ can correct 3 errors while $R(2, 4)$ is a single-error correcting code. The Reed Decoding Algorithm can be used to decode these RM codes.

In general, RM codes have proved easy to encode and decode. Nevertheless, when it comes to first-order RM codes, encoding becomes exceedingly simple. A circuit can handle the process of encoding a message into a codeword. An example of an encoder for $R(1, 4)$ is presented here:

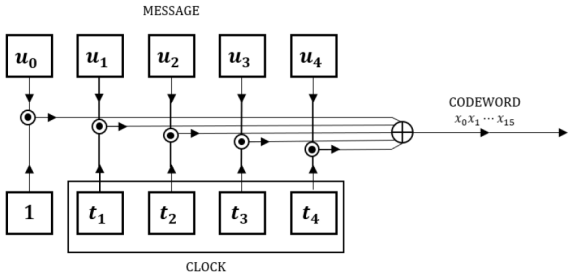


Figure 12.3: Encoder (Clock circuit) for $R(1, 4)$

The procedure is similar to encoding using generator matrix. For $R(1, 4)$, an $[16, 5, 8]$ code, a message including 5 message symbols is encoded into the codeword $(x_0x_1 \dots x_{15})$. The generator matrix, consisting of 5 basic vectors, and the message (aG) equals the codeword (x). The clock circuit in 12.3 accomplishes this by counting from 0 to 15 through

$$t_1t_2t_3t_4 = 0000, 0001, 0010, 0011, 0100, 0101, \dots, 1111, 0000, 0001, \dots$$

As a result, the circuit forms $u_0 1 + t_1u_1 + t_2u_2 + t_3u_3 + t_4u_4$ which is the codeword $(x_0x_1 \dots x_{15})$.

12.1.4 Decoding

Reed-Muller codes have the advantage of being easier to decode than many other codes. RM codes belong to the class of geometrical codes and can therefore be decoded by majority logic. Consequently, the following examples of decoding methods and algorithms are based on majority logic decoding.

The Reed Decoding Algorithm works for all RM codes. Here we present an example of decoding $R(1, 4)$

with this algorithm. In fig. 12.2 we can see that, without any errors:

$$\begin{aligned} a_1 &= y_0 + y_1 \\ &= y_2 + y_3 \\ &\dots \\ &= y_{14} + y_{15}, \\ a_2 &= y_0 + y_2 \\ &= \dots \end{aligned}$$

We observe that there are 8 votes for every a_i. It is clear that $R(1,4)$ can correct up to 3 errors by majority logic. If $R(2,4)$, then the 6 message symbols, a_{12} to a_{34}, each have 4 votes and, as a result, the code can only correct one error by majority logic. We now continue to a_0:

$$x'y - a_4v_4 - \dots - a_1v_1 = a_01 + \text{error},$$

$a_0 = 0$ or 1 based on the number of 1's in x'.

Example 12.1.7. Assume that we have a $[8,4,4]$ code, in other words $R(1,3)$, and that we receive 10101101. Decode the received message using Reed's Algorithm.

$a_1 = 1 = 1 = 0 = 1$, by majority logic, $a_1 = 1$.
$a_2 = 0 = 0 = 1 = 0$, so $a_2 = 0$.
$a_3 = 0 = 1 = 1 = 1$, so $a_3 = 1$.
$x' = 10101101 - 00001111 - 00000000 - 01010101 = 11110111$.
$a_0 = 1$ (modulo 2).
The correct message is $a = 1101$ and the codeword is$x = y+$ error$= 10101101 + 00001000 = 10100101$.
We can see that this is an example of a nonsystematic code.

Definition 12.1.8. The input consists of a function $f : F_2^m \to F_2$ such that there exists a polynomial P of degree r with $\Delta(f,P) < \frac{2^{m-r}}{2}$. The output of Reed's Algorithm is the polynomial P.

Theorem 12.1.9. In the case of no errors, $a_\sigma = \sum_{P \in U_i} x_p$, $i = 1, \dots, 2^{m-r}$, where σ indicates a string of r symbols.

If more than $\frac{1}{2}(2^{m-r} - 1)$ errors occur, then the Reed Decoding Algorithm can correct the same number of errors. For example, if no more than 3 errors occur in $R(1,4)$ then the algorithm can correct them. The theorem consequently implies that all a_σ can be recovered correctly if no more than $\frac{1}{2}(2^{m-r} - 1)$ errors occur. If $r > 1$, then all a_{σ_r} can be recovered and subsequently the rest of the a's by majority logic decoding.

Unit 13

Course Structure

- Markovian decision Process
- Powers of Stochastic Matrices
- Regular matrices

13.1 Introduction

A Markov Process consists of a set of objects and a set of states such that

(i) at any given time, each object must be in a state (distinct objects need not be in distinct states).

(ii) the probability that an object moves from one state to another state which may be the same as the first state, in one time period depends only on those two states.

The integral numbers of time periods past the moment when the process is started represent the stages of the process which may be finite or infinite.

If the number of states is finite or countably infinite, the Markov process is called a **Markov Chain**. A finite Markov chain is one having a finite number of states. We denote the probability of moving from state i to state j in one time period by p_{ij}. For an N state Markov chain, where N is a fixed positive integer, the $N \times N$ matrix $P = [p_{ij}]$ is the **stochastic** or **transition matrix** associated with the process. Necessarily, the elements of each row of P sum to unity.

Theorem 13.1.1. Every stochastic matrix has 1 as an eigen value (possible multiple and none of the eigen values exceed 1 in absolute value).

Because of the way P is defined, it proves convenient in this chapter to indicate N-dimensional vectors as row vectors.

According to the theorem, there exists a vector $X \neq 0$ such that $XP = X$. This left eigen value is called a fixed point of P.

13.2 Powers of Stochastic Matrices

We denote the nth power of a matrix P by

$$P^n \equiv [p_{ij}^{(n)}],$$

where $p_{ij}^{(n)}$ represents the probability that an object moves from state i to state j in n-time periods. P^n is obviously a stochastic matrix.

We write $X^{(0)} = [x_1^{(0)}, x_2^{(0)}, \ldots, x_N^{(0)}]$ which represents the proportion of objects in each state of the beginning of the process whereas

$$X^{(n)} = [x_1^{(n)}, x_2^{(n)}, \ldots, x_N^{(n)}],$$

where, $x_i^{(N)}$ represents the proportion of objects in state i at the end of nth time period, $1 \leq i \leq N$.

$X^{(n)}$ is related to $X^{(0)}$ by the relation $X^{(n)} = X^{(0)}P^n$.

Example 13.2.1. Grapes in Kashmir are classified as either superior, average or poor. Following a superior harvest, the probabilities of having a superior, average and poor harvest in the next year are 0, 0.8 and 0.2. Following an average harvest, the probabilities of a superior, average and poor harvest are 0.2, 0.6 and 0.1. Following a poor harvest, the probabilities of a superior, average and poor harvest are 0.1, 0.8 and 0.1. Determine the probabilities of a superior harvest for each of the next five years if the most recent harvest was average.

Solution. The transition matrix is given by

$$\begin{array}{c} \\ S \\ A \\ P \end{array} \begin{array}{c} \begin{array}{ccc} \text{superior}(S) & \text{average}(A) & \text{poor}(P) \end{array} \\ \begin{pmatrix} 0 & 0.8 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.8 & 0.1 \end{pmatrix} \end{array}$$

Since the most recent harvest rate was average, so,

$$X^{(0)} = \begin{array}{c} \begin{array}{ccc} \text{S} & \text{A} & \text{P} \end{array} \\ \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} \end{array}$$

initial probability distribution. Thus,

$$X^{(5)} = X^{(0)}P^5.$$

Now,

$$\begin{aligned} P^2 &= \begin{bmatrix} 0 & 0.8 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.8 & 0.1 \end{bmatrix} \begin{bmatrix} 0 & 0.8 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.8 & 0.1 \end{bmatrix} \\ &= \begin{bmatrix} 0+0.16+0.02 & 0+0.48+0.16 & 0+0.16+0.02 \\ 0+0.12+0.02 & 0.16+0.36+0.16 & 0.04+0.12+0.02 \\ 0+0.16+0.01 & 0.08+0.48+0.08 & 0.02+0.16+0.01 \end{bmatrix} \\ &= \begin{bmatrix} 0.18 & 0.64 & 0.18 \\ 0.14 & 0.68 & 0.18 \\ 0.17 & 0.64 & 0.19 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} P^4 &= \begin{bmatrix} 0.18 & 0.64 & 0.18 \\ 0.14 & 0.68 & 0.18 \\ 0.17 & 0.64 & 0.19 \end{bmatrix} \begin{bmatrix} 0.18 & 0.64 & 0.18 \\ 0.14 & 0.68 & 0.18 \\ 0.17 & 0.64 & 0.19 \end{bmatrix} \\ &= \begin{bmatrix} 0.1526 & 0.6656 & 0.1818 \\ 0.1510 & 0.6672 & 0.1818 \\ 0.1525 & 0.6656 & 0.1819 \end{bmatrix}. \end{aligned}$$

$$\begin{aligned} P^5 &= \begin{bmatrix} 0.1526 & 0.6656 & 0.1818 \\ 0.1510 & 0.6672 & 0.1818 \\ 0.1525 & 0.6656 & 0.1819 \end{bmatrix} \begin{bmatrix} 0.18 & 0.64 & 0.18 \\ 0.14 & 0.68 & 0.18 \\ 0.17 & 0.64 & 0.19 \end{bmatrix} \\ &= \begin{bmatrix} 0.151558 & 0.666624 & 0.181818 \\ 0.151494 & 0.666688 & 0.181818 \\ 0.151557 & 0.666624 & 0.181819 \end{bmatrix}. \end{aligned}$$

Thus,

$$\begin{aligned} X^{(5)} &= \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} P^5 \\ &= \begin{bmatrix} 0.151494 & 0.666688 & 0.181818 \end{bmatrix}. \end{aligned}$$

Hence the probability of a superior harvest for each of the next five years is 0.151494. ■

Definition 13.2.2. (Regular Matrix:) A stochastic matrix is regular if one of its powers contains only positive entries.

Theorem 13.2.3. If a stochastic matrix is regular, then 1 is an eigen value of multiplicity one, and all other eigen values λ_i satisfy $|\lambda_i| \leq 1$.

Example 13.2.4. Is the stochastic matrix

$$P = \begin{bmatrix} 0 & 1 \\ 0.4 & 0.6 \end{bmatrix}$$

regular?

Solution.

$$P^2 = \begin{bmatrix} 0 & 1 \\ 0.4 & 0.6 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0.4 & 0.6 \end{bmatrix} = \begin{bmatrix} 0.40 & 0.60 \\ 0.24 & 0.76 \end{bmatrix}.$$

Since each entry of P^2 is positive, hence P is regular. ■

Unit 14

Course Structure

- Ergodic Matrices

14.1 Ergodic Matrix

Definition 14.1.1. (Ergodic Matrix:) A stochastic matrix P is ergodic if $\lim_{n\to\infty} P^n$ exists, that is, each $P_{ij}^{(n)}$ has a limit as $n \to \infty$. We denote $L = \lim_{n\to\infty} P^n$. Obviously, P is a stochastic matrix. $X^{(\infty)}$ is defined by the equation $X(\infty) = X^{(0)}L$.

The components of $X^{(\infty)}$ are limiting state distributions and represent the approximate proportions of objects in the various states of a Markov chain after a large number of time periods.

Theorem 14.1.2. A stochastic matrix is ergodic if and only if the only eigen value λ of magnitude 1 is 1 itself and if $\lambda = 1$ has multiplicity k, then there exists k linearly independent (left) eigen vectors associated with this eigen value.

Theorem 14.1.3. A regular matrix is ergodic but the converse is not true in general.

If P is regular with limit matrix L, then the rows of L are identical with one another, each being the unique left eigen vector of P associated with the eigen value $\lambda = 1$ and having the sum of its components equal to unity.

Let us denote this eigen vector by E_1. Now, if P is regular, then regardless of the initial distribution $X^{(0)}$, we can write $X^{(\infty)} = E_1 (= X^{(0)}L)$.

Example 14.1.4. Is the stochastic matrix

$$P = \begin{bmatrix} 0 & 1 \\ 0.4 & 0.6 \end{bmatrix}$$

ergodic? Calculate $L = \lim_{n\to\infty} P^n$, if it exists.

Solution. Since each entry of

$$P^2 = \begin{bmatrix} 0.40 & 0.60 \\ 0.24 & 0.76 \end{bmatrix}$$

is positive, P is regular and therefore, ergodic; hence $L = \lim_{n\to\infty} P^n$ exists. Now,

$$\begin{aligned} \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 0.40 & 0.60 \\ 0.24 & 0.76 \end{bmatrix} &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \\ \Rightarrow\ x_1 - 0.4x_2 &= 0 \qquad (14.1.1) \\ \text{and}\ \ x_1 + x_2 &= 1. \qquad (14.1.2) \end{aligned}$$

Solving equation (14.1.1) and (14.1.2), we get,

$$x_1 = \frac{2}{7} \quad \text{and} \quad x_2 = \frac{5}{7}.$$

Thus,

$$E_1 = \begin{bmatrix} \frac{2}{7} & \frac{5}{7} \end{bmatrix} \quad \text{and} \quad \lim_{n\to\infty} P^n = L = \begin{bmatrix} \frac{2}{7} & \frac{5}{7} \\ \frac{2}{7} & \frac{5}{7} \end{bmatrix}.$$

■

Theorem 14.1.5. If every eigen value of a matrix P yields linearly independent (left) eigen vectors in number equal to its multiplicity, then there exists a non-singular matrix M, whose rows are left eigen vectors of P, such that $D \equiv MPM^{-1}$ is a diagonal matrix. The diagonal elements of D are the eigen values of P, repeated according to multiplicity.

We have,

$$\begin{aligned} L &= \lim_{n\to\infty} P^n \\ &= (M^{-1}M) \lim_{n\to\infty} P^n (M^{-1}M) \\ &= M^{-1} \left(\lim_{n\to\infty} MP^nM^{-1} \right) M \\ &= M^{-1} \left(\lim_{n\to\infty} (MPM^{-1})^n \right) M \\ &= M^{-1} \left(\lim_{n\to\infty} D^n \right) M \\ &= M^{-1} \begin{bmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & \ddots & & & & \\ & & & 1 & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{bmatrix}_{N\times N} M. \end{aligned}$$

The diagonal matrix on the right has k $1's$ and $(N-k)$ $0's$ on the main diagonal.

Example 14.1.6. Is the stochastic matrix

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.4 & 0 & 0.6 & 0 \\ 0.2 & 0 & 0.1 & 0.7 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

regular? Is it ergodic? Calculate $L = \lim_{n\to\infty} P^n$, if it exists.

Solution. The characteristic equation of P is

$$\begin{aligned}
\begin{vmatrix} 1-\lambda & 0 & 0 & 0 \\ 0.4 & -\lambda & 0.6 & 0 \\ 0.2 & 0 & 0.1-\lambda & 0.7 \\ 0 & 0 & 0 & 1-\lambda \end{vmatrix} &= 0 \\
\Rightarrow (1-\lambda)(-\lambda)(0.1-\lambda)(1-\lambda) &= 0 \\
\Rightarrow \lambda &= 1,\ 1,\ 0.1,\ 0.
\end{aligned}$$

Thus, $\lambda_1 = 1$ (multiplicity 2), $\lambda_2 = 0.1$, $\lambda_3 = 0$ are the eigen values of P. Hence P is not regular.

The left eigen vectors for the double eigen value $\lambda_1 = 1$ are $[1, 0, 0, 0]$ and $[0, 0, 0, 1]$, which are linearly independent. Hence P is ergodic. Thus, $L = \lim_{n\to\infty} P^n$ exists.

We now fine the eigen vectors corresponding to $\lambda_2 = 0.1$ and $\lambda_3 = 0$.

$$\begin{aligned}
\begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.4 & 0 & 0.6 & 0 \\ 0.2 & 0 & 0.1 & 0.7 \\ 0 & 0 & 0 & 1 \end{bmatrix} &= 0.1 \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \\
\Rightarrow (1-0.1)x_1 + 0.4x_2 + 0.2x_3 &= 0 \\
-0.2x_2 &= 0 \\
0.6x_2 + (0.1-0.1)x_3 &= 0 \\
0.7x_3 + (1-0.1)x_4 &= 0 \\
\Rightarrow 0.9x_1 + 0.4x_2 + 0.2x_3 &= 0 \\
-0.1x_2 &= 0 \\
0.6x_2 &= 0 \\
0.7x_3 + 0.9x_4 &= 0.
\end{aligned}$$

Solving these equations, we get,

$$x_1 = -2, \quad x_2 = 0, \quad x_3 = 9, \quad x_4 = -7.$$

Thus, the eigen vector corresponding to λ_2 is $[-2, 0, 9, -7]$. Again,

$$\begin{aligned}
\begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.4 & 0 & 0.6 & 0 \\ 0.2 & 0 & 0.1 & 0.7 \\ 0 & 0 & 0 & 1 \end{bmatrix} &= 0 \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \\
\Rightarrow x_1 + 0.6x_2 + 0.2x_3 &= 0 \\
0.6x_2 + 0.1x_3 &= 0 \\
0.7x_3 + x_4 &= 0.
\end{aligned}$$

Solving these equations, we get

$$x_1 = 4, \quad x_2 = 5, \quad x_3 = -30, \quad x_4 = 21.$$

Thus, the eigen vector corresponding to λ_3 is $[4, 5, -30, 21]$.

To make P diagonalizable, we consider

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -2 & 0 & 9 & -7 \\ 4 & 5 & -30 & 21 \end{bmatrix} \text{ and } D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

We now find M^{-1}.

$$\begin{aligned}
[M : I] \quad &= \quad \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ -2 & 0 & 9 & -7 & 0 & 0 & 1 & 0 \\ 4 & 5 & -30 & 21 & 0 & 0 & 0 & 1 \end{array}\right] \\
&\xrightarrow[R_4 \to R_2]{R_2 \to R_4} \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 4 & 5 & -30 & 21 & 0 & 0 & 0 & 1 \\ -2 & 0 & 9 & -7 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}\right] \\
&\xrightarrow[R_3 \to R_3 + 2R_1]{R_2 \to R_2 - 4R_1} \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 5 & -30 & 21 & -4 & 0 & 0 & 1 \\ 0 & 0 & 9 & -7 & 2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}\right] \\
&\xrightarrow[R_3 \to \frac{1}{9} R_3]{R_2 \to \frac{1}{5} R_2} \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & -6 & \frac{21}{5} & -\frac{4}{5} & 0 & 0 & \frac{1}{5} \\ 0 & 0 & 1 & -\frac{7}{9} & \frac{2}{9} & 0 & \frac{1}{9} & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}\right] \\
&\xrightarrow{R_2 \to R_2 + 6R_3} \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -\frac{7}{15} & \frac{8}{15} & 0 & \frac{2}{3} & \frac{1}{5} \\ 0 & 0 & 1 & -\frac{7}{9} & \frac{2}{9} & 0 & \frac{1}{9} & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}\right] \\
&\xrightarrow[R_3 \to R_3 + \frac{7}{9} R_4]{R_2 \to R_2 + \frac{7}{15} R_4} \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & \frac{8}{15} & \frac{7}{15} & \frac{2}{3} & \frac{1}{5} \\ 0 & 0 & 1 & 0 & \frac{2}{9} & \frac{7}{9} & \frac{1}{9} & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}\right].
\end{aligned}$$

Thus

$$M^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{8}{15} & \frac{7}{15} & \frac{2}{3} & \frac{1}{5} \\ \frac{2}{9} & \frac{7}{9} & \frac{1}{9} & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Thus,

$$
\begin{aligned}
L &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{8}{15} & \frac{7}{15} & \frac{2}{3} & \frac{1}{5} \\ \frac{2}{9} & \frac{7}{9} & \frac{1}{9} & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -2 & 0 & 9 & -7 \\ 4 & 5 & -30 & 21 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{8}{15} & \frac{7}{15} & 0 & 0 \\ \frac{2}{9} & \frac{7}{9} & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -2 & 0 & 9 & -7 \\ 4 & 5 & -30 & 21 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{8}{15} & 0 & 0 & \frac{7}{15} \\ \frac{2}{9} & 0 & 0 & \frac{7}{9} \\ 0 & 0 & 0 & 1 \end{bmatrix}.
\end{aligned}
$$

■

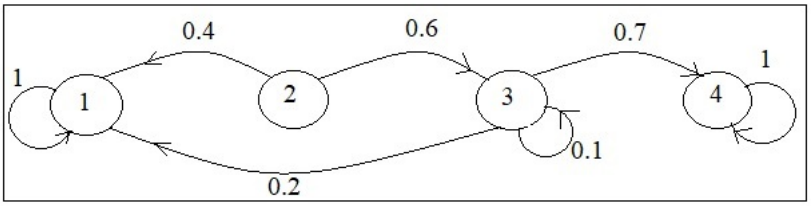
Example 14.1.7. Construct the state-transition diagram for the Markov chain

$$
P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.4 & 0 & 0.6 & 0 \\ 0.2 & 0 & 0.1 & 0.7 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}
$$

Solution. [A state-transition diagram is an oriented network in which the nodes represent states and the arcs represent possible transitions.]

Labelling the states by $1, 2, 3, 4$, we have the following state-transition diagram.

The number on each arc is the probability of the transition. ■



Example 14.1.8. Prove that if P is regular, then all the rows of $L = \lim_{n\to\infty} P^n$ are identical.

Solution. Given, $L = \lim_{n\to\infty} P^n$. Also, we have, $L = \lim_{n\to\infty} P^{n-1}$. Consequently,

$$
L = \lim_{n\to\infty} P^n = \lim_{n\to\infty} (P^{n-1})P = (\lim_{n\to\infty} P^{n-1})P = LP
$$

which implies that every row of L is a left eigen vector of P corresponding to the eigen value $\lambda = 1$.

Now, P being regular, all such eigen vectors are scalar multiples of a single vector.

On the other hand, L being stochastic, each row of it sums to unity. Thus it follows that all the rows are identical. ■

Example 14.1.9. Prove that if λ is an eigen value of a stochastic matrix P, then $|\lambda| \leq 1$.

Solution. Let $E = [e_1 \quad e_2 \quad \dots \quad e_N]^T$ be a right eigen vector corresponding to λ. Then $PE = \lambda E$, and considering the jth component of both sides of this equality, we conclude that

$$\sum_{k=1}^{N} p_{jk} e_k = \lambda e_j. \tag{14.1.3}$$

Let e_i be that component of E having the greatest magnitude, that is,

$$|e_i| = \max\{|e_1|, |e_2|, \dots, |e_N|\}. \tag{14.1.4}$$

By definition, $E \neq 0$, so that $|e_i| > 0$. Thus, it follows from (14.1.3), with $j = i$ and from (14.1.4) that,

$$|\lambda||e_i| = |\lambda e_i| = \left|\sum_{k=1}^{N} p_{ik} e_k\right| \leq \sum_{k=1}^{N} p_{ik}|e_k| \leq |e_i| \sum_{k=1}^{N} p_{ik} = |e_i|,$$

which implies that $|\lambda| \leq 1$. ■

Example 14.1.10. Formulate the following process as a Markov chain:

The manufacturer of Hi-Glo toothpaste currently controls 60% of the market in a particular city. Data from the previous year show that 88% of Hi-Glo's customers remained loyal to Hi-Glo, while 12% of Hi-Glo's customers switched to rival brands. In addition, 85% of the competition's customers remained loyal to the competition, while the other 15% switched to Hi-Glo. Assuming that these trends continue, determine Hi-Glo's share of the market

(a) in 5 years and (b) over the long run.

Solution. We take state 1 to be consumption of Hi-Glo toothpaste and state 2 to be consumption of a rival brand. Then p_{11} is the probability that a Hi-Glo customer remains loyal to Hi-Glo, that is, 0.88; p_{12} is the probability that a Hi-Glo customer switches to another brand, that is, 0.12; p_{21} is the probability that the customer of another brand switches to Hi-Glo, that is, 0.15; p_{22} is the probability that customer of another brand remains loyal to the competition, that is, 0.85.

The stochastic matrix (Markov chain) defined by these transition probabilities is

$$P = \begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{pmatrix} 0.88 & 0.12 \\ 0.15 & 0.85 \end{pmatrix} \end{array}$$

The initial probability distribution vector is $X^{(0)} = [0.60 \quad 0.40]$, where, the components $x_1^{(0)} = 0.60$ and $x_2^{(0)} = 0.40$ represent the proportions of people initially in states 1 and 2, respectively.

(a) Thus,

$$\begin{aligned} X^{(5)} &= X^{(0)} P^5 \\ &= \begin{bmatrix} 0.60 & 0.40 \end{bmatrix} \begin{bmatrix} 0.6477 & 0.3523 \\ 0.4404 & 0.5596 \end{bmatrix} \\ &= \begin{bmatrix} 0.5648 & 0.4352 \end{bmatrix}. \end{aligned}$$

After 5 years, Hi-Glo's share of the market will have declined to 56.48%. Now,

$$P = \begin{bmatrix} 0.88 & 0.12 \\ 0.15 & 0.85 \end{bmatrix}$$

is regular, since each entry of the first power of P is positive, that is, P is positive. Hence P is ergodic. So, $\lim_{n\to\infty} P^n = L$(say) exists. Now, the left eigen vector corresponding to $\lambda = 1$ is given by

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 0.88 & 0.12 \\ 0.15 & 0.85 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}$$
$$\Rightarrow 0.12x_1 - 0.15x_2 = 0 \quad \text{and} \quad x_1 + x_2 = 1.$$

Solving, we get,

$$x_1 = \frac{5}{9} \quad \text{and} \quad x_2 = \frac{4}{9}$$

and thus

$$E_1 = \begin{bmatrix} x_1 & x_2 \end{bmatrix} = \begin{bmatrix} \frac{5}{9} & \frac{4}{9} \end{bmatrix}.$$

Hence,

$$L = \lim_{n\to\infty} P^n = \begin{bmatrix} \frac{5}{9} & \frac{4}{9} \\ \frac{5}{9} & \frac{4}{9} \end{bmatrix}.$$

(b)

$$\begin{aligned} X^{(\infty)} &= X^{(0)} L \\ &= \begin{bmatrix} 0.60 & 0.40 \end{bmatrix} \begin{bmatrix} \frac{5}{9} & \frac{4}{9} \\ \frac{5}{9} & \frac{4}{9} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{3} + \frac{2}{9} & \frac{12}{45} + \frac{16}{45} \end{bmatrix} = \begin{bmatrix} \frac{5}{9} & \frac{4}{9} \end{bmatrix} = E_1. \end{aligned}$$

Therefore, over the long run, Hi-Glo's share of the market will stabilize at $\frac{5}{9}$, that is, approximately 55.56%.

■

Example 14.1.11. Solve the previous problem, if Hi-Glo currently controls 90% of the market

(a)

$$\begin{aligned} X^{(5)} &= X^{(0)} P^5 \\ &= \begin{bmatrix} 0.90 & 0.10 \end{bmatrix} \begin{bmatrix} 0.6477 & 0.3523 \\ 0.4404 & 0.5596 \end{bmatrix} \\ &= \begin{bmatrix} 0.6270 & 0.3730 \end{bmatrix}. \end{aligned}$$

Therefore, after 5 years, Hi-Glo controls approximately 68% of the market.

(b) Since P is regular,

$$X^{(\infty)} = E_1 = \begin{bmatrix} \frac{5}{9} & \frac{4}{9} \end{bmatrix}.$$

Example 14.1.12. The geriatric ward of a hospital lists its patients as bedridden or ambulatory. Historical data indicate that over a 1-week period, 30% of all ambulatory patients are discharged, 40% remain ambulatory, and 30% are remanded to complete bed rest. During the same period, 50% of all the bedridden patients become ambulatory, 20% remain bedridden, and 30% die. Currently the hospital has 100 patients in its geriatric ward, with 30 bedridden and 70 ambulatory. Determine the status of the patients

(a) after 2 weeks, and

(b) over the long run

(The status of a discharged patient does not change if the patient die).

Solution. We take state 1 to be discharged, sate 2 to be ambulatory, state 3 to be bedridden or bed rest and state 4 to be died patients. Consider 1 time period to be 1 week.

The transition probabilities given by the following transition matrix:

$$P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1(\text{Discharged}) & 2(\text{Ambulatory}) & 3(\text{Bedridden}) & 4(\text{Died}) \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.3 & 0.4 & 0.3 & 0 \\ 0 & 0.5 & 0.2 & 0.3 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

Since, currently the hospital has 100 patients in its geriatric ward, with 30 bedridden and 70 ambulatory, so the initial probability distribution vector is

$$X^{(0)} = \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 0.7 & 0.3 & 0 \end{pmatrix}$$

Now,

$$\begin{aligned} P^2 &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.3 & 0.4 & 0.3 & 0 \\ 0 & 0.5 & 0.2 & 0.3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.3 & 0.4 & 0.3 & 0 \\ 0 & 0.5 & 0.2 & 0.3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.42 & 0.31 & 0.18 & 0.09 \\ 0.15 & 0.30 & 0.19 & 0.36 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \end{aligned}$$

(a)

$$\begin{aligned} X^{(2)} &= X^{(0)}P^2 \\ &= \begin{bmatrix} 0 & 0.7 & 0.3 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.42 & 0.31 & 0.18 & 0.09 \\ 0.15 & 0.30 & 0.19 & 0.36 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 0.339 & 0.307 & 0.183 & 0.171 \end{bmatrix}. \end{aligned}$$

After 2 weeks, there are approximately 34% discharged, 30% ambulatory, 18% bedridden and 17% dead patients.

Now, the characteristic equation of P is

$$\begin{aligned} |P - \lambda I| &= 0 \\ \Rightarrow \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0.3 & 0.4 & 0.3 & 0 \\ 0 & 0.5 & 0.2 & 0.3 \\ 0 & 0 & 0 & 1 \end{vmatrix} &= 0 \\ \Rightarrow (1-\lambda)^2(\lambda^2 - 0.6\lambda - 0.07) &= 0 \\ \Rightarrow \lambda &= 1,\ 1,\ 0.7,\ -0.1. \end{aligned}$$

Since $\lambda_1 = 1$ (multiplicity 2), $\lambda_2 = 0.7$, $\lambda_3 = -0.1$ are the eigen values of P, so P is not regular.

The left eigen vectors for the double eigen value 1 are $[1 \;\; 0 \;\; 0 \;\; 0]$ and $[0 \;\; 0 \;\; 0 \;\; 1]$ which are linearly independent. Hence P is ergodic. Therefore,

$$L = \lim_{n\to\infty} P^n.$$

Now,

$$\begin{aligned}
\begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.3 & 0.4 & 0.3 & 0 \\ 0 & 0.5 & 0.2 & 0.3 \\ 0 & 0 & 0 & 1 \end{bmatrix} &= 0.7 \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \\
\Rightarrow (1-0.7)x_1 + 0.3x_2 &= 0 \\
(0.4-0.7)x_2 + 0.5x_3 &= 0 \\
0.3x_2 + (0.2-0.7)x_3 &= 0 \\
0.3x_3 + (1-0.7)x_4 &= 0 \\
\Rightarrow 0.3x_1 + 0.3x_2 &= 0 \\
0.3x_2 - 0.5x_3 &= 0 \\
0.3x_3 + 0.3x_4 &= 0.
\end{aligned}$$

Solving the above equations, we get

$$x_1 = -x_2 = -\frac{5}{3}x_3 = \frac{5}{3}x_4.$$

Let $x_4 = 3$. Then we get

$$x_1 = 5, \quad x_2 = -5, \quad x_3 = -3.$$

Thus,

$$\begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} = \begin{bmatrix} 5 & -5 & -3 & 3 \end{bmatrix}$$

is the eigen vector corresponding to $\lambda_2 = 0.7$.

Now,

$$\begin{aligned}
\begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.3 & 0.4 & 0.3 & 0 \\ 0 & 0.5 & 0.2 & 0.3 \\ 0 & 0 & 0 & 1 \end{bmatrix} &= -0.1 \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \\
\Rightarrow (1+0.1)x_1 + 0.3x_2 &= 0 \\
(0.4+0.1)x_2 + 0.5x_3 &= 0 \\
0.3x_2 + (0.2+0.1)x_3 &= 0 \\
0.3x_3 + (1+0.1)x_4 &= 0 \\
\Rightarrow 1.1x_1 + 0.3x_2 &= 0 \\
x_2 + x_3 &= 0 \\
0.3x_3 + 1.1x_4 &= 0.
\end{aligned}$$

Solving the equations, we get,

$$x_1 = -\frac{3}{11}x_2 = \frac{3}{11}x_3 = -x_4.$$

Taking $x_2 = 11$, we get

$$x_1 = -3, \quad x_2 = 11, \quad x_3 = 3, \quad x_4 = 3.$$

Thus, $[-3 \ \ 11 \ \ 3 \ \ 3]$ is the eigen vector corresponding to $\lambda_3 = -0.1$.

To make P diagonalizable, we consider

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 5 & -5 & -3 & 3 \\ -3 & 11 & 3 & 3 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & -0.1 \end{bmatrix}.$$

To find M^{-1}:

$$\begin{aligned}
[M : I] \quad &= \quad \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 5 & -5 & -3 & 3 & 0 & 0 & 1 & 0 \\ -3 & 11 & 3 & 3 & 0 & 0 & 0 & 1 \end{array}\right] \\
&\xrightarrow[R_3 \to R_3 - 5R_1]{R_2 \leftrightarrow R_4} \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ -3 & 11 & 3 & 3 & 0 & 0 & 0 & 1 \\ 0 & -5 & -3 & 3 & -5 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}\right] \\
&\xrightarrow{R_2 \to R_2 + 3R_1} \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 11 & 3 & 3 & 3 & 0 & 0 & 1 \\ 0 & -5 & -3 & 3 & -5 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}\right] \\
&\xrightarrow[R_3 \to R_3 - 3R_4]{R_2 \to R_2 - 3R_4} \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 11 & 3 & 0 & 3 & -3 & 0 & 1 \\ 0 & -5 & -3 & 0 & -5 & -3 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}\right] \\
&\xrightarrow{R_2 \to \frac{1}{11} R_2} \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & \frac{3}{11} & 0 & \frac{3}{11} & -\frac{3}{11} & 0 & \frac{1}{11} \\ 0 & -5 & -3 & 0 & -5 & -3 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}\right] \\
&\xrightarrow{R_3 \to R_3 + 5R_2} \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & \frac{3}{11} & 0 & \frac{3}{11} & -\frac{3}{11} & 0 & \frac{1}{11} \\ 0 & 0 & -\frac{18}{11} & 0 & -\frac{40}{11} & -\frac{48}{11} & 1 & \frac{5}{11} \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}\right] \\
&\xrightarrow[R_3 \to -\frac{11}{18} R_3]{R_2 \to R_2 + \frac{1}{6} R_3} \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -\frac{1}{3} & -1 & \frac{1}{6} & \frac{1}{66} \\ 0 & 0 & 1 & 0 & \frac{20}{9} & \frac{8}{3} & -\frac{11}{18} & -\frac{5}{18} \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}\right].
\end{aligned}$$

Thus,

$$M^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{3} & -1 & \frac{1}{6} & \frac{1}{66} \\ \frac{20}{9} & \frac{8}{3} & -\frac{11}{18} & -\frac{5}{18} \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Thus,

$$\begin{aligned}\lim_{n\to\infty} P^n = L &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{3} & -1 & \frac{1}{6} & \frac{1}{66} \\ \frac{20}{9} & \frac{8}{3} & -\frac{11}{18} & -\frac{5}{18} \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 5 & -5 & -3 & 3 \\ -3 & 11 & 3 & 3 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{3} & -1 & 0 & 0 \\ \frac{20}{9} & \frac{8}{3} & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 5 & -5 & -3 & 3 \\ -3 & 11 & 3 & 3 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{3} & 0 & 0 & -1 \\ \frac{20}{9} & 0 & 0 & \frac{8}{3} \\ 0 & 0 & 0 & 1 \end{bmatrix}.\end{aligned}$$

(b) Thus, the status of the patients over the long run is

$$\begin{aligned}X^{(\infty)} &= X^{(0)}L \\ &= \begin{bmatrix} 0 & 0.7 & 0.3 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{3} & 0 & 0 & -1 \\ \frac{20}{9} & 0 & 0 & \frac{8}{3} \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{13}{30} & 0 & 0 & \frac{1}{10} \end{bmatrix} = \begin{bmatrix} 0.43 & 0 & 0 & 0.1 \end{bmatrix}.\end{aligned}$$

Therefore, over the long run, there are 43% discharged patients and 10% patients die. No ambulatory or bedridden patients remain in the geriatric ward.

■

Example 14.1.13. The training programme for production supervisors at a particular company consists of two phases. Phase 1, which involves 3 weeks of classroom work, is followed by Phase 2, which is a 3 week apprenticeship program under the direction of working supervisors. From past experience, the company expects only 60% of those beginning classroom training to be graduated into the apprenticeship phase, with the remaining 40% dropped completely from the training program. Of those who make it to the apprenticeship phase, 70% are graduated as supervisors, 10% are asked to repeat the second phase, and 20% are dropped completely from the program. How many supervisors can the company expect from its current training programme if it has 45 people in the classroom phase and 21 people in the apprenticeship phase?

Solution. We consider one time period to be 3 weeks and define states 1 through 4 as the conditions of being dropped, a classroom trainee, an apprentice, and a supervisor, respectively. If we assume that discharged individuals never re-enter the training programme and that supervisors remain supervisors, then the transition probabilities are given by the Markov chain

$$P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.4 & 0 & 0.6 & 0 \\ 0.2 & 0 & 0.1 & 0.7 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

. Since there are $45 + 21 = 66$ people in the training programme currently, so the initial probability vector is given by

$$X^{(0)} = \left[0,\ \frac{45}{66},\ \frac{21}{66},\ 0\right].$$

We have from example 14.1.6,

$$\lim_{n\to\infty} P^n = L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{8}{15} & 0 & 0 & \frac{7}{15} \\ \frac{2}{9} & 0 & 0 & \frac{7}{9} \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

$$\begin{aligned} X^{(\infty)} &= X^{(0)}L \\ &= \begin{bmatrix} 0 & \frac{45}{66} & \frac{21}{66} & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{8}{15} & 0 & 0 & \frac{7}{15} \\ \frac{2}{9} & 0 & 0 & \frac{7}{9} \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 0.4343 & 0 & 0 & 0.5657 \end{bmatrix}. \end{aligned}$$

Eventually, 43.43% of those currently in training (or about 29 people) will be dropped from the programme and 56.67% (or about 37 people) will become supervisors. ■

Example 14.1.14. Solve the previous problem if all 66 people are currently in the classroom phase of training programme.

Solution. Here, $X^{(0)} = [0\ \ 1\ \ 0\ \ 0]$. Thus,

$$\begin{aligned} X^{(\infty)} &= X^{(0)}L \\ &= \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{8}{15} & 0 & 0 & \frac{7}{15} \\ \frac{2}{9} & 0 & 0 & \frac{7}{9} \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{8}{15} & 0 & 0 & \frac{7}{15} \end{bmatrix}. \end{aligned}$$

Thus, $\frac{8}{15} \times 66 \simeq 35$ people will ultimately drop from the program and the remaining $66 - 35 = 31$ people eventually become supervisors. ■

Unit 15

Course Structure

- Geometric programming
- General form of GP (Unconstrained GP)(Primal Problem)

15.1 Geometric Programming

We shall focus our attention on a rather interesting technique called *Geometric Programming* for solving a special type of non-linear programming problem. This technique is initially derived from inequalities rather than the calculus and its extension. This technique was given the name geometric programming because the geometric arithmetic mean inequality was the basis of its development. The advantage here is that it is usually much simpler to work with the dual problem than the primal problem. Geometric programming derives its name from the fact that it is based on the certain geometric concept such as orthogonality and arithmetic geometric inequality. It was developed in early 1960's by Duffin, Peterson and Zener for solving the class of optimization problem that involve special type of functions called posynomial (positive+ polynomial).

A real expression of the form

$$C_j \prod_{i=1}^{n} (x_i)^{a_{ij}}$$

where c_j, a_{ij} are real and $X = (x_1, x_2, \ldots, x_m)^T > 0$ is called monomial in X.

Example: $5.7x_1^3 x_2 - 4x_3^{2.5}$ is a monomial.

Posynomial and Signomial: A generalised polynomial that consist of a finite number of monomials such as

$$f(x) = \sum_{j=1}^{n} C_j \prod_{i=1}^{m} (x_i)^{a_{ij}}$$

is said to be posynomial if all the coefficients C_j are positive; is called the signomial if the coefficients C_j are negative.

The G.P approach instead of solving a non-linear programming problem first finds the optimal value of the objective function by solving its dual problem and then determines an optimal solution to the given NLPP from the optimal solution of the dual.

15.1.1 General form of G.P (Unconstrained G.P) (Primal Problem)

$$\begin{aligned}&\min f(x) = \sum_{j=1}^{n} c_j u_j(x)\\ &\text{such that } \; x_i \geq 0 \quad \text{with } \; c_j > 0\\ &\text{and } \; u_j(x) = \prod_{i=1}^{n} (x_j)^{a_{ij}},\end{aligned}$$

where a_{ij} may be any real number.

15.1.2 Necessary conditions for optimality

The necessary conditions for optimality can be obtained by taking partial derivatives with respect to each x_r and equating the result with 0. Thus

$$\frac{\partial f(x)}{\partial x_r} = \sum_{j=1}^{n} c_j \frac{\partial u_j(x)}{\partial x_r} = 0$$

But,

$$\frac{\partial}{\partial x_r} u_j(x) = \frac{a_{rj}}{x_r} u_j(x).$$

Putting this result in the previous equation, we get,

$$\frac{\partial f(x)}{\partial x_r} = \frac{1}{x_r} \sum_{j=1}^{n} a_{rj} c_j u_j(x) = 0.$$

Let, $f^*(x)$ be the minimum value of $f(x)$. Since, each x_r and c_j is positive, therefore $f^*(x)$ will also be positive. Defining $\dfrac{\partial f(x)}{\partial x_r}$ by $f^*(x)$ we get,

$$\sum_{j=1}^{n} \frac{a_{rj} c_j u_j(x)}{f^*(x)} = 0.$$

Now, we take a simple transformation of variable as

$$y_j = \frac{c_j u_j(x)}{f^*(x)}, \quad j = 1, 2, \ldots, n.$$

Using this transformation, the necessary conditions for local minimum becomes,

$$\sum_{j=1}^{n} a_{rj} y_j = 0; \quad r = 1, 2, \ldots, m. \tag{15.1.1}$$

Thus, due to the definition of y_j, we obtain

$$\sum_{j=1}^{n} y_j = \frac{1}{f^*(x)} \sum_{j=1}^{n} c_j u_j(x) = 1. \tag{15.1.2}$$

At the optimal solution, conditions (15.1.1) and (15.1.2) are the necessary conditions for optimality of non-linear function and also known as orthogonality and normality conditions respectively. This condition give a unique value of y_j for $m+1=n$ and all equations are independent but for $n > (m+1)$, the value of y_j no longer remains independent.

[Degree of G.P difficulty (D.D) of G.P is equal to number of terms in G.P -(1 + number of variables in G.P]

$$\therefore D.D = n - (m+1), \quad (> 0 \quad \text{infinite solution}).$$

Conditions (15.1.1) and (15.1.2) can be expressed as

$$AY = b,$$

where

$$A = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Thus, we require to form the normality and orthogonality condition $AY = B$. This means that the original NLP problem is reduced to one of finding the set of values of Y that satisfy this linear non-homogeneous equation. Hence, to determine the unique value of y_j for the purpose of minimizing effect.

(i) Rank (A, b) > Rank (A), there will be no solution, where (A, b) denote the augmented matrix.

(ii) Rank (A, b) = Rank (A), then a unique solution.

(iii) Rank $(A) < n$, i.e $n > m+1$, that is infinite number of solutions exist.

To find the minimum value of $\mathbf{f(x)}$

At the optimal solution we know that

$$f^*(x) = \frac{c_j u_j(x)}{y_j} = \frac{1}{y_j} c_j \prod_{i=1}^{n} (x_i)^{a_{ij}}$$

Raising both side to power of y_j and taking the product we get,

$$\sum_{j=1}^{n} \left\{ f^*(x) \right\}^{y_j} = \prod_{j=1}^{n} \left\{ \frac{1}{y_j} c_j \prod_{i=1}^{n} (x_i)^{a_{ij}} \right\}^{y_j},$$

Now, since $\sum_{j=1}^{n} y_j = 1$, therefore

$$\prod_{j=1}^{n} \{f^*(x)\}^{y_j} = \left[f^*(x) \right]^{\sum_{j=1}^{n} y_j} = f^*(x)$$

In R.H.S of the above equation we have

$$\begin{aligned}\prod_{j=1}^{n}\left[\left(\frac{c_j}{y_j}\right)\prod_{i=1}^{m}(x_i)^{a_{ij}}\right]^{y_j} &= \prod_{j=1}^{n}\left(\frac{c_j}{y_j}\right)^{y_j}\prod_{j=1}^{n}\left\{\prod_{i=1}^{m}(x_i)^{a_{ij}}\right\}^{y_j}\\ &= \prod_{j=1}^{n}\left(\frac{c_j}{y_j}\right)^{y_j}\prod_{i=1}^{m}(x_i)^{\sum_{j=1}^{n}a_{ij}y_j}\\ &= \prod_{j=1}^{n}\left(\frac{c_j}{y_j}\right)^{y_j}\prod_{i=1}^{m}(x_i)^{0} \text{ [By Eq. (15.1.1)]}\end{aligned}$$

Thus,

$$\min f(x) = f^*(x) = \prod_{j=1}^{n}\left(\frac{c_j}{y_j}\right)^{y_j} \quad \text{and}$$

$$\therefore f(x) \geq \prod_{j=1}^{n}\left(\frac{c_j}{y_j}\right)^{y_j}.$$

where y_j must satisfy the orthogonality and normality conditions. For the given value of f^* and unique value of y_j, the solution to a set of equations can be obtained from

$$c_j\prod_{i=1}^{m}(x_i)^{a_{ij}} = y_j f^*(x).$$

Dual Problem:

$$\begin{aligned}\max g(y) &= \prod_{j=1}^{n}\left(\frac{c_j}{y_j}\right)^{y_j}\\ \text{subject to} \quad &\sum_{j=1}^{n}a_{ij}y_j = 0\\ \text{and} \quad &\sum_{j=1}^{n}y_j = 1\\ &y_j \geq 0.\end{aligned}$$

Theorem 15.1.1. If x is a feasible solution vector of the unconstraint of a primal geometric programming and y is a feasible solution vector for DP (Dual problem), then

$$f(x) \geq g(y). \qquad \text{(Primal Dual inequality)}$$

Proof. The expression for $f(x)$ can be written as

$$f(x) = \sum_{j=1}^{n}\frac{C_j\prod_{i=1}^{m}(x_i)^{a_{ij}}}{y_j}.$$

Here, weights are $y_1, y_2, \ldots, y_n$ and the positive terms are

$$\frac{C_1\prod_{i=1}^{m}(x_i)^{a_{i1}}}{y_1}, \quad \frac{C_2\prod_{i=1}^{m}(x_i)^{a_{i2}}}{y_2}, \ldots, \quad \frac{C_n\prod_{i=1}^{m}(x_i)^{a_{in}}}{y_n}.$$

Now, applying weighted Arithmetic Mean- Geometric mean inequality,

$$\left(\frac{\dfrac{C_1\prod_{i=1}^{m}(x_i)^{a_{i1}}}{y_1}\cdot y_1+\dfrac{C_2\prod_{i=1}^{m}(x_i)^{a_{i2}}}{y_2}\cdot y_2+\cdots+\dfrac{C_n\prod_{i=1}^{m}(x_i)^{a_{in}}}{y_n}\cdot y_n}{y_1+y_2+\cdots+y_n}\right)^{y_1+y_2+\cdots+y_n}$$

$$\geq \left(\frac{C_1\prod_{i=1}^{m}(x_i)^{a_{i1}}}{y_1}\right)^{y_1}\cdot\left(\frac{C_2\prod_{i=1}^{m}(x_i)^{a_{i2}}}{y_2}\right)^{y_2}\cdots\left(\frac{C_n\prod_{i=1}^{m}(x_i)^{a_{in}}}{y_n}\right)^{y_n}$$

$$\text{or,}\quad f(x)\geq\prod_{j=1}^{n}\left(\frac{C_j\prod_{i=1}^{m}(x_i)^{a_{ij}}}{y_j}\right)^{y_j}\qquad[\text{since}\quad y_1+y_2+\cdots+y_n=1\ \text{ for normality condition}]$$

$$\text{or,}\quad f(x)\geq\prod_{j=1}^{n}\left(\frac{C_j}{y_j}\right)^{y_j}\prod_{i=1}^{m}(x_i)^{\sum_{j=1}^{n}a_{ij}y_j}\tag{15.1.3}$$

$$\text{or,}\quad f(x)\geq\prod_{j=1}^{n}\left(\frac{C_j}{y_j}\right)^{y_j}\qquad\left[\sum_{i=1}^{n}a_{ij}y_j=0,\ \text{ orthogonality condition}\right]$$

$$\text{or,}\quad f(x)\geq g(y).$$

For constraint, after above

$$g_i(x)=\sum_{r=1}^{P(i)}y_{ir}\left(\frac{C_{ir}\prod_{i=1}^{n}(x_i)^{a_{irj}}}{y_{ir}}\right)$$

Applying weighted arithmetic mean geometric mean inequality, we have

$$\left(\frac{g_i(x)}{\sum_{r=1}^{P(i)}y_{ir}}\right)^{\sum_{r=1}^{P(i)}y_{ir}}\geq\prod_{i=1}^{m}\prod_{r=1}^{P(i)}\left(\frac{C_{ir}\prod_{i=1}^{n}(x_i)^{a_{irj}}}{y_{ir}}\right)^{y_{ir}}$$

$$(g_i(x))^{\sum_{r=1}^{P(i)}y_{ir}}\geq\prod_{i=1}^{m}\prod_{r=1}^{P(i)}\left(\frac{C_{ir}}{y_{ir}}\right)^{y_{ir}}\prod_{i=1}^{n}(x_i)^{\sum_{r=1}^{P(i)}a_{irj}y_{ir}}\left(\sum_{r=1}^{P(i)}y_{ir}\right)^{y_{ir}}.$$

Since $g_i(x)\leq 1$ (constraint), so,

$$1\geq(g_i(x))^{\sum_{r=1}^{P(i)}y_{ir}}.$$

Hence,

$$1 \geq \prod_{i=1}^{m}\prod_{r=1}^{P(i)}\left(\frac{C_{ir}}{y_{ir}}\right)^{y_{ir}} \prod_{i=1}^{n}(x_i)^{\sum_{r=1}^{P(i)} a_{irj}y_{ir}} \left(\sum_{r=1}^{P(i)} y_{ir}\right)^{y_{ir}}. \tag{15.1.4}$$

Multiplying (15.1.3) and (15.1.4), we have

$$f(x) \geq \prod_{j=1}^{n}\left(\frac{C_j}{y_j}\right)^{y_j} \prod_{1=1}^{m}\left[\prod_{r=1}^{P(i)}\left(\frac{C_{ir}}{y_{ir}}\right)^{y_{ir}} \left(\sum_{r=1}^{P(i)} y_{ir}\right)^{y_{ir}}\right] (x_i)^{\sum_{i=1}^{n} a_{ij}y_j + \sum_{i=1}^{m}\sum_{r=1}^{P(i)} a_{irj}y_{ir}}.$$

Using orthogonality condition,

$$\sum_{i=1}^{n} a_{ij}y_j + \sum_{i=1}^{m}\sum_{r=1}^{P(i)} a_{irj}y_{ir} = 0.$$

Thus, we have,

$$f(x) \geq \prod_{j=1}^{n}\left(\frac{C_j}{y_j}\right)^{y_j} \prod_{1=1}^{m}\left[\prod_{r=1}^{P(i)}\left(\frac{C_{ir}}{y_{ir}}\right)^{y_{ir}} \left(\sum_{r=1}^{P(i)} y_{ir}\right)^{y_{ir}}\right]$$

or, $\quad f(x) \geq g(y).$

□

Example 15.1.2. Solve the following NLPP by geometric programming technique.

$$\min z = 7x_1x_2^{-1} + 3x_2x_3^{-2} + 5x_1^{-3}x_2x_3 + x_1x_2x_3$$
$$x_1, x_2, x_3 \geq 0$$

Solution.

$$A = \begin{bmatrix} 1 & 0 & -3 & 1 \\ -1 & 1 & 1 & 1 \\ 0 & -2 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

and we get $AY = b$ with

$$y_1 = \frac{1}{2},\ y_2 = \frac{1}{6},\ y_3 = \frac{5}{24},\ y_4 = \frac{3}{24},\ f^*(x) = \frac{761}{50}$$
$$x_1^* = 1.315,\ x_2^* = 1.21,\ x_3^* = 1.2$$

Now $AY = b$ gives

$$\begin{bmatrix} 1 & 0 & -3 & 1 \\ -1 & 1 & 1 & 1 \\ 0 & -2 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

which leads to the following system of equations

$$y_1 - 3y_3 + y_4 = 0 \tag{15.1.5}$$
$$-y_1 + y_2 + y_3 + y_4 = 0 \tag{15.1.6}$$
$$-2y_2 + y_3 + y_4 = 0 \tag{15.1.7}$$
$$y_1 + y_2 + y_3 + y_4 = 1 \tag{15.1.8}$$

Now, (15.1.6)-(15.1.8) gives

$$-y_1 + y_2 + y_3 + y_4 - y_1 - y_2 - y_3 - y_4 = -1$$
$$\Rightarrow -2y_1 = -1 \Rightarrow y_1 = \frac{1}{2}$$

Now, (15.1.6)-(15.1.7) gives

$$-y_1 + y_2 + y_3 + y_4 + 2y_2 - y_3 - y_4 = 0$$
$$\Rightarrow -y_1 + 3y_2 = 0 \Rightarrow 3y_2 = y_1$$
$$\Rightarrow 3y_2 = \frac{1}{2} \Rightarrow y_2 = \frac{1}{6}.$$

Now, (15.1.5)-(15.1.7) gives

$$y_1 - 3y_3 + y_4 + 2y_2 - y_3 - y_4 = 0$$
$$\Rightarrow y_1 + 2y_2 - 4y_3 = 0 \Rightarrow 4y_3 = y_1 + 2y_2$$
$$\Rightarrow 4y_3 = \frac{1}{2} + \frac{1}{3} \Rightarrow 4y_3 = \frac{5}{6} \Rightarrow y_3 = \frac{5}{24}.$$

Now,

$$\begin{aligned} y_4 &= 1 - (y_1 + y_2 + y_3) \\ &= 1 - \left(\frac{1}{2} + \frac{1}{6} + \frac{5}{24}\right) \\ &= 1 - \frac{12 + 4 + 5}{24} \\ &= 1 - \frac{21}{24} \\ &= \frac{3}{24} \end{aligned}$$

$$\therefore y_1 = \frac{1}{2}, \quad y_2 = \frac{1}{6}, \quad y_3 = \frac{5}{24}, \quad y_4 = \frac{3}{24}$$

$$\begin{aligned} f^*(x) &= \left(\frac{7}{1/2}\right)^{1/2} \times \left(\frac{3}{1/6}\right)^{1/6} \times \left(\frac{5}{5/24}\right)^{5/24} \times \left(\frac{1}{3/24}\right)^{3/24} \\ &= (14)^{1/2} \times (18)^{1/6} \times (24)^{5/24} \times (8)^{3/24} \\ &= 3.74 \times 1.62 \times 1.94 \times 1.297 \\ &= 15.245 \\ &= \frac{761}{50} \end{aligned}$$

Now

$$c_j \prod_{i=1}^{m} (x_i)^{a_{ij}} = y_j f^*(x)$$

$$\therefore 7x_1 x_2^{-1} = \frac{1}{2} \times \frac{761}{50}$$

$$\Rightarrow x_1 x_2^{-1} = \frac{761}{700} \tag{15.1.9}$$

$$\text{and} \quad 3x_2 x_3^{-2} = \frac{1}{6} \times \frac{761}{50}$$

$$\Rightarrow x_2 x_3^{-2} = \frac{761}{900}$$

$$5x_1^{-3} x_2 x_3 = \frac{5}{24} \times \frac{761}{50}$$

$$\Rightarrow x_1^{-3} x_2 x_3 = \frac{761}{1200} \tag{15.1.10}$$

$$\text{and} \quad x_1 x_2 x_3 = \frac{3}{24} \times \frac{761}{50}$$

$$\Rightarrow x_1 x_2 x_3 = \frac{761}{400} \tag{15.1.11}$$

Now (15.1.10) and (15.1.11) gives

$$\frac{x_1^{-3} x_2 x_3}{x_1 x_2 x_3} = \frac{761/1200}{761/400}$$

$$\Rightarrow x_1^{-4} = \frac{1}{3}$$

$$x_1 = \left(\frac{1}{3}\right)^{-1/4} = 3^{1/4} = 1.316.$$

$$\therefore x_1^* = 1.316$$

Now from, (15.1.9) we get

$$x_1 x_2^{-1} = \frac{761}{700}$$

$$\Rightarrow x_2^{-1} = \frac{761}{700} \times \frac{1}{x_1}$$

$$\Rightarrow x_2 = \frac{700}{761} \times x_1$$

$$\Rightarrow x_2 = \frac{700}{761} \times 1.3616$$

$$\Rightarrow x_2 = 1.21$$

$$\therefore x_2^* = 1.21$$

Now, from (15.1.9) we get

$$x_2 x_3^{-2} = \frac{761}{900}$$

$$x_3^2 = \frac{900}{761} x_2$$

$$x_3 = \sqrt{\frac{900}{761}} \times \sqrt{1.21} = 1.2$$

$$\therefore x_3^* = 1.2$$

■

Example 15.1.3. Solve the following NLPP by the geometric programming.

$$\min f(x) = 5x_1x_2^{-1} + 2x_1^{-1}x_2 + 5x_1 + x_2^{-1}; \quad x_1, x_2 \geq 0$$

Solution. The given function may be written as

$$f(x) = 5x_1x_2^{-1} + 2x_1^{-1}x_2 + 5x_1^1x_2^0 + x_1^0x_2^{-1}.$$
$$(c_1, c_2, c_3, c_4) = (5, 2, 5, 1)$$

The orthogonality and normality conditions are given by

$$\begin{bmatrix} 1 & -1 & 1 & 0 \\ -1 & 1 & 0 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Since $n > m + 1$, this equations do not give y_j directly. Solving for y_1, y_2 and y_3 in terms of y_4 we get,

$$\begin{bmatrix} 1 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 0 \\ y_4 \\ 1 - y_4 \end{bmatrix}$$

$$\text{or } \; y_1 = (1 - 3y_4)/2 = 0.8(1 - 3y_4); \quad y_2 = 0.5(1 - y_4); \quad y_3 = y_4.$$

The corresponding dual problem may be written as

$$\max f(y) = \left[\frac{5}{0.5(1-3y_4)}\right]^{0.5(1-3y_4)} \left[\frac{2}{0.5(1-y_4)}\right]^{0.5(1-3y_4)} \left[\frac{5}{y_4}\right]^{y_4} \left[\frac{1}{y_4}\right]^{y_4}$$

Since, maximization of $f(y)$ is equivalent to $\log f(y)$, taking $\log$ both sides we have

$$\begin{aligned} \log f(y) = {} & 0.5(1-3y_4)\{\log 10 - \log(1-3y_4)\} + 0.5(1-y_4)\{\log 4 - \log(1-y_4)\} \\ & + y_5(\log 5 - \log y_4) + y_4\{\log 1 - \log y_4\} \end{aligned} \tag{15.1.12}$$

The value of y_4 maximizing $\log f(y)$ must be unique, because the primal problem has a unique minimum. Differentiating (15.1.12) with respect to y_4 and equating to zero, we have

$$\begin{aligned} \frac{\partial}{\partial y_4} f(y) \quad = \quad & -\frac{3}{2}\log 10 - \left\{-\frac{3}{2} + \left(-\frac{3}{2}\right)\log(1-3y_4)\right\} \\ & -\frac{1}{2}\log 4 - \left\{-\frac{1}{2} + \left(-\frac{1}{2}\right)\log(1-y_4)\right\} \\ & + \log 5 - \{1 + \log y_4\} + \log 1 - \{1 + \log y_4\} = 0 \end{aligned}$$

Then after simplification, we have

$$-\log\left\{\frac{2 \times 10^{3/2}}{5}\right\} + \log\left\{\frac{(1-3y_4)^{3/2}(1-y_4)^{1/2}}{y_4^2}\right\} = 0.$$
$$\Rightarrow \frac{\sqrt{(1-3y_4)^3(1-y_4)}}{y_4^2} = 12.6$$

After solving we have $y_4^* = 0.16$. Hence

$$y_1^* = 0.26, \quad y_2^* = 0.42, \quad y_3^* = 0.16$$

$$\begin{aligned} \text{The value of } f^*(y) &= f^*(x) \\ &= \left(\frac{5}{0.26}\right)^{0.26} \left(\frac{2}{0.42}\right)^{0.42} \left(\frac{5}{0.16}\right)^{0.16} \left(\frac{1}{0.16}\right)^{0.160} \\ &= 9.661 \end{aligned}$$

$$u_1 = y_1^* f^*(x), \quad u_2 = y_2 f^*(x), \quad u_3 = y_3 f^*(x), \quad u_4 = y_4 f^*(x)$$

$$\begin{aligned} 5x_1 &= 0.16 \times 9.661 \\ \Rightarrow x_1^* &= \frac{0.16 \times 9.661}{5} = 0.309 \end{aligned}$$

and

$$\begin{aligned} x_2^{-1} &= 0.42 \times 9.661 \\ \Rightarrow x_2^* &= \frac{1}{0.16 \times 9.661} = 0.647 \end{aligned}$$

■

Unit 16

Course Structure

- Constraint Geometric Programming Problem

16.1 Constraint Geometric Programming Problem

$$\min z = f(x)$$
$$\text{such that } \; g_i(x) = \sum_{r=1}^{P(i)} c_{ij} u_{ir}(x) = 1, \quad i = 1, 2, \ldots, M.$$

where $P(i)$ denotes the number of terms in the i-th constraint and $u_{ir}(x) = \prod_{j=1}^{n} (x_i)^{a_{irj}}$.

Forming Lagrange function to obtain normality and orthogonality condition,

$$F(x, \lambda) = f(x) + \sum_{i=1}^{M} \lambda_i [g_i(x) - 1]$$

and require the conditions,

(i) $\dfrac{\partial F}{\partial x_t} = \dfrac{\partial f(x)}{\partial x_t} + \sum_{i=1}^{M} \lambda_i \dfrac{\partial g_i(x)}{\partial x_t} = 0.$

(ii) $\dfrac{\partial F}{\partial \lambda_i} = g_i(x) - 1 = 0; \; i = 1, 2, \ldots, M.$

So, long as right hand side in the second constraint $g_i(x) = 1$, it can be obtained in this form by simple transformation. However, $g_i(x) = 0$ is not admissible because solution space required $x > 0$. Considering once again condition (i), we have

$$\frac{\partial F}{\partial x_t} = \sum_{j=1}^{n} \frac{c_j a_{tj} c_j(x)}{x_t} + \sum_{i=1}^{M} \lambda_i \left[\sum_{r=1}^{P(i)} \frac{c_{ir} a_{irt} u_{ir}(x)}{x_t} \right].$$

Introducing variables y_j for objective and y_{ir} for constraints as follows:

$$y_j = \frac{c_j u_j(x)}{f^*(x)} \quad \text{and} \quad y_{ir} = \frac{\lambda_i c_{ir} u_{ir}(x)}{f^*(x)}$$

By substituting the values of y_j and y_{ir} in $\dfrac{\partial F}{\partial x_t} = 0$, we obtain the orthogonality conditions and normality condition as

$$\sum_{j=1}^{n} a_{tj} y_j + \sum_{i=1}^{M} \sum_{r=1}^{P(i)} a_{irt} y_{ir} = 0; \;\; t = 1, 2, \ldots, n. \quad \text{(Orthogonality Conditions)}$$

$$\sum_{j=1}^{n} y_j = 1 \quad \text{(Normality Condition)}$$

We have seen in earlier discussion that y_j were all positive, because $y_j = \dfrac{c_j u_j(x)}{f^*(x)} > 0$. However, in the equality constraint case, y_j are again positive. But, y_{ir} may be negative because λ_i need not be non-negative. To formulate a dual function it is desirable to all $y_{ir} > 0$. But if one of the y_{ir} is negative, then its sign can be reversed by writing the term in the Lagrange function as $\lambda_q\{1 - g_q(x)\}$. Once again normality and orthogonality conditions can be derived by solving a system of linear equations

$$\sum_{j=1}^{n} a_{tj} y_j$$

When these equations have a unique solution, the optimal of the original problem can be obtained from the definition of y_j and y_{ir} in terms of $f^*(x)$ and x. In case, these equations have an infinite number of solution, we tend to maximize the dual function given by

$$\max f(y) = \prod_{j=1}^{n} \left(\frac{c_j}{y_j}\right)^{y_j} \prod_{i=1}^{M} \left[\prod_{r=1}^{P(i)} \left(\frac{c_{rj}}{y_{ij}}\right)^{y_{rj}}\right] \prod_{i=1}^{M} (v_i)$$

where $v_i = \sum_{r=1}^{P(i)} y_{ir}$ such that the orthogonality and normality constraints.

In the above functions the constraints are linear and therefore it is easy to obtain the optimal solution. Moreover, we may also work with $\log$ of the dual function which is linear in the variable $\delta_i = \log y_j$ and $\delta_{ir} = \log y_{ir}$.

Example 16.1.1. Solve the following NLPP by G.P.

$$\begin{aligned} &\min f(x) = 2x_1 x_2^{-3} + 4x_1^{-1} x_2^{-2} + \frac{32}{3} x_1 x_2 \\ &\text{such that} \quad x_1^{-1} x_2^2 = 0 \\ &\qquad\qquad x_1, x_2 \geq 0. \end{aligned}$$

Solution. Given problem derive as

$$\begin{aligned} &\min f(x) = 2x_1 x_2^{-3} + 4x_1^{-1} x_2^{-2} + \frac{32}{3} x_1 x_2 \\ &\text{such that } \; 0.1 x_1^{-1} x_2^2 = 1 \\ &x_1, x_2 \geq 0. \end{aligned}$$

Dual problem:

$$\max f(y) = \left(\frac{2}{y_1}\right)^{y_1}\left(\frac{4}{y_2}\right)^{y_2}\left(\frac{32}{3y_3}\right)^{y_3}\left(\frac{0.1}{y_4}\right)^{y_4}(y_4)^{y_4}$$

such that

$$y_1 + y_2 + y_3 = 1$$
$$y_1 - y_2 + y_3 - y_4 = 0$$
$$-3y_1 - 2y_2 + y_3 + 2y_4 = 0$$

Expressing each of the variable in the objective function in terms of y_1, we get

$$\max f(y_1) = \left(\frac{2}{y_1}\right)^{y_1}\left(\frac{4}{1-\frac{4}{3}y_1}\right)^{1-\frac{4}{3}y_1}\left(\frac{32}{y_1}\right)^{\frac{1}{3}y_1}(0.1)^{\frac{8}{3}y_1-1}$$

where

$$y_2 = 1 - \frac{4}{3}y_1$$
$$y_3 = \frac{y_1}{3}$$
$$y_4 = \frac{8}{3}y_1 - 1$$

Taking $\log$ both sides of $f(y_1)$ and differentiating with respect to y_1, we have,

$$\begin{aligned} F(y_1) &= \log f(y_1) \\ &= y_1 \log\left(\frac{2}{y_1}\right) + \left\{1 - \left(\frac{4}{3}\right)y_1\right\}\log 4 - \log\left(1 - \frac{4}{3}y_1\right) \\ &\quad \frac{y_1}{3}\left\{\log 32 - \log y_1\right\} + \left(\frac{8}{3}y_1 - 1\right)\log(0.1) \end{aligned}$$

Now,

$$\frac{dF}{dy_1} = \log\left(\frac{2}{y_1}\right) + 2 - \left(\frac{16}{3}\right)y_1 + \log\left(\frac{32}{y_1}\right) + \frac{8}{3}\log(0.1) = 0$$
$$\Rightarrow y_1 = 0.662$$

The values of the other variables are

$$y_1 = 0.662, \quad y_2 = 0.217, \quad y_3 = 0.221, \quad y_4 = 0.766$$

Using the relation $y_j = \dfrac{c_j u_j}{f^*(x)}$ we obtain

$$y_1 = \frac{c_1 u_1}{f^*(x)} = \frac{2x_1 x_2^{-1}}{f^*(x)}$$
$$y_2 = \frac{c_2 u_2}{f^*(x)} = \frac{4x_1^{-1}x_2^{-1}}{f^*(x)}$$
$$y_3 = \frac{c_3 u_3}{f^*(x)} = \frac{32x_1 x_2}{3f^*(x)}$$
$$y_4 = \frac{c_4 u_4}{f^*(x)} = \frac{x_1^{-1}x_2^2}{f^*(x)}$$

■

Exercise 16.1.2. Solve the following NLPP by G.P.

1.

$$\min f(x) = 5x_1x_2^{-1}x_3^2 + x_1^{-2}x_2^{-1} + 10x_2^2 + 2x_1^{-1}x_2x_3^{-2}$$
$$x_1, x_2, x_3 \geq 0$$

Answer: $x_1 = 1.26,\ x_2 = 0.41,\ x_3 = 0.59$ and $\min f(x) = 10.28$

2.

$$\min f(x) = 2x_1 + 4x_2 + \frac{10}{x_1x_2}$$
$$x_1, x_2 \geq 0$$

Answer: $x_1 = 14.1,\ x_2 = 23$ and $\min f(x) = 112.9$

3.

$$\min z = \frac{3x_1}{x_2} + \frac{x_2^2}{x_1} + x_1^2x_2$$
such that
$$\frac{1}{4}x_1^2x_2^{-1} + \frac{1}{9}x_2x_1 = 1$$
$$2\left(\frac{1}{x_1^2}\right) + 4\left(\frac{x_2}{x_1^2}\right) = 2$$
$$x_1, x_2 \geq 0.$$

Unit 17

Course Structure

- Inventory Control/Problem/Model
- The Economic Order Quantity (EOQ) model without shortage

17.1 Inventory Control/Problem/Model

17.1.1 Production Management

In our daily lives, we observe that a small retailer knows roughly the demand of his customers in a month or a week or a day, and accordingly places orders on the wholesaler to meet the demand of his customer. But this is not the case with a manager of a big departmental store or a big retailer because the stocking in such cased depends upon various factors namely demand, time of ordering, lag between orders and actual receives etc. So, the real problem is to have a compromise between over stocking and under stocking. The study of such type of problems is known as material management or production management or inventory control. In broad sense, inventory may be defined as the stock of goods, commodities or other economic resources that are stored or reserved in order to ensure smooth and efficient running of business affairs. The inventory may be kept in any of the following forms:

(i) Raw-material inventory

(ii) Working process inventory

(iii) Finished good inventory

(iv) Inventory also includes furniture, machinery etc.

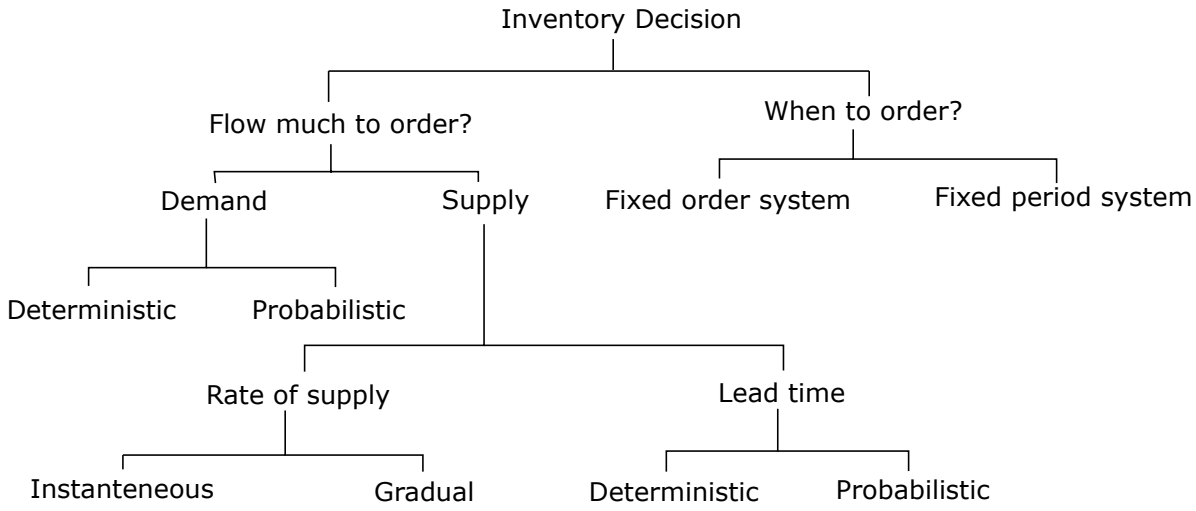
The term inventory may be classified in two main categories, viz.

(1) Direct Inventory

(2) Indirect Inventory

Indirect inventory includes those items which are necessarily required for manufacturing but do not become the component of finished products like oil, grease, lubricants, petrol, office materials, etc.

17.1.2 Inventory Decisions



Lead time: Time between placing an order an actual received.

17.1.3 Inventory related cost:

(1) **Holding Cost (**C_1 **or** C_n**)**: The cost associated with carrying or holding the goods in stock is known as holding cost or carrying cost, which is usually denoted by C_1 per unit of goods per unit time.

(2) **Shortage or stockout cost (**C_2 **or** C_s**)**: The penalty cost which is incurred as a result of running out of stock or shortage is known as shortage or stockout cost. It is usually denoted by C_2 per unit of goods for a specified period. This cost arises due to shortage of goods, sales may be lost, goodwill may be lost and so on.

(3) **Set up or ordering cost (**C_3 **or** C_0**)**: This includes the fixed cost associated with obtaining goods during placing of an order or purchasing or manufacturing or setting up a machinery before starting production. It is usually denoted by C_3 or C_0 per production run (cycle).

17.1.4 Why inventory is maintained?

Mathematically the problem of maintaining the inventory arises due to the fact that if a person decides to have a large stock, his holding cost C_1 increases but his shortage cost C_2 and set up cost C_3 decrease. On the other hand if he has small stock, his holding cost C_1 decreases but shortage cost C_2 and set up cost C_3 increase. Similarly, if he decides to order very frequently, the ordering cost increases when the other cost may decrease. So, it becomes necessary to have a compromise between over stocking and under stocking by making optimum decision by controlling value of some variables.

17.1.5 Variables in Inventory Problems

(i) Controlled variable: $q,\ t$

(ii) Uncontrolled variable: $C_1,\ C_2,\ C_3$, Demand (R), Lead time.

17.1.6 Some Notations

C_1 = Holding cost per quantity per unit time.
C_2 = Shortage cost per quantity per unit time.
C_3 = Set up cost per order.
R = Demand rate.
K = Production rate.
t = Scheduling time period which is variable.
t_p = Prescribed time period.
D = Total demand or annual demand.
q = Quantity already present in the beginning.
L = Lead time.

17.2 The Economic Order Quantity (EOQ) model without shortage

17.2.1 Model I(a): Economic lot size model with uniform demand

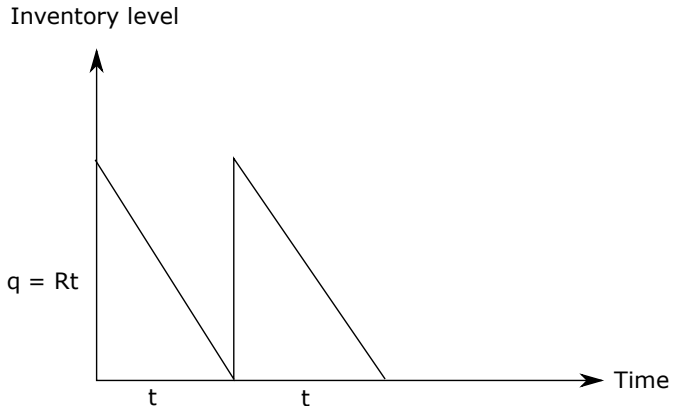
Assumptions:

(i) Demand is uniform at a rate R quantity units per unit time.

(ii) Lead time is zero.

(iii) Production rate is infinite, i.e., instantaneous.

(iv) Shortages are not allowed.



Let each production cycle be made at fixed interval t and therefore the quantity q already present in the beginning should be

$$q = Rt, \tag{17.2.1}$$

where R is a demand rate. Since, the stock in small time dt is $Rt\ dt$, therefore, the stock in total time t will be

$$\int_0^t R\, t\, dt = \frac{1}{2}Rt^2 = \frac{1}{2}qt.$$

Thus,

$$\text{The cost of holding inventory per production run} = C_1\frac{1}{2}qt = C_1\frac{1}{2}Rt^2 \tag{17.2.2}$$

$$\text{The set up cost} = C_3 \text{ per production run for interval } t.$$

$$\text{Total cost} = \frac{1}{2}C_1Rt^2 + C_3 \tag{17.2.3}$$

Therefore, total average cost is given by

$$C(t) = \frac{\frac{1}{2}C_1Rt^2 + C_3}{t} = \frac{1}{2}C_1Rt + \frac{C_3}{t} \text{ (Cost Equation)} \tag{17.2.4}$$

The condition of minimum or maximum of $C(t)$,

$$\begin{aligned} &\frac{d}{dt}\Big[C(t)\Big] = 0 \\ \Rightarrow\ &\frac{1}{2}C_1R - \frac{C_3}{t^2} = 0 \\ \Rightarrow\ &t^* = \sqrt{\frac{2C_3}{C_1R}} \end{aligned} \tag{17.2.5}$$

Also, $\frac{d^2}{dt^2}C(t) = \frac{2C_3}{t^3}$, which is obviously positive for the value of t^*. Hence, $C(t)$ is minimum for optimum time interval t^* and optimum quantity to be produced or ordered at each interval t^* is given by

$$q^* = Rt^* = R\sqrt{\frac{2C_3}{C_1R}} = \sqrt{\frac{2C_3R}{C_1}} \tag{17.2.6}$$

which is called optimal lost size formula and the corresponding minimum cost

$$\begin{aligned} C^*_{\min} &= \frac{1}{2}RC_1\sqrt{\frac{2C_3}{C_1R}} + C_3\sqrt{\frac{C_1R}{2C_3}} \\ &= \sqrt{\frac{C_1C_3R}{2}} + \sqrt{\frac{C_1C_3R}{2}} \\ &= \sqrt{2C_1C_3R} \quad \text{per unit time.} \end{aligned}$$

Note 17.2.1. The cost equation (17.2.4) can also be written as

$$C(q) = \frac{1}{2}C_1q + C_3\frac{R}{q} \quad \text{where} \quad q = Rt.$$

17.2.2 Model I(b): Economic lot size with different rates of demand in different cycles

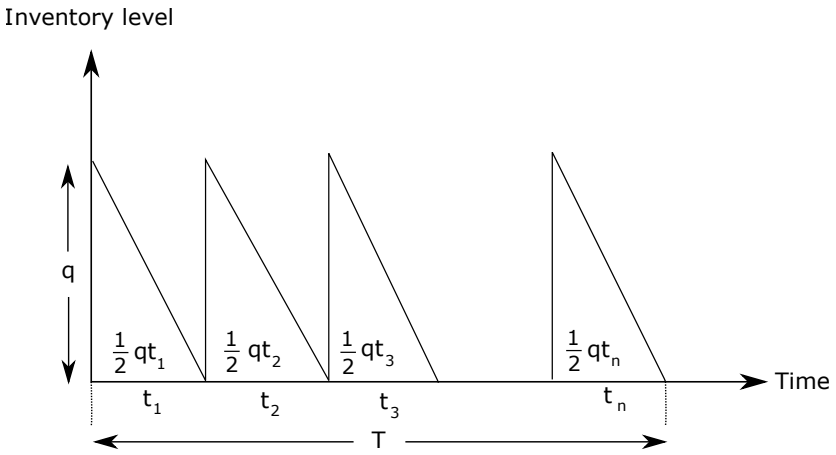
In model I(a), the total demand D is prescribed over the total period T instead of demand rate being constant for each production cycle, that is rate of demand being different in different production cycles.

Let q be the fixed quantity produced in each production cycle. Since, D is the total demand prescribed over the time period T, the number of production cycle will be $n = D/q$. Also, let the total time period $T = t_1 + t_2 + t_3 + \cdots + t_n$. Obviously, the carrying cost for the period T will be

$$\left(\frac{1}{2}qt_1\right)C_1 + \left(\frac{1}{2}qt_2\right)C_1 + \cdots + \left(\frac{1}{2}qt_n\right)C_1 = \frac{1}{2}C_1q(t_1 + t_2 + \cdots + t_n) = \frac{1}{2}C_1qT$$

Set up cost will be equal to $\frac{D}{q}C_3$. Thus, we obtain the cost equation for period T.

$$C(q) = \frac{1}{2}C_1 qT + \frac{D}{q}C_3$$



For minimum cost

$$\begin{aligned} &\frac{dC(q)}{dq} = 0 \\ \Rightarrow\ &\frac{1}{2}C_1 T - \frac{C_3}{q^2}D = 0 \\ \Rightarrow\ &q^* = \sqrt{\frac{2C_3\left(\frac{D}{T}\right)}{C_1}} \end{aligned}$$

Also, $\frac{d^2C}{dq^2} = \frac{2C_3 D}{q^3} > 0$, which minimizes the total cost $C(q)$ and the corresponding minimum value will be

$$\begin{aligned} C_{\min} &= \frac{1}{2}C_1 T\sqrt{\frac{2C_3\left(\frac{D}{T}\right)}{C_1}} + C_3 D\sqrt{\frac{C_1}{2C_3\left(\frac{D}{T}\right)}} \\ &= \sqrt{\frac{C_1 C_3 TD}{2}} + \sqrt{\frac{C_1 C_3 TD}{2}} \\ &= \sqrt{2C_1 C_3 DT} \end{aligned}$$

Hence, the minimum total average cost will be

$$\begin{aligned} C_{\min} &= \frac{\sqrt{2C_1 C_3 DT}}{T} \\ &= \sqrt{\frac{2C_1 C_3 D}{T}} \end{aligned}$$

Note 17.2.2. Here we observed that the fixed demand rate R in model I(a) is replaced by the average demand rate D/T.

Example 17.2.3. You have to supply your customer 100 units of a certain product every Monday. You obtained the product from a local supplier at Rs. 60 per unit. The cost of ordering and transportation from the supplier is Rs. 150 per order. The cost of carrying inventory is estimated at 15% per year of the cost of the product carried.

(i) Describe graphically the inventory system.

(ii) Find the lot size which will minimize the cost of the system.

(iii) How frequently should order be placed?

(iv) Determine the number of orders.

(v) Determine the optimum cost.

Solution. Here

$$\begin{aligned} R &= 100 \text{ units/week.} \\ C_3 &= 150 \text{ per order.} \\ C_1 &= \text{Rs. } \frac{15 \times 60}{100 \times 52} \text{ per unit per week} \\ &= \text{Rs. } \frac{9}{52} \end{aligned}$$

(i)

$$C(t) = 60R + \frac{1}{2}C_1 Rt + \frac{C_3}{t}.$$

(ii)

$$\begin{aligned} q^* &= \sqrt{\frac{2C_3R}{C_1}} \\ &= \sqrt{\frac{2 \times 150 \times 100 \times 52}{9}}. \\ &= 416 \text{ units} \end{aligned}$$

(iii)

$$t^* = \frac{q^*}{R} = \frac{416}{100} = 4.16 \text{ weeks}$$

(iv)

$$\eta = \frac{R}{q^*} = \frac{100}{416} \text{ orders per week}$$

(v)

$$\begin{aligned} C_{\min} &= 60R + \sqrt{2C_1C_3R} \\ &= (60 \times 100) + \sqrt{2 \times \frac{9}{52} \times 150 \times 100} \\ &= 6000 + 72 \\ &= \text{Rs. } 6072 \end{aligned}$$

■

Example 17.2.4. An aircraft company uses rebate at an approximate customer rate of 2500 kg per year. Each unit costs Rs. 30 per kg and the company personal estimate that it cost Rs. 130 to place an order and that the carrying cost of inventory is 10% per year. How frequently should orders be placed? Also determine the optimum size of each order.

Solution. Here

$$\begin{aligned} R &= 2500 \text{ kg per year.} \\ C_3 &= \text{Rs. } 130 \\ C_1 &= \text{Cost of each unit } \times \text{ inventory carrying cost} \\ &= \text{Rs. } 30 \times \frac{1}{30} \\ &= \text{Rs. } 3 \text{ per unit per year} \end{aligned}$$

$$\begin{aligned} q^* &= \sqrt{\frac{2C_3R}{C_1}} \\ &= \sqrt{\frac{2 \times 130 \times 2500}{3}} \\ &= 466 \text{ units} \end{aligned}$$

$$\therefore t^* = \frac{q^*}{R} = \frac{466}{2500} = 0.18 \text{ year} = 0.18 \times 12 \text{ months} = 2.16 \text{ months}$$

■

17.2.3 Model I(c): Economic lot size with finite rate of Replenishment (finite production) [EPQ model]

Some Notations:

$$\begin{aligned} C_1 &= \text{Holding cost per unit item per unit time.} \\ R &= \text{Demand rate.} \\ K &= \text{Production rate is finite, uniform and greater than R.} \\ t &= \text{interval between production cycle.} \\ q &= Rt \end{aligned}$$

In this model, each production cycle time t consists of two parts: t_1 and t_2, where

(i) t_1 is the period during which the stock is growing up at a rate of $(K - R)$ items per unit time.

(ii) t_2 is the period during which there is supply but there is only a constant demand at the rate of R.

It is evident from the graphical situation (see fig. 17.1) that

$$t_1 = \frac{Q}{K - R} \qquad \text{and} \qquad t_2 = \frac{Q}{R}$$

$$\begin{aligned} t &= t_1 + t_2 \\ &= \frac{Q}{K - R} + \frac{Q}{R} \\ &= \frac{QK}{R(K - R)} \end{aligned}$$

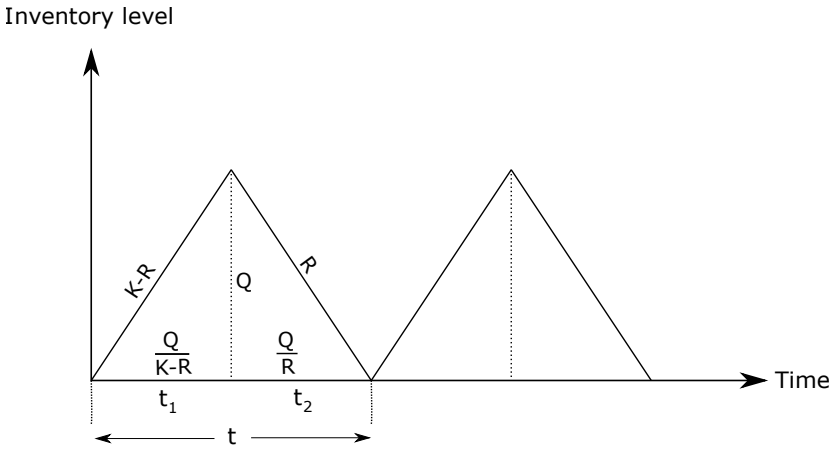


Figure 17.1

which gives

$$\begin{aligned} Q &= \frac{K-R}{K}Rt \\ &= \frac{K-R}{K}q \quad [\because q = Rt] \end{aligned}$$

Now, Holding cost for the time period t is $\frac{1}{2}C_1Qt$ and the set up cost for period t is C_3.

$\therefore$ The total average cost is

$$\begin{aligned} C(t) &= \frac{\frac{1}{2}C_1Qt + C_3}{t} \\ C(q) &= \frac{1}{2}C_1\left(\frac{K-R}{K}\right)q + C_3\frac{R}{q} \quad [\because q = Rt] \end{aligned} \tag{17.2.7}$$

For optimum value of q, we have

$$\begin{aligned} &\frac{dC}{dq} = 0 \\ &\Rightarrow \frac{1}{2}\left(1 - \frac{R}{K}\right)C_1 - \frac{C_3R}{q^2} = 0 \\ &\Rightarrow q = \sqrt{\frac{2C_3RK}{C_1(K-R)}} = \sqrt{\frac{2C_3R}{C_1\left(1-\frac{R}{K}\right)}} \end{aligned}$$

Now, $\quad \frac{d^2C}{dq^2} = \frac{2C_3R}{q^3} > 0$

$\therefore q^* = \sqrt{\frac{2C_3R}{C_1\left(1-\frac{R}{K}\right)}}$ (optimal lot size)

and $\quad t^* = \frac{q^*}{R} = \sqrt{\frac{2C_3}{C_1R\left(1-\frac{R}{K}\right)}}$

and the corresponding minimum total average cost

$$C_{\min} = \sqrt{2C_1\left(1 - \frac{R}{K}\right)C_3R}$$

Note 17.2.5. 1. If $K = R$, $C_{\min} = 0$, which implies that there will be no carrying cost and set up cost.

2. If $K \to \infty$, i.e., production rate is infinite, then this model becomes exactly same as Model I(a).

Example 17.2.6. A contractor has to supply 10,000 bearings per day to an auto-mobile manufacturer. He finds that when he starts a production run, he can produce 25,000 bearings per day. The cost of holding a bearing in stock for one year is 20 paisa and set up cost of a production run is Rs. 180. How frequently (time) should production run be made?

Solution.

$$\begin{aligned} R &= 10000 \quad \text{bearings per day} \\ K &= 25000 \quad \text{bearing per day} \\ C_1 &= \text{Rs. } \frac{0.20}{365} \text{ per bearing per day} \\ &= \text{Rs. } 0.0005 \text{ per bearing per day.} \\ C_3 &= \text{Rs. } 180 \text{ per run.} \end{aligned}$$

$$\therefore t^* = \sqrt{\frac{2 \times 180}{0.0005 \times 10000} \times \frac{3}{5}} = 0.3 \text{ day}$$

■

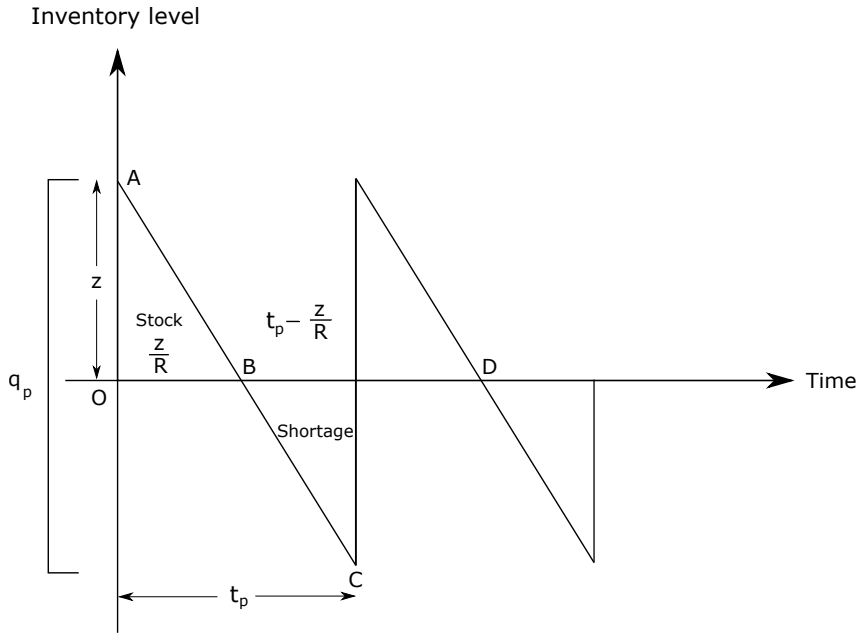
Unit 18

Course Structure

- Model II(a): EOQ model with constant rate of demand scheduling time constant.
- Model II(b): EOQ model with constant rate of demand scheduling time variable.
- Model II(c): EPQ model with shortages.

18.1 Model II(a) : EOQ model with constant rate of demand scheduling time constant

Model II is the extension of Model I allowing shortages.



Some Notations:

C_1 = Holding cost
C_2 = Shortage cost
R = Demand rate
t_p = Scheduling time period is constant
q_p = Fixed lot size (Rt_p)
z = Order level to which the inventory raised in the beginning of each scheduling period.

Here z is the variable. Production rate is infinite. Lead Time is zero.

In this model, we can easily observe that the inventory carrying cost C_1 and also the shortage cost C_2 will be involved only when $0 \leq z \leq q_p$.

$$\begin{aligned}\text{Holding cost per unit time} &= C_1(\Delta OAB)/t_p \\ &= \frac{C_1}{t_p}\left(\frac{1}{2}\cdot z\cdot\frac{z}{R}\right) \\ &= \frac{1}{2}\frac{z^2C_1}{Rt_p} \quad (\because q_p = Rt_p)\end{aligned}$$

$$\begin{aligned}\text{Shortage cost per unit time} &= C_2(\Delta BDC)/t_p \\ &= \frac{C_2}{t_p}\left(\frac{1}{2}\cdot BD\cdot DC\right) \\ &= \frac{C_2}{t_p}\left[\frac{1}{2}(t_p-\frac{z}{R})(q_p-z)\right] \\ &= \frac{1}{2}\frac{C_2}{q_p}(q_p-z)^2\end{aligned}$$

$$\text{Total average cost is } C(z) = \frac{1}{2}\frac{z^2C_1}{q_p} + \frac{1}{2}\frac{C_2}{q_p}(q_p-z)^2 + \frac{C_3}{t_p}$$

Note 18.1.1. Since, the set up cost C_3 and period t_p are constant, the average set up cost $\frac{C_3}{t_p}$ also being constant, will be considered in the cost equation.

Now

$$\frac{dC}{dz} = \frac{1}{2}\cdot\frac{C_1}{q_p}\cdot 2z + \frac{1}{2}\frac{C_2}{q_p}2(q_p-z)(-1) = 0$$
$$\Rightarrow z = \frac{C_2}{C_1+C_2}q_p = \frac{C_2}{C_1+C_2}Rt_p.$$

$$\frac{d^2C}{dz^2} = \frac{C_1}{q_p} + \frac{C_2}{q_p} = \frac{C_1+C_2}{q_p} > 0.$$
$$\therefore z^* = \frac{C_2}{C_1+C_2}Rt_p$$
$$C_{\min} = \frac{C_1C_2}{2(C_1+C_2)}Rt_p.$$

18.2 Model II(b) : EOQ model with constant rate of demand scheduling time variable

Assumptions:

(i) R is the demand rate.

(ii) Production is instantaneous.

(iii) $q = Rt$.

(iv) t is the scheduling time period which is variable.

(v) z is the order level.

(vi) Lead time is zero.

Formulate the model. Show that the optimal order quantity per run which minimizes the total cost is

$$q = \sqrt{\frac{2RC_3(C_1 + C_2)}{C_1C_2}}$$

Since, all the assumptions in this model are same as in Model II(a), except with the difference that the scheduling time period t is not constant here, so, it now becomes important to consider the average set up cost $\frac{C_3}{t}$ in the cost equation.

Thus the cost equation becomes

$$C(z,t) = \frac{C_1z^2}{2Rt} + \frac{1}{2}\frac{C_2}{Rt}(Rt - z)^2 + \frac{C_3}{t}.$$

For the optimization, $\frac{\partial C}{\partial z} = 0$ and $\frac{\partial C}{\partial t} = 0$ which gives

$$\frac{1}{t}\left(\frac{2C_1z}{2R} - \frac{2C_3}{2R}(Rt - z)\right) = 0$$
$$\therefore z = \frac{C_2Rt}{C_1 + C_2}$$

Now

$$-\frac{1}{t^2}\left(\frac{C_1z^2}{2R} + \frac{C_2}{2R}(Rt - z)^2 + C_3\right) + \frac{1}{t}\left(0 + \frac{C_2}{2R}2(Rt - z) + 0\right) = 0$$
$$\Rightarrow \quad -\frac{1}{t^2}\left(\frac{C_1z^2}{2R} + \frac{C_2}{2R}(Rt - z)^2 + C_3\right) + \frac{C_2}{t}(Rt - z) = 0$$

Multiplying this equation by $2Rt^2$ and simplifying we get,

$$-(C_1 + C_2)z^2 + C_2R^2t^2 = 2RC_3$$

Substituting the value of z in the given equation, we have

$$
\begin{aligned}
&-\frac{R^2t^2C_2^2}{C_1+C_2}+C_2R^2t^2=2RC_3\\
&\Rightarrow C_2R^2t^2\left(1-\frac{C_2}{C_1+C_2}\right)=2RC_3\\
&\Rightarrow C_2R^2t^2\left(\frac{C_1}{C_1+C_2}\right)=2RC_3\\
&\Rightarrow t=\sqrt{\frac{2C_3(C_1+C_2)}{RC_1C_2}}
\end{aligned}
$$

For minimum cost, we may further verify that

$$
\begin{aligned}
&\frac{\partial^2C}{\partial t^2}\cdot\frac{\partial^2C}{\partial z^2}-\left(\frac{\partial^2C}{\partial t\partial z}\right)^2>0\\
&\text{and}\quad \frac{\partial^2C}{\partial t^2}>0 \qquad \frac{\partial^2C}{\partial z^2}>0
\end{aligned}
$$

Hence

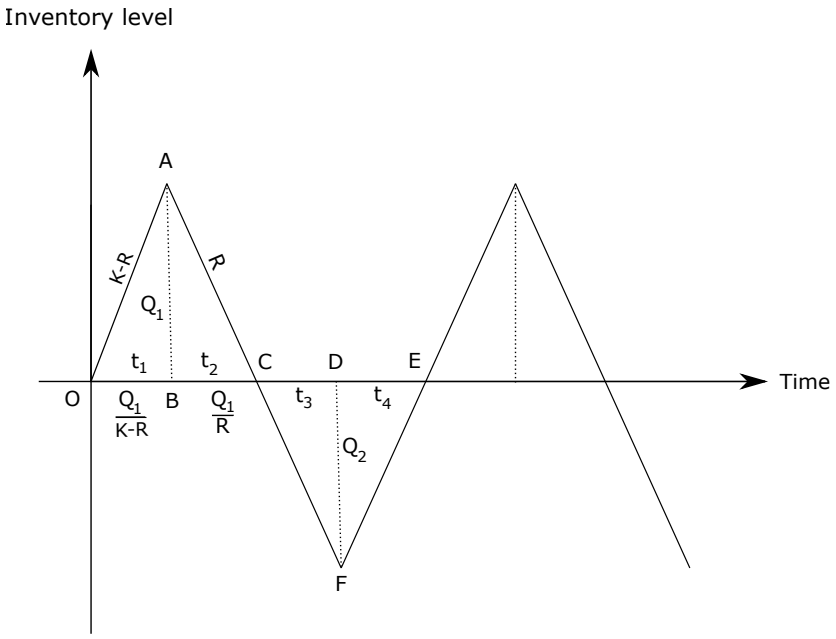
$$
\begin{aligned}
t^* &= \sqrt{\frac{2C_3(C_1+C_2}{RC_1C_2}}\\
q^* &= Rt^*=R\sqrt{\frac{2C_3(C_1+C_2)}{RC_1C_2}}\\
&= \sqrt{\frac{2RC_3(C_1+C_2)}{C_1C_2}} \quad \text{(EOW/lot size)}\\
C_{\min} &= \frac{C_1}{2Rt^*}\left(\frac{C_2Rt^*}{C_1+C_2}\right)^2+\frac{1}{2}\frac{C_2}{Rt^*}\left(Rt^*-\frac{C_2Rt^*}{C_1+C_2}\right)^2+\frac{C_3}{t^*}\\
&= \frac{C_1C_2^2R^2}{2(C_1+C_2)^2Rt^*}t^{*2}+\frac{1}{2}\frac{C_2}{Rt^*}\left(\frac{C_1Rt^*}{C_1+C_2}\right)^2+\frac{C_3}{t^*}\\
&= \frac{C_1C_262}{2(C_1+C_2)^2}(Rt^*)+\frac{C_2C_1^2}{2(C_1+C_2)^2}(Rt^*)+\frac{C_3}{t^*}\\
&= \frac{1}{2}\frac{C_1C_2}{(C_1+C_2)^2}(C_1+C_2)(Rt^*)+\frac{C_3}{t^*}\\
&= \frac{1}{2}\frac{C_1C_2}{(C_1+C_2)}\sqrt{\frac{2RC_3(C_1+C_2)}{C_1C_2}}+C_3\sqrt{\frac{RC_1C_2}{2C_3(C_1+C_2)}}\\
&= \sqrt{\frac{C_1C_2RC_3}{2(C_1+C_3}}+\sqrt{\frac{RC_1C_2C_3}{2(C_1+C_2)}}\\
&= 2\sqrt{\frac{RC_1C_2C_3}{2(C_1+C_2)}}\\
&= \sqrt{2C_1C_3R\times\frac{C_2}{C_1+C_2}}\\
&= \sqrt{2C_1C_3R}\sqrt{\frac{C_2}{C_1+C_2}}
\end{aligned}
$$

Further, it is interesting to note that the minimum cost is less than that already given by Model I(a) $\sqrt{2C_1C_3R}$. (Draw figure as like Model II(a) replaced by t).

18.3 Model II(c) : EPQ model with shortages

The production lot size model with shortage.

Assumptions:



(i) R is the demand rate.

(ii) Lead time is zero.

(iii) Production rate (K) is finite, $K > R$.

(iv) Inventory carrying cost $C_1 = IP$ (For EPQ), $P =$ Finite production cost.

(v) Shortages are allowed and backlogged.

(vi) Shortage cost is Rs. C_2 per quantity unit per unit time.

(vii) Set up cost is Rs. C_3 per order or per set up.

$$\text{Holding Cost} = C_1 \times \Delta\text{OAC} = C_1 \times \frac{1}{2}Q_1(t_1 + t_2).$$

$$\text{Shortage Cost} = C_2 \left(\frac{1}{2}Q_2(t_3 + t_4)\right)$$

and Set up cost C_3.

Thus, the total average cost,

$$\begin{aligned} C &= \frac{\frac{1}{2}C_1Q_1(t_1+t_2)+\frac{1}{2}C_2Q_2(t_3+t_4)+C_3}{t_1+t_2+t_3+t_4} \\ t_1 &= \frac{Q_1}{K-R}, \quad t_2=\frac{Q_1}{R} \\ &= \frac{Rt_2}{K-R}, \quad Q_1=Rt_2. \end{aligned} \tag{18.3.1}$$

Again,

$$\begin{aligned} Q_2 &= Rt_3, & t_4 &= \frac{Q_2}{K-R}. \\ Q_2 &= (K-R)t_4 & &= \frac{Rt_3}{K-R}. \end{aligned}$$

Finally,

$$\begin{aligned} q=Rt &= R(t_1+t_2+t_3+t_4) \\ &= R\left(\frac{Rt_2}{K-R}+t_2+t_3+\frac{Rt_3}{K-R}\right) \\ &= \frac{(t_2+t_3)KR}{K-R} \\ C &= \frac{\frac{1}{2}\left\{C_1(Rt_2)\left(\frac{Rt_2}{K-R}+t_2\right)+C_3Rt_3\left(t_3+\frac{Rt_3}{K-R}\right)\right\}+C_3}{\frac{Rt_2}{K-R}+t_2+t_3+\frac{Rt_3}{K-R}} \\ &= \frac{\frac{1}{2}\left\{\frac{C_1t_2^2RK}{K-R}+\frac{C_2t_3^2RK}{K-R}\right\}+C_3}{(t_2+t_3)\left(1+\frac{R}{K-R}\right)} \\ &= \frac{\frac{1}{2}(C_1t_2^2+C_2t_3^2)\left(\frac{RK}{K-R}\right)+C_3}{(t_2+t_3)\left(\frac{K}{K-R}\right)} \\ &= \frac{\frac{1}{2}(C_1t_2^2+C_2t_3^2)RK+C_3(K-R)}{K(t_2+t_3)}. \end{aligned}$$

This is a function of t_2 and t_3 $C(t_2,t_3)$

$$\frac{\partial C}{\partial t_2}=0, \quad \frac{\partial C}{\partial t_3}=0,$$

$$t_2^*=\sqrt{\frac{2C_3C_2(1-R/K)}{(R(C_1+C_2)C_1}}, \qquad q^*=\sqrt{\frac{2RC_3(C_1+C_2)}{(C_1C_2}\left(\frac{1}{1-R/K}\right)}$$

$$t_3^*=\sqrt{\frac{2C_3C_1(1-R/K)}{(R(C_1+C_2)C_2}}, \qquad C_{\text{min}}=\sqrt{\frac{2RC_1C_2C_3(1-R/K)}{C_1+C_2}}$$

$$C=\frac{\frac{1}{2}(C_1t_2^2+C_2t_3^2)RK+C_3(K-R)}{K(t_2+t_3)}.$$

Now,

$$\begin{aligned}
& \frac{\partial C}{\partial t_2} = 0. \\
\Rightarrow \quad & \frac{K(t_2+t_3)\left[\frac{1}{2}C_1 \times 2t_2\right]RK - \left[\frac{1}{2}(C_1t_2^2 + C_2t_3^2)RK + C_3(K-R)\right]K}{K^2(t_2+t_3)^2} = 0 \\
\Rightarrow \quad & K(t_2+t_3)\cdot C_1t_2RK - \left[\frac{1}{2}(C_1t_2^2 + C_2t_3^2)RK + C_3(K-R)\right]K = 0 \\
\Rightarrow \quad & C_1t_2^2RK^2 + C_1t_2t_3RK^2 - \frac{1}{2}C_1t_2^2RK^2 - \frac{1}{2}C_2t_3^2RK^2 - C_3K(K-R) = 0 \\
\Rightarrow \quad & \frac{1}{2}C_1t_2^2RK^2 + C_1t_2t_3RK^2 - \frac{1}{2}C_2t_3^2RK^2 - C_3K(K-R) = 0 \\
\Rightarrow \quad & \frac{1}{2}C_1t_2^2RK^2 + C_1t_2t_3RK^2 - \frac{1}{2}C_2t_3^2RK^2 = C_3K(K-R) \\
\Rightarrow \quad & \frac{1}{2}RK^2(C_1t_2^2 + 2C_1t_2t_3 - C_2t_3^2) = C_3K(K-R) \\
\Rightarrow \quad & C_1t_2^2 + 2C_1t_2t_3 - C_2t_3^2 = \frac{2C_3(1-R/K)}{R} \\
\Rightarrow \quad & C_1t_2^2 + 2C_1t_2t_3 + C_1t_3^2 - C_1t_3^2 - C_2t_3^2 = \frac{2C_3(1-R/K)}{R} \\
\Rightarrow \quad & C_1(t_2+t_3)^2 - t_3^2(C_1+C_2) = \frac{2C_3(1-R/K)}{R} \\
\Rightarrow \quad & C_1(t_2+t_3)^2 = \frac{2C_3(1-R/K)}{R} + t_3^2(C_1+C_2) \\
\Rightarrow \quad & (t_2+t_3)^2 = \frac{2C_3(1-R/K)}{RC_1} + \frac{t_3^2(C_1+C_2)}{C_1} \\
\Rightarrow \quad & t_2+t_3 = \sqrt{\frac{2C_3(1-R/K)}{RC_1} + \frac{t_3^2(C_1+C_2)}{C_1}}.
\end{aligned}$$

Also,

$$\begin{aligned}
& K(t_2+t_3)\left[\frac{1}{2}\times C_2 \times 2C_3\right]RK - \left[\frac{1}{2}(C_1t_2^2 + C_2t_3^2)RK + C_3(K-R)\right]K = 0 \\
\Rightarrow \quad & C_2t_2t_3RK^2 + C_2t_3^2RK^2 - \frac{1}{2}C_1t_2^2RK^2 - \frac{1}{2}C_2t_3^2RK^2 - C_3(K-R)K = 0 \\
\Rightarrow \quad & \frac{1}{2}C_2t_3^2RK^2 + C_2t_2t_3RK^2 - \frac{1}{2}C_1t_2^2RK^2 = C_3(K-R)K \\
\Rightarrow \quad & C_2t_3^2 + 2C_2t_2t_3 + C_2t_2^2 - (C_1+C_2)t_2^2 = \frac{2C_3(1-R/K)}{R} \\
\Rightarrow \quad & C_2(t_2+t_3)^2 - (C_1+C_2)t_2^2 = \frac{2C_3(1-R/K)}{R}.
\end{aligned}$$

Now,

$$
\begin{aligned}
& C_1(t_2+t_3)^2-(C_1+C_2)t_3^2=\frac{2C_3(1-R/K)}{R}\\
& C_2(t_2+t_3)^2-(C_1+C_2)t_2^2=\frac{2C_3(1-R/K)}{R}\\
\Rightarrow\ & C_1(t_2+t_3)^2-(C_1+C_2)t_3^2=C_2(t_2+t_3)^2-(C_1+C_2)t_2^2\\
\Rightarrow\ & C_1(t_2+t_3)^2-C_2(t_2+t_3)^2=(C_1+C_2)t_3^2-(C_1+C_2)t_2^2\\
\Rightarrow\ & (t_2+t_3)^2(C_1-C_2)=(C_1+C_2)(t_3^2-t_2^2)\\
\Rightarrow\ & (t_2+t_3)^2(C_1-C_2)=(C_1+C_2)(t_3-t_2)(t_3+t_2)\\
\Rightarrow\ & (t_2+t_3)(C_1-C_2)=(C_1+C_2)(t_3-t_2)\\
\Rightarrow\ & C_1t_2-C_2t_2+C_1t_3-C_2t_3=C_1t_3+C_2t_3-C_1t_2-C_2t_2\\
\Rightarrow\ & 2C_1t_2=2C_2t_3\\
\Rightarrow\ & 2C_1t_2=2C_2t_3\\
\Rightarrow\ & t_2=\frac{C_2}{C_1}t_3
\end{aligned}
$$

Thus,

$$
\begin{aligned}
& C_2(t_2+t_3)^2-(C_1+C_2)t_2^2=\frac{2C_3(1-R/K)}{R}\\
\Rightarrow\ & C_2\left(\frac{C_2}{C_1}t_3+t_3\right)^2-(C_1+C_2)\left(\frac{C_2}{C_1}t_3\right)^2=\frac{2C_3(1-R/K)}{R}\\
\Rightarrow\ & C_2\left(\frac{C_2}{C_1}+1\right)^2t_3^2-(C_1+C_2)\frac{C_2^2}{C_1^2}t_3^2=\frac{2C_3(1-R/K)}{R}\\
\Rightarrow\ & \frac{C_2(C_2+C_1)^2}{C_1^2}t_3^2-\frac{(C_1+C_2)C_2^2}{C_1^2}t_3^2=\frac{2C_3(1-R/K)}{R}\\
\Rightarrow\ & \frac{t_3^2}{C_1^2}(C_1+C_2)\left[C_2(C_1+C_2)-C_2^2\right]=\frac{2C_3(1-R/K)}{R}\\
\Rightarrow\ & \frac{t_3^2(C_1+C_2)}{C_1^2}C_1C_2=\frac{2C_3(1-R/K)}{R}\\
\Rightarrow\ & \frac{t_3^2(C_1+C_2)}{C_1}C_2=\frac{2C_3(1-R/K)}{R}\\
\Rightarrow\ & t_3^2=\frac{2C_1C_3(1-R/K)}{R(C_1+C_2)C_2}\\
\Rightarrow\ & t_3^*=\sqrt{\frac{2C_1C_3(1-R/K)}{R(C_1+C_2)C_2}}.
\end{aligned}
$$

Now,

$$\begin{aligned} t_2^* &= \frac{C_2}{C_1} t_3^* \\ &= \frac{C_2}{C_1}\sqrt{\frac{2C_1C_3(1-R/K)}{R(C_1+C_2)C_2}} \\ &= \sqrt{\frac{2C_1C_3(1-R/K)C_2^2}{C_1^2R(C_1+C_2)C_2}} \\ &= \sqrt{\frac{2C_2C_3(1-R/K)}{R(C_1+C_2)C_1}}. \end{aligned}$$

Now,

$$\begin{aligned} q^* &= \frac{KR}{K-R}\left[\sqrt{\frac{2C_2C_3(1-R/K)}{R(C_1+C_2)C_1}} + \sqrt{\frac{2C_1C_3(1-R/K)}{R(C_1+C_2)C_2}}\right] \\ &= \frac{R}{(1-R/K)}\left[\sqrt{\frac{2C_3(1-R/K)}{R(C_1+C_2)}}\left\{\sqrt{\frac{C_2}{C_1}} + \sqrt{\frac{C_1}{C_2}}\right\}\right] \\ &= \sqrt{\frac{2C_3}{R(C_1+C_2)(1-R/K)}} \times \frac{(C_1+C_2)R}{\sqrt{C_1C_2}} \\ &= \sqrt{\frac{2C_3(C_1+C_2)^2R^2}{R(C_1+C_2)C_1C_2(1-R/K)}} \\ &= \sqrt{\frac{2RC_3(C_1+C_2)}{C_1C_2(1-R/K)}} = \sqrt{\frac{2RC_3(C_1+C_2)}{C_1C_2}\left(\frac{1}{1-R/K}\right)}. \end{aligned}$$

So,

$$\begin{aligned} C_{\min} &= \frac{\frac{1}{2}(C_1t_2^{*2} + C_2t_3^{*2})RK + C_3(K-R)}{K(t_2^* + t_3^*)} \\ &= \frac{\frac{1}{2}\left[\frac{2C_1C_2C_3(1-R/K)}{R(C_1+C_2)C_1} + \frac{2C_1C_2C_3(1-R/K)}{R(C_1+C_2)C_2} + C_3(K-R)\right]}{K\left(\sqrt{\frac{2C_3(1-R/K)}{R(C_1+C_2)}} \cdot \frac{(C_1+C_2)}{\sqrt{C_1C_2}}\right)} \\ &= \frac{\left[\frac{C_2C_3(1-R/K)}{R(C_1+C_2)} + \frac{C_1C_3(1-R/K)}{R(C_1+C_2)} + \frac{C_3(K-R)}{2}\right]}{K\sqrt{\frac{2C_3(1-K/R)(C_1+C_2)}{RC_1C_2}}} \\ &= \frac{2C_2C_3(1-R/K) + 2C_1C_3(1-R/K) + C_3K(1-R/K)R(C_1+C_2)}{K\sqrt{\frac{2C_3(1-R/K)(C_1+C_2)}{RC_1C_2}}}. \end{aligned}$$

Example 18.3.1. The demand of an item is uniform at a rate of 25 units per month. The fixed cost is Rs. 15 each time a production run is made (Setup cost). The production cost if Rs. 1 per item and inventory carrying cost is Rs. 0.30 per item per month. If the shortage cost is Rs. 1.50 per item per month, determine how often to make a production run and of what size it should be?

Solution. We have,

$$\begin{aligned} R &= 25 \text{ units per month} \\ C_3 &= \text{Rs. } 15 \text{ per run} \\ I &= \text{Rs. } 0.30 \text{ per item per month. (Inventory carrying cost)} \\ C_2 &= \text{Rs. } 1.50 \text{ per item per month} \\ P &= \text{Rs. } 1 \text{ per item.} \end{aligned}$$

Thus,

$$C_1 = \text{Rs. } 0.30 \text{ per item per month.}$$

Here, the demand of an item is uniform. So,

$$q^* = \sqrt{\frac{2RC_3(C_1+C_2)}{C_1C_2}} = \sqrt{\frac{2\times 25\times 15\times (0.30+1.50)}{0.30\times 1.50)}} \approx 54 \text{ units.}$$

and

$$t^* = \sqrt{\frac{2C_3(C_1+C_2)}{RC_1C_2}} = \sqrt{\frac{2\times 15\times (0.30+1.50)}{25\times 0.30\times 1.50}} = 2.19 \text{ months.}$$

■

Unit 19

Course Structure

- Model III: Multi-item inventory model

19.1 Model III: Multi-item inventory model

So far, we have considered each item separately but if there exists a relationship among the items under some limitations, then it is not possible to consider them separately. After constructing the cost equation in such models, we use the method of Lagrange's multiplier to minimize the cost. We consider the problem with the following assumptions

1. n is the number of items to be considered and no lead time.
2. R_{1i} is the uniform demand rate for the ith item $(i = 1, 2, \ldots, n)$.
3. C_{1i} is the holding cost of the ith item
4. Shortages are not allowed
5. C_{3i} is the setup cost for the ith item
6. q_i is the total quantity to be produced of the ith item.

Now, proceeding exactly as in the model I(a), we get,

$$C_i(t) = \frac{1}{2}C_{1i}R_i t + \frac{C_{3i}}{t},$$
$$\text{or,} \quad C_i(q_i) = \frac{1}{2}C_{1i}q_i + \frac{C_{3i}R_i}{q_i} \tag{19.1.1}$$

Then total cost

$$C = \sum_{i=1}^{n}\left\{\frac{1}{2}C_{1i}q_i + \frac{C_{3i}R_i}{q_i}\right\} \tag{19.1.2}$$

To determine the optimum value of q_i, we have

$$\begin{aligned} & \frac{\partial C}{\partial q_i} = 0 \\ \Rightarrow \quad & \frac{1}{2}C_{1i} - \frac{C_{3i}R_i}{q_i^2} = 0 \\ \Rightarrow \quad & q_i = \sqrt{\frac{2c_{3i}R_i}{C_{1i}}}. \end{aligned}$$

Thus,

$$\frac{\partial^2 C}{\partial q_i^2} > 0, \ \ \forall\, q_i.$$

The total cost is minimum. Hence, the optimum cost of

$$q_i^* = \sqrt{\frac{2C_{3i}R_i}{C_{1i}}}, \ \ (i = 1, 2, \ldots, n) \tag{19.1.3}$$

We now proceed to consider the effect of limitations, which are,

1. limitation on investment
2. limitation on stock unit
3. limitation on warehouse floor space

19.1.1 Model III(a): Limitation on Investment

In this case, there is an upper limit M (in Rs.) on the amount to be invested on inventory. Let C_{4i} be the unit price of the ith item. Then

$$\sum_{i=1}^{n} C_{4i}q_i \leq M \tag{19.1.4}$$

Now, our problem is to minimize the total cost C given by equation (19.1.2) subject to the constraint (19.1.4). In this situation, two cases may arise.

Case I: When $\sum_{i=1}^{n} C_{4i}q_i \leq M$ and $q_i^* = \sqrt{\frac{2C_{3i}R_i}{C_{1i}}}$.

In this case, there is no difficulty and hence q_i^* is the optimal solution.

Case II: When $\sum_{i=1}^{n} C_{4i}q_i > M$ and $q_i^* = \sqrt{\frac{2C_{3i}R_i}{C_{1i}}}$.

In this case, q_i^* are not required optimal solutions. Thus, we shall use the Lagrange's multiplier technique.

$$L = \sum_{i=1}^{n}\left(\frac{1}{2}C_{1i}q_i + \frac{C_{3i}R_i}{q_i}\right) + \lambda\left(\sum_{i=1}^{n} C_{4i}q_i - M\right).$$

Here, λ is the Lagrangian multiplier.

The necessary condition

$$\begin{aligned}
&\frac{\partial L}{\partial q_i} = 0 \ (i = 1, 2, \ldots, n) \\
&\frac{\partial L}{\partial \lambda} = 0 \\
\Rightarrow \quad &\frac{1}{2}C_{1i} - \frac{C_{3i}R_i}{q_i^2} + \lambda C_{4i} = 0, \quad \text{and} \quad C_{4i}q_i - M = 0 \\
\Rightarrow \quad &q_i^* = \sqrt{\frac{2C_{3i}R_i}{C_{1i} + 2\lambda C_{4i}}} \quad \text{and} \quad C_{4i}q_i^* = M.
\end{aligned}$$

q_i^* depends on λ. λ can be found by *trial and error method*. By trying positive successive values of λ, the values of λ^* should result in simultaneous value of q_i^* satisfying the given constraint by equality sense.

Example 19.1.1. Consider a shop producing three items, the items are produced in lots. The demand rate for each item is constant and can be assumed to be deterministic. No back order (shortages) are allowed. The following data are given below.

Item	1	2	3
H.C	20	20	20
S.C	50	40	60
Cost per unit item	6	7	5
Yearly demand rate	10,000	12,000	7,500

Determine approximately the EOQ when the total value of average inventory levels of three items if Rs. 1,000.

Solution.

$$\begin{aligned}
q_1^* &= \sqrt{\frac{2C_{31}R}{C_{11}}} = \sqrt{\frac{2 \times 50 \times 10,000}{20}} = 100\sqrt{5} \approx 223 \\
q_2^* &= 40\sqrt{30} \approx 216 \\
q_3^* &= 150\sqrt{2} \approx 210.
\end{aligned}$$

Since the average optimal inventory at any time is $q_i^*/2$, the investment over the average inventory is obtained by replacing q_i by $q_i^*/2$, that is,

$$\sum_{i=1}^{n} C_{4i}\left(\frac{1}{2}q_i^*\right) = \text{Rs. } \left(6 \times \frac{223}{2} + 7 \times \frac{216}{2} + 5 \times \frac{210}{2}\right) = \text{Rs. } 1950.$$

We observe that the amount of Rs. 1950 is greater than the upper limit of Rs. 1000. Thus, we try to find the suitable value of λ by trial and error method for computing q_i^*.

If we put $\lambda = 4$, we get

$$\begin{aligned}
q_1^* &= \sqrt{\frac{2 \times 50 \times 10,000}{20 + 2 \times 4 \times 6}} = 121 \\
q_2^* &= 112 \\
q_3^* &= 123.
\end{aligned}$$

$$\text{Cost of average inventory} = 6 \times \frac{121}{2} + 7 \times \frac{112}{2} + 5 \times \frac{123}{2} = \text{Rs. } 1112.50.$$

Again, if we put $\lambda = 5$, then

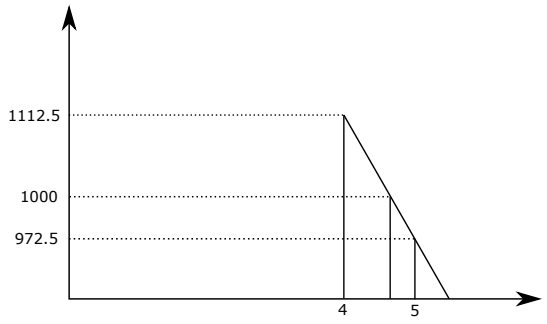
$$q_1^* = 111$$
$$q_2^* = 102$$
$$q_3^* = 113.$$

and

$$\text{Corresponding cost} = \text{Rs. } 972.50$$

which is less than Rs. 1000.

From this, we conclude that, the most suitable value of λ lies between 4 and 5.



To find the most suitable value of λ, we draw a graph between cost and the value of λ as shown in the figure. This graph indicates that $\lambda = 4.7$ is the most suitable value corresponding to which the cost of inventory is Rs. 999.5, which is sufficiently close to Rs. 1000. Hence, for $\lambda = 4.7$, we obtain

$$q_1^* = 114$$
$$q_2^* = 105$$
$$q_3^* = 116.$$

■

19.1.2 Model III(b): Limitation on inventory

In this case, the upper limit of average number of all units in stock is N (say). Hence we have, since the average number of units at any time is $q_i/2$.

$$\text{Min } C = \sum_{i=1}^{n} \left(\frac{1}{2} C_{1i} q_i + \frac{C_{3i} R_i}{q_i} \right)$$
$$\text{subject to } \frac{1}{2} \sum_{i=1}^{n} q_i \leq N.$$

Here also, two cases arise.

Case I: $\frac{1}{2} \sum_{i=1}^{n} q_i \leq N$ and $q_i^* = \sqrt{\frac{2C_{3i}R_i}{C_{1i}}}$, there is no difficulty and the optimum values of q_i^*.

Case II: $\frac{1}{2}\sum_{i=1}^{n} q_i > N$, then q_i^* are not the required values. So, we use Lagrange's multiplier technique. Here, Lagrangian function

$$L = \sum_{i=1}^{n}\left(\frac{1}{2}C_{1i}q_i + \frac{C_{3i}R_i}{q_i}\right) + \lambda\left(\frac{1}{2}\sum_{i=1}^{n} q_i - N\right)$$

where $\lambda > 0$ is a Lagrangian multiplier.

For the minimum value of L, the necessary conditions are

$$\frac{\partial L}{\partial q_i} = \frac{1}{2}C_{1i} - \frac{C_{3i}R_i}{q_i^2} + \frac{\lambda}{2} = 0$$
$$\frac{\partial L}{\partial \lambda} = \frac{1}{2}\sum_{i=1}^{n} q_i - N = 0, \;\; i = 1, 2, \ldots, n.$$

Solving, we get

$$q_i^* = \sqrt{\frac{2C_{3i}R_i}{C_{1i} + \lambda}}$$
$$\frac{1}{2}\sum_{i=1}^{n} q_i = N.$$

To obtain the value of q_i^*, we obtain the value of λ by successive trial and error method and satisfying the given constraint in equality sign.

Example 19.1.2. A company producing three items have a limited storage space of 750 items of all types in average. Determine the optimal production quantity for each item separately when the following information is given

Product	1	2	3
H.S(Rs.)	0.05	0.02	0.04
S.C(Rs.)	50	40	60
D.R(per unit)	100	120	75

Solution. We have

$$q_1^* = 447$$
$$q_2^* = 693$$
$$q_3^* = 464.$$

$$\text{The total average inventory is} = \frac{1}{2}(447 + 693 + 464) = 802 \text{ units},$$

which is greater than 750 units per year. Thus, we have to find the value of the parameter λ by trial and error method.

From these, we observe that the average inventory level is less than the available amount of items. So we try for some other values of λ,

$$\lambda = 0.004,\ 0.003,\ 0.002,\ \text{etc.}$$

For $\lambda = 0.002$,

$$
\begin{aligned}
&q_1^* = 428 \\
&q_2^* = 628 \\
&q_3^* = 444 \\
&\text{Average inventory level} = \frac{1}{2}(428 + 628 + 444) = 750,
\end{aligned}
$$

which is equivalent to the given amount of average inventory. Hence, the optimal solutions are

$$
\begin{aligned}
q_1^* &= 428 \\
q_2^* &= 628 \\
q_3^* &= 444.
\end{aligned}
$$

For $\lambda = 0.004$,

$$
\begin{aligned}
q_1^* &= \sqrt{\frac{2 \times 50 \times 100}{0.05 + 0.004}} = 430 \\
q_2^* &= \sqrt{\frac{2 \times 40 \times 120}{0.02 + 0.004}} = 632 \\
q_1^* &= \sqrt{\frac{2 \times 60 \times 75}{0.04 + 0.004}} = 452 \\
&\text{Average inventory level} = \frac{1}{2}(430 + 632 + 452) = 757.
\end{aligned}
$$

For $\lambda = 0.003$,

$$
\begin{aligned}
q_1^* &= \sqrt{\frac{2 \times 50 \times 100}{0.05 + 0.003}} = 434 \\
q_2^* &= \sqrt{\frac{2 \times 40 \times 120}{0.02 + 0.003}} = 646 \\
q_1^* &= \sqrt{\frac{2 \times 60 \times 75}{0.04 + 0.003}} = 457 \\
&\text{Average inventory level} = \frac{1}{2}(434 + 646 + 457) = 768.5.
\end{aligned}
$$

■

19.1.3 Model III(c): Limitation on floor space

$A =$ The maximum storage area available for the n items.
$a_i =$ Storage area required per unit of the ith item.

Thus, the total storage requirement constraint becomes

$$\sum_{i=1}^{n} a_i q_i \leq A, \;\; q_i \geq 0.$$

Hence, our problem becomes,

$$\text{Min } C = \sum_{i=1}^{n} \left(\frac{1}{2} C_{1i} q_i + \frac{C_{3i} R_i}{q_i} \right),$$

$$\text{Subject to } \sum_{i=1}^{n} a_i q_i \leq A.$$

$$q_i \geq 0.$$

Case I: If $\sum_{i=1}^{n} a_i q_i \leq A$, then $q_i^* = \sqrt{\frac{2C_{3i}R_i}{C_{1i}}}$. Here we have no difficulty. Hence q_i^* is the optimal solution.

Case II: If $\sum_{i=1}^{n} a_i q_i > A$, then the optimal value q_i^* are not the required value. So we use the Lagrange's multiplier technique. The Lagrangain function is

$$L = \sum_{i=1}^{n} \left(\frac{1}{2} C_{1i} q_i + \frac{C_{3i} R_i}{q_i} \right) + \lambda \left(\sum_{i=1}^{n} a_i q_i - A \right)$$

where $\lambda > 0$ is a Lagrangian multiplier.

The necessary conditions for minimum value of L are

$$\frac{\partial L}{\partial q_i} = 0, \quad \frac{\partial L}{\partial \lambda} = 0.$$

Then, solving we have

$$q_i^* = \sqrt{\frac{2C_{3i}R_i}{C_{1i} + 2\lambda a_i}}, \ \ i = 1, 2, \ldots, n, \quad \text{and} \quad \sum_{i=1}^{n} a_i q_i^* = A.$$

The second equation implies that q_i^* must satisfy the storage constraint in equality sense. The determination of λ by usual trial and error method automatically gives the optimal value of q_i^*.

Unit 20

Course Structure

- Model IV: Deterministic inventory model with price breaks of quantity discount
- Probabilistic Inventory Model

20.1 Model IV: Deterministic inventory model with price breaks of quantity discount

Notations:

$P =$ Cost per item of producing.
$I =$ Unit price per unit item.
$C_3 =$ Setup cost.
$R =$ demand rate.
$t =$ Interval between placing orders.
$q =$ Quantity order.

Assumptions:

1. Demand rate R is constant.
2. Demand is both fixed and known.
3. No shortages are to be permitted.
4. The variable cost associated with the purchasing process.

Determine:

1. How often should be purchased (t^*)?
2. How many units should be purchased at any time (q^*)?

We have,

$$q = Rt \tag{20.1.1}$$

The number of inventories will be given by $\frac{1}{2}qt$.

$$\frac{1}{2}qt = \frac{1}{2}q\frac{q}{R} = \frac{q^2}{R} \tag{20.1.2}$$

The number of lot of inventories will be given by

$$\frac{1}{2}\frac{qt}{q} = \frac{1}{2}\frac{\frac{q^2}{R}}{q} = \frac{1}{2}\frac{q}{R} \tag{20.1.3}$$

$$\begin{aligned}
C_3 &= \text{Setup Cost.} \\
qP &= \text{the purchasing cost of q units.} \\
C_3\left(\frac{1}{2}\frac{q}{R}\right)I &= \text{Cost associated with setup of inventory for period t.} \\
qP\left(\frac{1}{2}\frac{q}{R}\right)I &= \text{Cost associated with purchase of inventory for period t.}
\end{aligned}$$

Therefore, total cost for period t is given by,

$$C_3 + qP + C_3\frac{1}{2}\frac{q}{R}I + qP\cdot\frac{1}{2}\frac{q}{R}I$$

Hence, average cost per unit time,

$$\begin{aligned}
C(q) &= \frac{1}{t}\left(C_3 + qP + C_3\frac{1}{2}\frac{q}{R}I + qP\cdot\frac{1}{2}\frac{q}{R}I\right) \\
C(q) &= \frac{C_3R}{q} + pR + \frac{C_3I}{2} + \frac{qPI}{2} \qquad \left(\text{since } t = \frac{q}{R}\right)
\end{aligned}$$

But the term $\frac{1}{2}C_3I$ being constant throughout the model, it maybe neglected for the purpose of minimization. Therefore,

$$C(q) = \frac{C_3R}{q} + PR + \frac{qPI}{2} \tag{20.1.4}$$

For minimum value of $C(q)$, $\frac{d}{dq}C(q) = 0$.

$$\begin{aligned}
&\frac{d}{dq}C(q) = 0 \\
\Rightarrow\quad &\frac{-C_3R}{q^2} + \frac{1}{2}PI = 0 \\
\Rightarrow\quad &q^* = \sqrt{\frac{2C_3R}{IP}}
\end{aligned} \tag{20.1.5}$$

Therefore,

$$C(q^*) = \sqrt{2C_3RPI} + PR \tag{20.1.6}$$

Purchase Cost (P) per item	Range of quantity
P_1	$1 \leq q_1 \leq b$
P_2	$q_2 \geq b$

Table 20.1

20.1.1 Model IV(a): Purchase inventory model with one price break

Consider the table 20.1

where b is the quantity at and beyond which the quantity discount applies. Obviously, $P_2 < P_1$. For any purchase quantity q_1 in the range $1 \leq q_1 < b$,

$$C(q_1) = \frac{C_3 R}{q_1} + P_1 R + \frac{P_1 q_1 I}{2} \tag{20.1.7}$$

Similarly, for q_2,

$$C(q_2) = \frac{C_3 R}{q_2} + P_2 R + \frac{P_2 q_2 I}{2} \tag{20.1.8}$$

Rule I Compute q_2^*, using (20.1.5). If $q_2 \geq b$, then the optimum lot size will be q_2^*.

Rule II If, $q_2 < b$, then the quantity discount no longer applies to the purchase quantity q_2^*. Compute q_1^*, then compare $C(q_1^*)$ and $C(b)$ given by,

$$\begin{aligned} C(q_1^*) &= \frac{C_3 R}{q_1^*} + P_1 R + \frac{P_1 q_1^* I}{2} \\ C(b) &= \frac{C_3 R}{b} + \frac{P_2 I b}{2} + P_2 R \end{aligned}$$

It shows that,

$$\frac{C_3 R}{b} + P_2 R < \frac{C_3 R}{q_1^*} + P_1 R \qquad [\text{since } q_1^* < b \text{ and } P_2 < P_1]$$

However, $\frac{P_2 I b}{2}$ may or may not be less than $\frac{P_1 I q_1^*}{2}$. Hence, we must compare the total cost. So, $q^* = b$.

Example 20.1.1. Find the optimum order quantity for a product for which the price breaks are as follows:

Quantity discount	Unit Cost (Rs.)
$0 \leq q_1 < 500$	10.00
$500 \leq q_2$	9.25

The monthly demand for a product is 200 units, the cost of storage is 2% of unit cost and cost of ordering is Rs. 350.

Solution.

$R = 200$ units per month

$I = Rs.\ 0.02$

$C_3 = Rs.\ 350$

$P_1 = Rs.\ 10.00$

$P_2 = Rs.\ 9.25$

$$q_2^* = \sqrt{\frac{2C_3 R}{P_2 I}} = \sqrt{\frac{2 \times 350 \times 200}{9.25 \times 0.02}} = 870 \text{ units} > b = 500.$$

Since $q_2^* = 870$ lies within the range $q_2 \geq 500$, hence the optimum purchase quantity will be $q_2^* = 870$ units. ■

Example 20.1.2. Same as the previous example with $C_3 = Rs.\ 100$. Thus,

$$q_2^* = \sqrt{\frac{2C_3R}{P_2I}} = \sqrt{\frac{2 \times 100 \times 200}{9.25 \times 0.02}} = 447 \text{ units} < b = 500.$$

Then compare C(447) with C(500), that is, the optimum cost of procuring the least quantity which will entitle or price break, that is,

$$C(q^*) = C(447) = Rs.\ 2090.42$$
$$C(500) = Rs.\ 1937.25.$$

Since C(500)<C(447), the optimum purchase quantity will be $q^* = b = 500$.

20.1.2 Model IV(b): Purchase inventory model with two price breaks

Purchase Cost P per item	Range of quantity
P_1	$1 \leq q_1 < b_1$
P_2	$b_1 \leq q_2 2 < b_2$
P_3	$b_2 \leq q_3$

Table 20.2

Consider the table 20.2, where b_1 and b_2 are the quantities which determine the price breaks. The working rule is as follows:

Step 1: Compute q_3^* and compare with b_2.

(i) If $q_3^* \geq b_2$, then the optimum purchase quantity is q_3^*.

(ii) If $q_3^* < b_2$, then go to step 2.

Step 2: Compute q_2^*, since $q_3^* < b_2$ and q_2^* is also less than b_2 because $q_1^* < q_2^* < \ldots < q_n^*$ in general. Thus, there are only two possibilities when $q_2^* < b_2$, that is, either $q_2^* \geq b_1$ or $q_2^* < b_1$.

(i) When $q_2^* < b_2$ but $\geq b_1$, then proceed as in case of one price- break only, that is, compare the cost $C(q_2^*)$ and $C(b_2)$ to obtain the optimum purchase quantity.

The quantity with least cost will naturally be optimum.

(ii) If $q_2^* < b_2$ and b_1, then go to step 3.

Step 3: If $q_2^* < b_2$ (and b_1 both). Then compute q_1^* which will satisfy the inequality $q_1^* < b_1$. In this case, compare the cost $C(q_1^*)$ with $C(b_1)$ and $C(b_2)$ both to determine the optimum purchase quantity.

Example 20.1.3. Find the optimum order quantity for a product for which the price breaks are in table 20.3. The monthly demand for a product is 200 units. The cost of storage is 2% of the unit cost. Cost of ordering is Rs. 350.

Quantity	:	$0 \le q_1 < 500$	$500 \le q_2 < 750$	$750 \le q_3$
Unit Price(Rs.)	:	10.00	9.25	8.75

Table 20.3

Solution.

$$R = 200 \text{ units per month}$$
$$I = Rs.\ 0.02$$
$$C_3 = Rs.\ 350$$
$$P_1 = Rs.\ 10.00$$
$$P_2 = Rs.\ 9.25$$
$$P_3 = Rs.\ 8.75$$
$$b_1 = 500$$
$$b_2 = 750$$
$$q_3^* = \sqrt{\frac{2 \times 350 \times 200}{8.75 \times 0.02}} = 894 > 750.$$

Thus, the optimum purchase quantity will be $q^* = 894$.

If we choose $C_3 = Rs.\ 100$, and all are the same, then

$$q_3^* = \sqrt{\frac{2 \times 100 \times 200}{8.75 \times 0.02}} = 478 < 750.$$

Step 2:

$$q_2^* = \sqrt{\frac{2 \times 100 \times 200}{9.25 \times 0.02}} = 465 < 500.$$

Again, we compute

$$q_1^* = \sqrt{\frac{2 \times 100 \times 200}{10 \times 0.02}} = 447 < 500.$$

Then, we compare C(447) with C(500) and C(750). Now,

$$C(447) = Rs.\ 2090.42, \quad C(500) = Rs.\ 1937.25, \quad C(750) = Rs.\ 1843.29$$

Thus, $C(750) < C(500) < C(q_1^*)$. This shows that, the optimum purchase quantity is $q^* = 750$ units. ■

20.2 Probabilistic Inventory Model

20.2.1 Instantaneous demand, no set up cost

Discrete Case

Find the optimum order level z which minimizes the total expected cost under the following assumptions

(i) t is the constant interval between orders. (daily, monthly, weekly, etc.)

(ii) z is the stock at the beginning of each period t

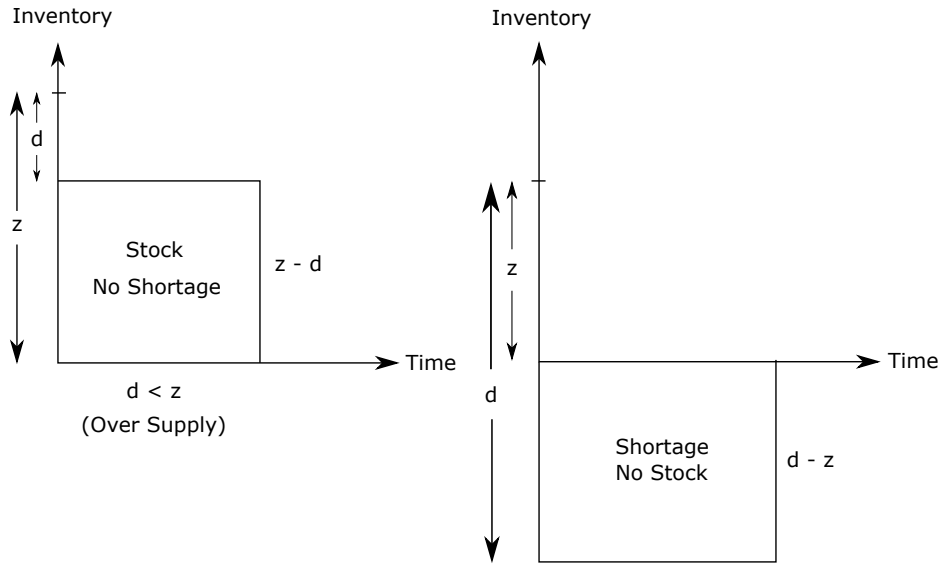
(iii) d is the estimated (random) demand at a discontinuous rate with probability P(d)

(iv) C_1 is holding cost

(v) C_2 is shortage cost

(vi) lead time zero

(vii) demand is instantaneous.



In the model with instantaneous demand, it is assumed that the total demand is fulfilled at the beginning of the period. Thus, depending on the demanded amount the inventory position may either be positive (surplus or stock) or negative (shortage).

Case I: $d \leq z$

$$\begin{aligned} \text{Holding cost} &= (z-d)\,C_1, \qquad \text{for } d \leq z \\ &= C_1 \times 0, \qquad \text{for } d > z \text{ (no stock)} \end{aligned}$$

Case II: $d > z$

$$\begin{aligned} \text{Shortage cost} &= C_2 \times 0 \qquad \text{for } d \leq z \text{ (no shortage)} \\ &= (d-z)\,C_2 \qquad \text{for } d > z \end{aligned}$$

To get the expected cost, we have to multiply the cost by given probability $P(d)$. Further to get the total expected cost we must sum over all the expected cost. So, the total expected cost per unit time is,

$$\begin{aligned} C(z) &= \sum_{d=0}^{z}(z-d)\,C_1 P(d) + \sum_{d=z+1}^{\infty} C_1 \cdot 0 \cdot P(d) + \sum_{d=0}^{z} C_2 \cdot 0 \cdot P(d) + \sum_{d=z+1}^{\infty} C_2 \cdot (d-z)\,P(d) \\ &= \sum_{d=0}^{z}(z-d)\,C_1 P(d) + \sum_{d=z+1}^{\infty} C_2 \cdot (d-z)\,P(d) \end{aligned} \tag{20.2.1}$$

For the minimum of $C(z)$, the following must be satisfied:

$$\Delta C\,(z-1) < 0 < \Delta C\,(z) \text{ (finite difference Calculus)} \tag{20.2.2}$$

But, we can difference (20.2.1) under the summation sign for $d = z+1$, the following condition satisfied

$$C_1\{(z+1)-d\}P(d) = C_2(d-(z+1))P(d).$$

Now,

$$\begin{aligned}
\Delta C(z) &= C_1\sum_{d=0}^{z}[((z+1)-d)-(z-d)]P(d) + C_2\sum_{d=z+1}^{\infty}[(d-(z+1))-(d-z)]P(d) \\
&= C_1\sum_{d=0}^{z}P(d) - C_2\sum_{d=z+1}^{\infty}P(d) \\
&= C_1\sum_{d=0}^{z}P(d) - C_2\left[\sum_{d=0}^{\infty}P(d) - \sum_{d=0}^{z}P(d)\right] \\
&= (C_1+C_2)\sum_{d=0}^{z}P(d) - C_2. \quad \left[\text{since } \sum_{d=0}^{\infty}P(d) = 1\right]
\end{aligned}$$

$$\begin{aligned}
&\Delta C(z) > 0 \\
\Rightarrow\ &(C_1+C_2)\sum_{d=0}^{z}P(d) - C_2 > 0 \\
\Rightarrow\ &\sum_{d=0}^{z}P(d) > \frac{C_2}{C_1+C_2}
\end{aligned} \tag{20.2.3}$$

Similarly,

$$\begin{aligned}
&\Delta C(z-1) < 0 \\
&\sum_{d=0}^{z-1}P(d) < \frac{C_2}{C_1+C_2}.
\end{aligned}$$

Combining, we get

$$\sum_{d=0}^{z-1}P(d) < \frac{C_2}{C_1+C_2} < \sum_{d=0}^{z}P(d). \tag{20.2.4}$$

Example 20.2.1. **(Newspaper boy problem)** A newspaper boy buys papers for Rs. 2.60 each and sells them for Rs. 3.60 each. He can not return unsold newspapers. Daily demand has the following probability distribution (Table 20.4).

No. of customers	:	23	24	25	26	27	28	29	30	31	32
Probability	:	0.01	0.03	0.06	0.10	0.20	0.25	0.15	0.10	0.05	0.05

Table 20.4

If each day, demand is independent of the previous days, how many papers should be ordered each day?

Solution. Let z=The number of newspapers ordered per day and d=demand that is, the number that could be sold per day if $z \geq d$, $P(d)$=The probability that the demand will be equal to on a randomly selected day,

$$\begin{aligned} C_1 &= \text{Cost per newspaper} \\ C_2 &= \text{Selling price per newspaper.} \end{aligned}$$

If the demand d exceeds z, his profit would become equal to $(C_2 - C_1)z$, and no newspaper will be let unsold. On the other hand, if d does not exceed z, his profit becomes equal to $(C_2 - C_1)d - (z - d)C_1$, where $(C_2 - C_1)d$ is for the sold papers and $(z - d)C_1$ for the unsold papers. Then the expected net profit per day becomes equal to

$$P(z) = \sum_{d=0}^{z} (C_2 d - C_1 z)P(d) + \sum_{d=z+1}^{\infty} (C_2 - C_1)zP(d).$$

where, $(C_2 d - C_1 z)P(d)$ is for $d \leq z$ and $(C_2 - C_1)zP(d)$ for $d > z$.

Using finite difference calculus, we know that the condition for maximum value of $P(z)$ is

$$\Delta P(z-1) > 0 > \Delta P(z).$$

$$\begin{aligned} \Delta P(z) &= \sum_{d=0}^{z} \left[\{C_2 d - C_1(z+1)\} - (C_2 d - C_1 z)\right] P(d) + \sum_{d=z+1}^{\infty} (C_2 - C_1)\{(z+1) - z\}P(d) \\ &= -C_1 \sum_{d=0}^{z} P(d) + (C_2 - C_1) \sum_{d=z+1}^{\infty} P(d) \\ &= -C_1 \sum_{d=0}^{z} P(d) + (C_2 - C_1) \left\{ \sum_{d=0}^{\infty} P(d) - \sum_{d=0}^{z} P(d) \right\} \\ &= -C_1 \sum_{d=0}^{z} P(d) + (C_2 - C_1) \left\{ 1 - \sum_{d=0}^{z} P(d) \right\} \\ &= -C_2 \sum_{d=0}^{z} P(d) + (C_2 - C_1). \end{aligned}$$

For, maximum of $P(z)$,

$$\begin{aligned} & \Delta P(z) < 0 \\ \text{or,} \quad & -C_2 \sum_{d=0}^{z} P(d) + (C_2 - C_1) < 0 \\ \text{or,} \quad & \sum_{d=0}^{z} P(d) > \frac{C_2 - C_1}{C_2}. \end{aligned} \tag{20.2.5}$$

Similarly, we can find,

$$\sum_{d=0}^{z-1} P(d) < \frac{C_2 - C_1}{C_2}.$$

Combining, we get,

$$\sum_{d=0}^{z} P(d) > \frac{C_2 - C_1}{C_2} > \sum_{d=0}^{z-1} P(d).$$

In this problem, $C_1 = Rs.\ 2.60,\ C_2 = Rs.\ 3.60$. The lower limit for demand d is 23 and upper limit is 32. Therefore, substituting these values in (20.2.5), we get,

$$\sum_{d=0}^{z} P(d) > \frac{3.60 - 2.60}{3.60} = 0.28.$$

Now, we can easily verify that this inequality holds for $z = 27$, that is,

$$\begin{aligned} \sum_{d=23}^{27} P(d) &= P(23) + P(24) + P(25) + P(26) + P(27) \\ &= 0.01 + 0.03 + 0.06 + 0.10 + 0.20 = 0.40 > 0.28. \end{aligned}$$

Similarly,

$$\sum_{d=23}^{26} P(d) = 0.20 < 0.28.$$

■

Continuous Case

This model is same as the previous model except that the stock levels are now assumed to be continuous quantities. So, instead of probability $P(d)$, we shall have $f(x)dx$ and in place of summation, we take integration, where $f(x)$ is the pdf (probability density function). The cost equation for this model becomes

$$C(z) = C_1 \int_0^z (z - x) f(x) dx + C_2 \int_z^\infty (x - z) f(x) dx. \tag{20.2.6}$$

The optimal value of z is obtained by equating z to zero the first derivative of $c(z)$, that is, $\dfrac{dC}{dz} = 0$.

Differentiating (20.2.6), we get,

$$\begin{aligned} \frac{dC}{dz} &= C_1 \int_0^z (1 - 0) f(x) dx + C_1 \left[(z - x) f(x) \frac{dx}{dz} \right]_0^z + C_2 \int_z^\infty (0 - 1) f(x) dx + C_2 \left[(x - z) f(x) \frac{dx}{dz} \right]_z^\infty \\ &= C_1 \int_0^z f(x) dx - C_2 \int_0^\infty f(x) dx \\ &= C_1 \int_0^z f(x) dx - C_2 \left[1 - \int_0^z f(x) dx \right] \\ &= (C_1 + C_2) \int_0^z f(x) dx - C_2. \end{aligned}$$

Thus,

$$\begin{aligned} & \frac{dC}{dz} = 0 \\ \Rightarrow\ & (C_1 + C_2) \int_0^z f(x) dx - C_2 \\ \Rightarrow\ & \int_0^z f(x) dx = \frac{C_2}{C_1 + C_2} \\ & \frac{d^2C}{dz^2} = (C_1 + C_2) \left[f(x) \frac{dx}{dz} \right]_0^z = (C_1 + C_2) f(x) > 0. \end{aligned}$$

Hence, we can get optimum value of z satisfying the sufficient condition for which the total expected cost C is minimum.

Example 20.2.2. A baking company sells cake by the kg weight, it makes a profit of Rs 5.00 per kg on each kg sold on the day it is baked. It disposes off all cakes not sold on the day it is baked at a loss of Rs. 1.20 per kg. If demand is known to be rectangular between 2000 and 3000 kgs, determine the optimal daily amount baked.

Solution.

$$\begin{aligned} C_1 &= \text{profit per kg cake} \\ C_2 &= \text{loss per kg cake for unsold cake} \\ x &= \text{Demand which is continuous with pdf } f(x), \end{aligned}$$

where,

$$\int_{x_1}^{x_2} f(x)dx = \text{the probability of an order within } x_1 \text{ to } x_2.$$

and z=stock level.

Then two cases arise.

Case I: If $x \leq z$, then clearly the demand x is satisfied and unsold $(z-x)$ quantities are returned with a loss of C_2 per kg, so, profit is C_1x and loss is $C_2(z-x)$. Hence the net profit becomes, $C_1x - C_2(z-x) = (C_1+C_2)x - C_2z$.

Case II: If $x > z$, then the net profit becomes C_1z. Thus, the total expected profit is given by

$$P(z) = \int_{x_1}^{z} \left[(C_1+C_2)x - C_2z\right] f(x)dx + \int_{z}^{x_2} C_1zf(x)dx = P_1(z) + P_2(z) \ \text{(say)}.$$

Now, for the maximum value of $P(z)$, we must have,

$$\frac{dP(z)}{dz} = \frac{d}{dz}P_1(z) + \frac{d}{dz}P_2(z) = 0$$

Now,

$$\begin{aligned} P_1(z) &= \int_{x_1}^{z} \left[(C_1+C_2)x - C_2z\right] f(x)dx \\ \frac{d}{dz}P_1(z) &= \int_{x_1}^{z} (0-C_2)f(x)dx + \left[\{(C_1+C_2)x - C_2z\}f(x)\frac{dx}{dz}\right]_{x_1}^{z} \\ &= -C_2\int_{x_1}^{z} f(x)dx + \{(C_1+C_2)x - C_2z\}f(x) \\ &= -C_2\int_{x_1}^{z} f(x)dx + C_1zf(z). \end{aligned}$$

Similarly,

$$\begin{aligned} \frac{d}{dz}P_2(z) &= \int_{z}^{x_2} C_1f(x)dx + \left[C_1zf(z)\frac{dx}{dz}\right]_{z}^{x_2} \\ &= C_1\int_{z}^{x_2} f(x)dx - C_1zf(z). \end{aligned}$$

Hence, we have,

$$\begin{aligned}
&\frac{dP(z)}{dz} = \left[-C_2\int_{x_1}^{z} f(x)dx + C_1 z f(z)\right] + \left[C_1\int_{z}^{x_2} C_1 f(x)dx - C_1 z f(z)\right] = 0 \\
\Rightarrow\quad & -C_2\int_{x_1}^{z} f(x)dx + C_1\int_{z}^{x_2} f(x)dx = 0 \\
\Rightarrow\quad & -C_2\int_{x_1}^{z} f(x)dx + C_1\left\{\int_{x_1}^{x_2} f(x)dx - \int_{x_1}^{z} f(x)dx\right\} = 0 \\
\Rightarrow\quad & -(C_1+C_2)\int_{x_1}^{z} f(x)dx + C_1 = 0 \\
\Rightarrow\quad & \int_{x_1}^{z} f(x)dx = \frac{C_1}{C_1+C_2} \qquad (20.2.7)
\end{aligned}$$

Also,

$$\frac{d^2P(z)}{dz^2} = -(C_1+C_2)f(z) < 0$$

satisfies the sufficient condition of maximum of $P(z)$.

In this problem,

$$C_1 = \text{Rs. } 5.00,\ \ C_2 = \text{Rs. } 1.20,\ \ x_1 = 2000,\ \ x_2 = 3000.$$

$$f(x) = \frac{1}{x_2 - x_1} = \frac{1}{1000}.$$

Substituting these values in equation (20.2.7), we have

$$\begin{aligned}
&\int_{2000}^{z} \frac{1}{1000}dx = \frac{5}{5+1.20} = 0.807 \\
\Rightarrow\quad & \frac{1}{1000}(z-2000) = 0.807 \\
\Rightarrow\quad & z = 2807 \text{ kg}.
\end{aligned}$$

■

References

1. An Introduction to Information Theory - F. M. Reza.
2. Operations Research : An Introduction - P. K. Gupta and D.S. Hira.
3. Graph Theory with Applications to Engineering and Computer Science - N. Deo.
4. Operations Research - K. Swarup, P. K. Gupta and Man Mohan.
5. Coding and Information Theory - Steven Roman.
6. Coding Theory, A First Course - San Ling r choaping Xing.
7. Introduction to Coding Theory - J. H. Van Lint
8. The Theory of Error Correcting Codes - Mac William and Sloane.
9. Information and Coding Theory - Grenth A. Jones and J. Marry Jones.
10. Information Theory, Coding and Cryptography - Ranjan Bose.