

DISCRETE STOCHASTIC PROCESSES

Lecture 11

Renewal and Renewal/Reward Processes – Chapter 3

Review: Little's Theorem

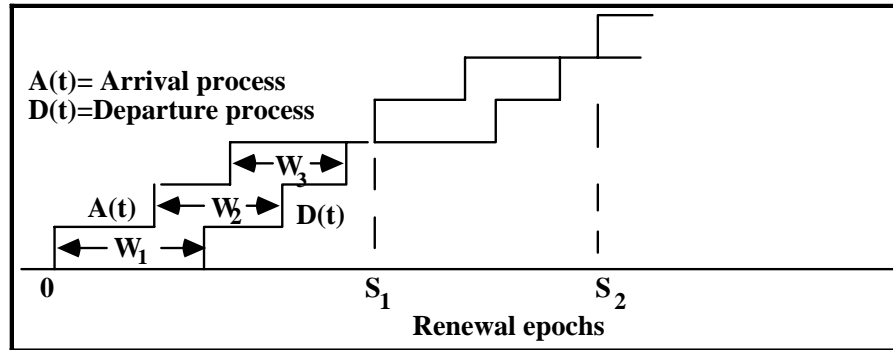
Average waiting time of M/G/1 queue - the PK Formula
“Pasta Property” - Poisson Arrivals See Time Averages
Mention “Delayed Renewal Processes”

Finite State Markov Chains – Chapter 4

Reading: Gallager, Chapter 4, Sections 4.1 – 4.3 (ignore proofs in 4.4.)

Classification of States

Little's Theorem



$L(t)$ = number of customers in system (service + queue) at time $t = R(t) = A(t) - D(t)$

Time averaged number of customers over $[0,t] = \bar{L}[0,t] = \frac{1}{t} \int_0^t L(\tau) d\tau$

W_n = wait (in queue and service) for n th customer. (Neither independent nor identically distributed! Why?)

Average wait for all customers arriving in $[0,t] = \bar{W}[0,t] = \frac{\sum_{k=1}^{A(t)} W_k}{A(t)}$

For t in empty period before S_1 : $\bar{L}[0,t] = \frac{1}{t} \int_0^t L(\tau) d\tau = \frac{1}{t} (W_1 + W_2 + W_3) = \frac{\sum_{k=1}^{A(t)} W_k}{t} = \frac{\sum_{k=1}^{A(t)} W_k}{A(t)} \frac{A(t)}{t}$

i.e., (time averaged occupancy) = (average wait per customer) * (average arrival rate)

$$\bar{L}[0,t] = \bar{W}[0,t] * \frac{A(t)}{t}$$

This is also true for t in any empty period before S_n , $n = 1, 2, 3, \dots$

Little's theorem asserts that this works in the limit of large t:

Little's Theorem (for G/G/1 Queue)

$$\lim_{t \rightarrow \infty} \bar{L}(t) = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t L(\tau) d\tau = \lim_{t \rightarrow \infty} \frac{\sum_{k=1}^{A(t)} W_k}{t} = \lim_{t \rightarrow \infty} \frac{\sum_{k=1}^{A(t)} W_k}{A(t)} \lim_{t \rightarrow \infty} \frac{A(t)}{t}, \quad \text{i.e.,}$$

$$\bar{L} = \bar{W} * \lambda$$

\bar{L} = time averaged number of customers in system (queue + service)

$$= \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t L(\tau) d\tau$$

\bar{W} = the average wait (in service + queue) of a customer

$$= \lim_{t \rightarrow \infty} \frac{\sum_{k=1}^{A(t)} W_k}{A(t)} = \lim_{n \rightarrow \infty} \frac{\sum_{k=1}^n W_n}{n}$$

$$\lambda = \text{customer average arrival rate} = \lim_{t \rightarrow \infty} \frac{A(t)}{t}$$

If, in addition, the **customer arrival process is non-arithmetic**, then the **system renewal process** is non-arithmetic, and

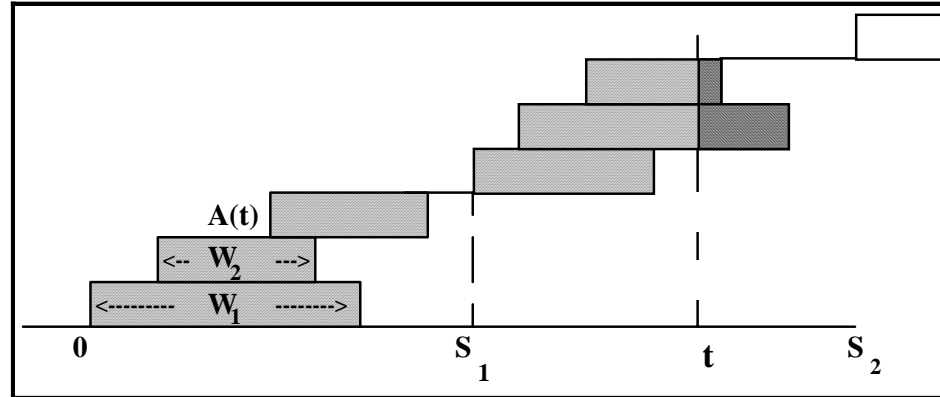
$$\bar{L} = \bar{W}\lambda$$

where it is also true for \bar{L} and \bar{W} that:

$$\bar{L} = \lim_{t \rightarrow \infty} E[L(t)]$$

$$\bar{W} = \lim_{t \rightarrow \infty} E[W(t)]$$

The argument does not depend on FCFS (FIFO) service.



By same argument,

$$\text{Average \# in queue} = (\text{Average wait in queue}) \cdot (\text{Arrival rate})$$

$$\text{Average \# in service} = (\text{Average time in service}) \cdot (\text{Arrival rate})$$

For single server, Average # in service = percentage of time server is busy

Very generally for queues,

$$\text{Server utilization} = (\text{Average time in service}) \cdot (\text{Arrival rate})$$

This argument also does not depend on the system being G/G/1. It works equally well for ***time in queue*** or ***time in system*** for a G/G/k system, provided the system is stable, i.e., expected time for system to empty out is finite (i.e., enough server capacity to handle the incoming crowd).

Time Average Waiting Time in queue for $M/G/1$ (The Pollaczek-Khinchin (PK) Formula)

$$\bar{W}^{(q)} = \frac{\lambda E[Z^2]}{2(1 - \lambda E[Z])}$$

$W^{(q)}$ is the s.s. expected wait in queue. Arrival process, $A(t)$, is Poisson with rate λ
Service time, Z_n , of n^{th} customer is independent of all arrival epochs and of all other service times.

We pick the renewal process defined by renewals occurring on arrivals to an empty system.

First form a reward processes from the residual life (i.e., residual service time) of the customer in service:

$R(t) = 0$ if no customer in service; $R(t) = x - t$ if customer in service at t departs at x , i.e., $R(t)$ is the residual life of the customer in service at time t . $L^{(q)}(t)$ is the # of customers in the queue at time t .

Let $U(t)$ be *the unfinished work in system at time t* ; this is $R(t)$ plus the aggregate amount of service time required for each customer in queue. It equals the waiting time in queue for a new customer if that customer were to arrive at time t

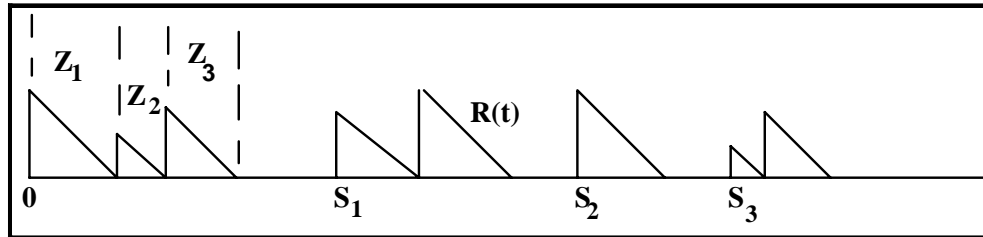
$$U(t) = R(t) + \sum_{i=1}^{L^{(q)}(t)} Z_i \quad (1)$$

$L^{(q)}(t)$ depends on the arrival epochs up to time t , and on the service times of customers that have entered service before t . $Z_1(t), \dots, Z_{L^{(q)}(t)}(t)$ are independent of these arrival epochs and service times, so are independent of $L(t)$ so

$$E[U(t)] = E[L^{(q)}(t)]E[Z] + E[R(t)] \quad (2)$$

We will first try to find:

$$\lim_{t \rightarrow \infty} E[U(t)] \quad (3)$$



$$\lim_{t \rightarrow \infty} \frac{\int_0^t R(\tau) d\tau}{t} = \lim_{t \rightarrow \infty} \frac{\sum_{i=1}^{A(t)} Z_i^2 / 2}{A(t)} = \frac{\lambda E[Z_i^2]}{2}$$

Renewal process is non-arithmetic because of Poisson arrivals. Thus by the Key Renewal Theorem,

$$\lim_{t \rightarrow \infty} E[R(t)] = \lambda E[Z^2] / 2 \quad (4)$$

Substituting (4) into (2) gives us, so far:

$$\lim_{t \rightarrow \infty} E[U(t)] = \lim_{t \rightarrow \infty} E[L^{(q)}(t)] E[Z] + \lambda E[Z^2] / 2 \quad (5)$$

From Little's Theorem we know that

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t L^{(q)}(\tau) d\tau = \lambda \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n W_n^{(q)}$$

Using the **Key Renewal Theorem** on **Little's Theorem** gives us

$$\lim_{t \rightarrow \infty} E \left[L^{(q)}(t) \right] = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t L^{(q)}(\tau) d\tau = \lambda \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n W_n^{(q)} = \lambda \lim_{n \rightarrow \infty} E \left[W_n^{(q)} \right] \quad (6)$$

The last equality takes some work as given by Exercise 3.31 of Gallager. (The SLLN is not directly applicable since the $W_n^{(q)}$ are not IID.)

In this problem $E \left[W_n^{(q)} \right]$ = expected wait **in the queue** of n^{th} customer = $E \left[U(t) \mid \text{the } n^{\text{th}} \text{ arrival epoch occurs instantaneously after (at) } t \right]$. Substituting (6) into (5) yields

$$\lim_{t \rightarrow \infty} E \left[U(t) \right] = \lambda \lim_{n \rightarrow \infty} E \left[W_n^{(q)} \right] + \lambda E \left[Z^2 \right] / 2 \quad (7)$$

$$\lim_{t \rightarrow \infty} E[U(t)] = \lambda \lim_{n \rightarrow \infty} E[W_n^{(q)}] + \lambda E[Z^2] / 2 \quad (7)$$

This is true for G/G/1 queues with non-arithmetic interarrivals.

We only now will use the memoryless property. For memoryless arrivals, the event of an arrival in $(t, t + d)$ is independent of $U(t)$ (and of everything else at t).

Thus $U(t)$ independent of whether or not t is an arrival epoch, so, for memoryless arrivals

$$\bar{W}^{(q)} = \lim_{n \rightarrow \infty} E[W_n^{(q)}] = \lim_{t \rightarrow \infty} E[U(t)],$$

i.e., the "Pasta" property holds. ,

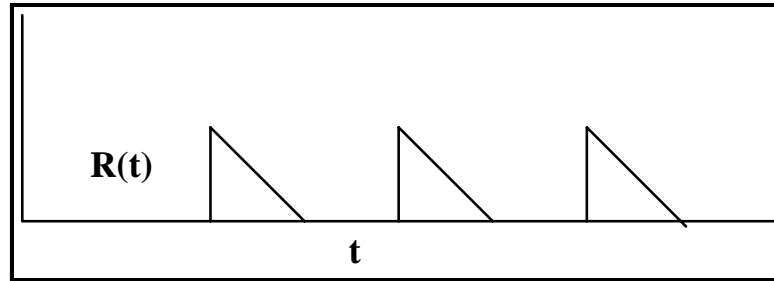
$$\bar{W}^{(q)} = \lambda E[Z] \bar{W}^{(q)} + \lambda E[Z^2] / 2$$

The Pollaczek-Khinchin (PK) Formula

$$\bar{W}^{(q)} = \frac{\lambda E[Z^2]}{2(1 - \lambda E[Z])}$$

Example 1: (Why the Poisson arrivals are needed).

Assume service times uniform in $(1 - \varepsilon, 1 + \varepsilon]$; interarrival times uniform in $(2 - \varepsilon, 2 + \varepsilon]$, $\varepsilon < 1/2$.



$$\lim_{t \rightarrow \infty} E[R(t)] = \lambda \bar{Z}^2 / 2 = (1 + \varepsilon^2 / 3) / 4$$

No customers ever wait in queue, so $E[W^{(q)}(t)] = 0$, but

$$\lim_{t \rightarrow \infty} E[U(t)] = (1 + \varepsilon^2 / 3) / 4.$$

Given that a customer is about to arrive, the previous customer left service about 1 unit of time previously, and the unfinished work is zero. The waiting time seen by an arriving customer does **not** equal average unfinished work.

The PK Formula gives us queueing truisms: The rate of arrivals must be less than the rate of departures or the queue blows up. As the rate of arrivals approaches the rate of departures the waiting time becomes infinite. This gives the queueing curve.

Even when the arrival rate is low enough, the expected waiting time may be infinite if $E[Z^2]$ is infinite. This is like a residual time effect. If someone takes a long time there are a lot of customers (proportional to the long waiting time) held up that long time.

Example 2: Assume Z is memoryless ($M/M/1$ queue). Then $\bar{W}^{(q)} = \frac{\lambda \bar{Z}^2}{1 - \lambda \bar{Z}}$

With service time mean $= 1/\mu$, $\bar{W} = \frac{(\lambda/\mu^2)}{1 - (\lambda/\mu)}$

Example 3: Assume Z is deterministic with service time d ($M/D/1$). Then

$$\bar{W}^{(q)} = \frac{\lambda d^2}{2(1 - \lambda d)}$$

FINITE STATE MARKOV CHAINS

Definition: A **Finite State Markov Chain** is an Integer Time Process, $\{X_n; n \geq 0\}$ in which X_n , for each $n \geq 0$ is a random variable with possible values $\{1, 2, \dots, J\}$ with the **Markov Property**

Here we restrict each rv X_n to take on values from a finite set of values. The names of these values are unimportant, so we take them to be $\{1, 2, \dots, J\}$.

Definition: **The Markov Property**

$$P(X_n = j | X_{n-1} = i, X_{n-2} = h, \dots, X_0 = m) = P(X_n = j | X_{n-1} = i) = P_{ij} \text{ for all } n, i, j, h, m, \dots$$

The definition does two things: X_n depends on the past only through X_{n-1} , and also the probabilities don't depend on n for $n \geq 1$ (homogeneous Markov chain).

THE NOTION OF STATE

X_n is called the **state** at time n . This characterizes everything from the past that is relevant for the future.

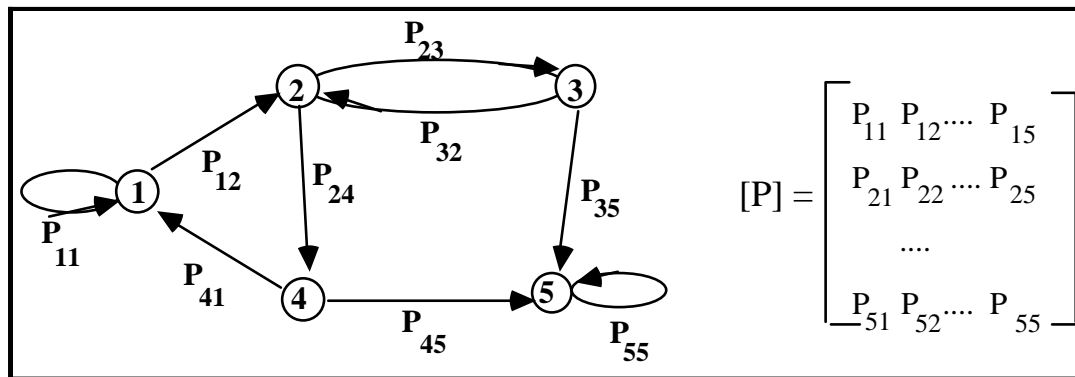
$$P(X_{n+1}, X_{n+2}, \dots | X_n, X_{n-1}, \dots) = P(X_{n+1} | X_n) P(X_{n+2} | X_{n+1}) \dots$$

Extension: Finite time memory. Suppose Y_n depends on Y_{n-1}, Y_{n-2} . Let X_n be the pair (Y_n, Y_{n-1}) .

$$P(X_n | X_{n-1}, X_{n-2}) = P(Y_n, Y_{n-1} | Y_{n-1}, Y_{n-2}, Y_{n-3}) = P(Y_n, Y_{n-1} | Y_{n-1}, Y_{n-2}) = P(X_n | X_{n-1})$$

Thus $\{X_n, n \geq 0\}$ Markov.

A Markov chain is completely described by set of transition probabilities P_{ij} plus initial probabilities $P(X_0)$. Sometimes view $\{P_{ij}\}$ graphically, sometimes as matrix.



The graph emphasizes the possible and impossible.

Let: $P_{ij}^n = P(X_n = j | X_0 = i)$

$$P_{ij}^2 = \sum_k P(X_2 = j, X_1 = k | X_0 = i)$$

$$= \sum_k P(X_2 = j | X_1 = k) P(X_1 = k | X_0 = i) = \sum_k P_{ik} P_{kj}$$

Note that this is the i, j term in the matrix product $[P] \cdot [P]$. Similarly P_{ij}^n is the i, j term in the matrix product $[P]^n$. Thus the matrix representation is useful in calculations.

$P_{ij}^n = 0$ iff no walk of n steps exists from i to j in graph. This can often be seen by inspection from the graph.

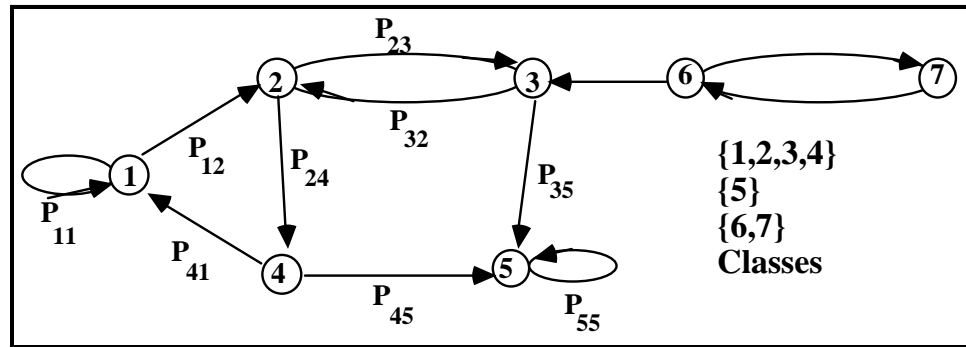
Classification of States

Definitions: State j is **accessible** from i ($i \rightarrow j$) if path from i to j in graph. States i and j **communicate** if ($i \leftrightarrow j$) if ($i \rightarrow j$) and ($j \rightarrow i$).

If ($i \leftrightarrow j$) and ($j \leftrightarrow k$) then ($i \leftrightarrow k$). Given a walk from i to j , and a walk from j to k , walk from i can be extended to go to k and $i \rightarrow k$. Similarly $k \rightarrow i$.

Definition: A **Class** of states is a non-empty set S of states such that $(i \leftrightarrow j)$ for each $i \in S, j \in S$ and also for $i \in S$, no $j \notin S$ with $(i \leftrightarrow j)$.

A class can be thought of as a **maximal** set of communicating states.



Definitions: A class S is **recurrent** if no $j \notin S$ is accessible from any $i \in S$.

A recurrent class can be thought of as a “trapping” class – once you get in, you never get out.

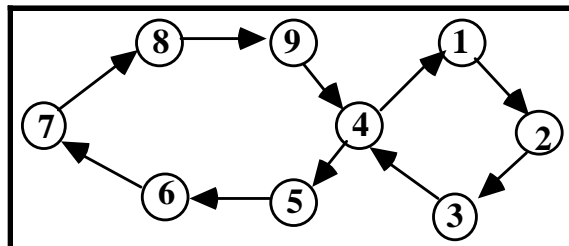
A class that is not recurrent is **transient**:

$\{5\}$ recurrent, $\{1, 2, 3, 4\}, \{6, 7\}$ transient

Transient means there is an outgoing edge from the class and no possible return.

States also classified in terms of **periodicity**.

Definition: The **period**, $d(i)$, of state i is defined as: $d(i) = \gcd \{n : P_{ii}^n > 0\}$

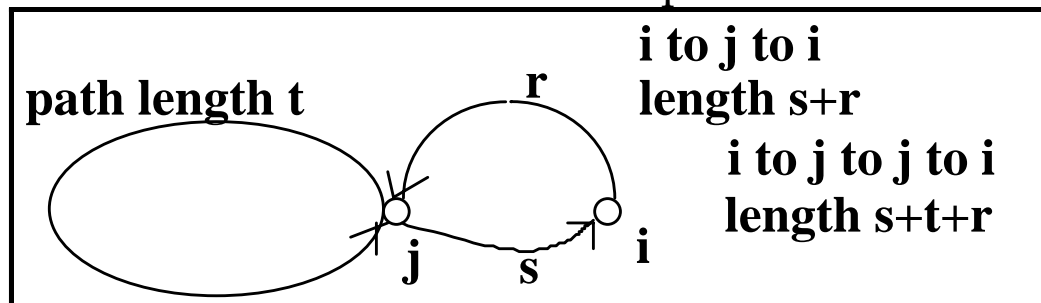


For example, $P_{44}^n > 0$ for $n = 4, 6, 8, 10, \dots$. The greatest common divisor is 2, $d(4) = 2$.

For state $i = 1$, $P_{11}^n > 0$ for $n = 4, 8, 10, 12, \dots$. $d(1) = 2$.

If $d(i) = 1$, i is defined to be aperiodic; otherwise it is periodic with period $d(i)$.

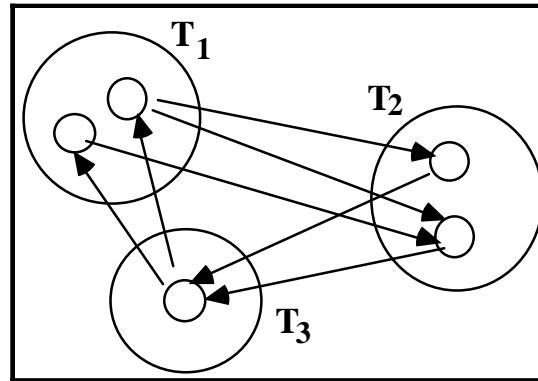
Theorem: All states in the same class have the same period.



Proof: Suppose i and j in same class, $P_{ij}^r > 0$, $P_{ji}^s > 0$, and let t be any integer with $P_{jj}^t > 0$.

Then $d(i)$ divides $r + t + s$ and also divides $r + s$. Thus $d(i)$ divides t for all t such that $P_{jj}^t > 0$. Thus $d(i) \leq d(j)$. Reverse roles of i and j to get $d(j) \leq d(i)$.

Theorem: If period of a class is $d > 1$, then class has partition T_1, T_2, \dots, T_d and all $i \in T_m$ have transitions only to T_{m+1} (or T_1 if $m = d$).



Proof: Let d be period of state 1. For $1 \leq m \leq d$, define

$$T_m = \left\{ j : P_{1j}^{nd+m} > 0 \text{ for some } n > 0 \right\}$$

For given j , assume $P_{1j}^{r'} > 0$ and $P_{j1}^s > 0$. Then a walk (through j) of $r + s$ steps exists from 1 to 1, and $r + s$ is divisible by d . If $P_{1j}^{r'} > 0$ also, $r' + s$ divisible by d , $r - r'$ divisible by d . If $r = nd + m$, then $r' = n'd + m$, so $j \in T_m$ for exactly one m .

Definition: A chain is **ergodic** if it has only one class, and that class is aperiodic.

Theorem: An ergodic chain of J states has $P_{ij}^m > 0$ for all i, j and all $m \geq J(J - 1)$.

Proof: For some fixed i , let $T(m) = \{j : P_{ij}^m > 0\}$. Note that $i \in T(n)$ for some $n \leq J$ and that $\{i\} \subseteq T(n) \subseteq T(2n) \dots$

This sequence of sets strictly increases up to a certain point (it can increase at most $J - 1$ times) and then remains constant.

$$T((J - 1)n) = T(Jn)$$

$T((J - 1)n)$ contains all J states. The rest of the proof (Theorem 4.4 in text) shows this.