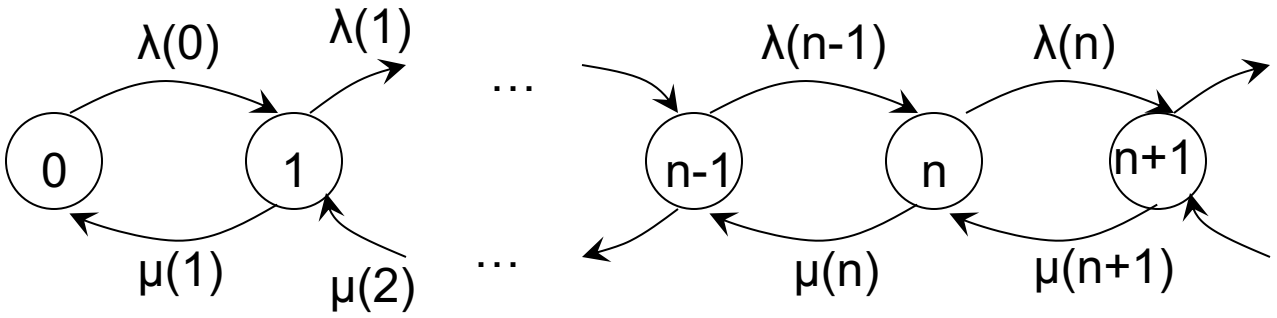
Chapter 5
Elementary Stochastic Analysis

Prof. Ali Movaghar

Birth and Death Processes

- $q_{k,k+1} = \lambda(k)$: Arrival (birth) rate in state k
- $q_{k,k-1} = \mu(k)$: Departure (death) rate in state k
- $q_{i,j} = 0$: for $|i-j|>1$
- $-q_{kk} = [\lambda(k) + \mu(k)]$



Birth and Death Processes (Con.)

- The rate arrival depends on the current system rate
  - Therefore interarrival times must be exponentially distributed.
  - Similarly, service-time distribution must also be exponential.
  - The probability of more than one arrival or service completion in Δt is negligiable.

Birth and Death Processes (Con.)

- The system is classical M/M :
  - Prob(one arrival in Δt | system is in state k)=$\lambda(k)\Delta t$
  - Prob(one service in Δt | system is in state k)=$\mu(k)\Delta t$
  - Prob(no arrival/services in Δt| system is in state k)= 1- $\lambda(k)\Delta t$ - $\mu(k)\Delta t$

Birth and Death Processes (Con.)

- P(n,t) : Probability of finding n customers in the system at time t.
- There are three ways for the system to be in state n at time t :
  - System is in state n-1 at time t and one arrival occurs at Δt
  - system is in state n+1 at time t and a service completion occurs during Δt
  - System is in state n at time t and no arrival/service completion occurs during Δt

Birth and Death Processes (Con.)

- $P(n, t+\Delta t) = P(n-1,t)\lambda(n-1)\Delta t + P(n+1,t)\mu(n+1)\Delta t + P(n,t)[1-\lambda(n)\Delta t - \mu(n)\Delta t]$
- Taking limit $\Delta t \rightarrow 0$, we get

$$\frac{dP(n,t)}{dt} = \lambda(n-1)P(n-1,t) + \mu(n+1)P(n+1,t) - [\lambda(n)+\mu(n)]P(n,t)$$

$$\frac{dP(0,t)}{dt} = \mu(1)P(1,t) - \lambda(0)P(0,t)$$

- The above differential difference equations describe the transient behavior of M/M system

Birth and Death Processes (Con.)

- To examine the steady-state solution, the system should be ergodic :

$$\exists k \, \forall n > k \, [\lambda(n) / \mu(n+1) < 1]$$

  - The work should be handled faster than it arrives

Birth and Death Processes (Con.)

- For steady state, we $\frac{dP(n,t)}{dt} = 0$ and P(n,t) is shown by P(n)
  - P(n-1)λ(n-1)+ P(n+1)μ(n+1)=P(n) [λ(n)+ μ(n)]
  - Under steady state, the effective rate with which the system enters state n should be equal to the effective rate with which it exits state n

Birth and Death Processes (Con.)

- We can derive equation in a more genera form:
  - For any closed boundary, the effective flow inward must equal the effective flow outward. (Global balance equation)
  - If the boundary contains states 0 through n-1, we get :

$$P(n-1)\lambda(n-1) = P(n)\mu(n)$$

so

$$P(n) = \frac{\lambda(0)\lambda(1)...\lambda(n-1)}{\mu(1)\mu(2)...\mu(n)} P(0)$$

Steady-State Analysis of M/M Systems

- We will derive detailed results for several important M/M systems
- Knedall's notation : M/G/c/FCFS/K/N
  - M : Poisson arrival
  - G : general service-time distribution
  - c : identical servers
  - FCFS : scheduling discipline
  - Storage capacity : K
  - N : population

Simple M/M/1/SI/∞/∞ Queue

- Here both λ(n) and μ(n) are independent of state n
  - Let $\rho=\lambda/\mu$, for ergodicity $\rho<1$
  - $P(n)= \rho^n P(0)$,
  - $\sum_{n=0}^{\infty} \rho^n P(0) = 1$ results $P(0)=(1-\rho)$
  - From two above : $P(n)= (1-\rho)\ \rho^n$
  - In a simple M/M/1 system, the queue length distribution is geometric with parameter ρ.

Simple M/M/1/SI/∞/∞ Queue (Con.)

- Various performance parameters can now be obtained :
  - Utilization (U) = 1- P(0) =ρ
  - Avg. queue length (Q) = E(X) = Σ nP(n) = $\rho/(\rho-1)$
  - Avg. response time (R)=$_{\text{little low}}$ E(X)/λ=1/$\mu(1-\rho)$
  - Avg. number waiting (L) = Σ (n-1) P(n)= $\rho^2/(1-\rho)$
  - Avg. waiting time (W) =$_{\text{little low}}$ L/λ = $\rho^2/\lambda(1-\rho)$

M/M/c/SI/∞/∞ Queue

- This is the multiple-server extension of M/M/1 system.
  - $\lambda(n) = \lambda$
  - $\mu(n) = \begin{cases} n\mu_0 & \text{for } n < c \\ c\mu_0 & \text{for } n \geq c \end{cases}$ where μ_0 is the basic service rate
  - $\rho = \lambda / c\mu_0$ so we get

$$P(n) = \begin{cases} \dfrac{\lambda^n P(0)}{\mu_0{}^n n!} = \dfrac{(c\rho)^n}{n!} P(0) & \text{for } n < c \\ \dfrac{\lambda^n P(0)}{\mu_0{}^n c! c^{n-c}} = \dfrac{(c\rho)^n}{c! c^{n-c}} P(0) & \text{for } n \geq c \end{cases}$$

M/M/c/SI/∞/∞ Queue (Con.)

- Using $\sum_{n=0}^{\infty} P(n) = 1$, we can compute P(0)

$$P(0)^{-1} = \sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} + \frac{(c\rho)^c}{c!(1-\rho)}$$

Simple M/G/∞/SI/∞/∞ Queue

- The arrival rate is state independent
  - $\lambda(n) = \lambda$
  - $\mu(n)=n\,\mu_0$

$$P(n) = \frac{\lambda^n}{\mu_0^n n!} P(0) = \frac{\rho^n P(0)}{n!}$$

  - $\rho=\lambda/\mu_0$. P(0) can be computed by the requirement that all probabilities sum to 1.

$$P(n) = \frac{\rho^n}{n!} e^{-\rho} \quad \text{since} \quad \sum_{n=0}^{\infty} \rho^n / n! = e^{\rho}$$

Simple M/G/∞/SI/∞/∞ Queue (Con.)

- Thus the distribution is Poisson with mean ρ.
  - The Avg. queue length $Q=\rho$
  - Utilization $U=\rho$
  - By little ‘s law Avg. Response time $R=1/\mu_0$

Simple M/M/c/SI/K/∞ Queue

- This is the finite storage case where system hold at most K customers:
  - $$\lambda(n) = \begin{cases} \lambda_0 & \text{for } n < K \\ 0 & \text{for } n \geq K \end{cases}$$
  - $\mu(n)=\min(n,c,K)\mu_0$
  - P(n) can be easily abstained as before.
  - Note that this system has only min(c,K)+1 states.

Finite Population Systems

- Let N denote the total number of customers in the "universe".
- the arrival rate λ depends on n and the underlying physical situation.
  a) Each customer has its own independent arrival rate λ_0.
     - Thus the overall arrival is proportional to the number of customers left in universe: $\lambda(n)=(N-n)\ \lambda_0$
  b) The customers are released sequentially. Thus arrival rate is constant

Finite Population Systems (Con.)

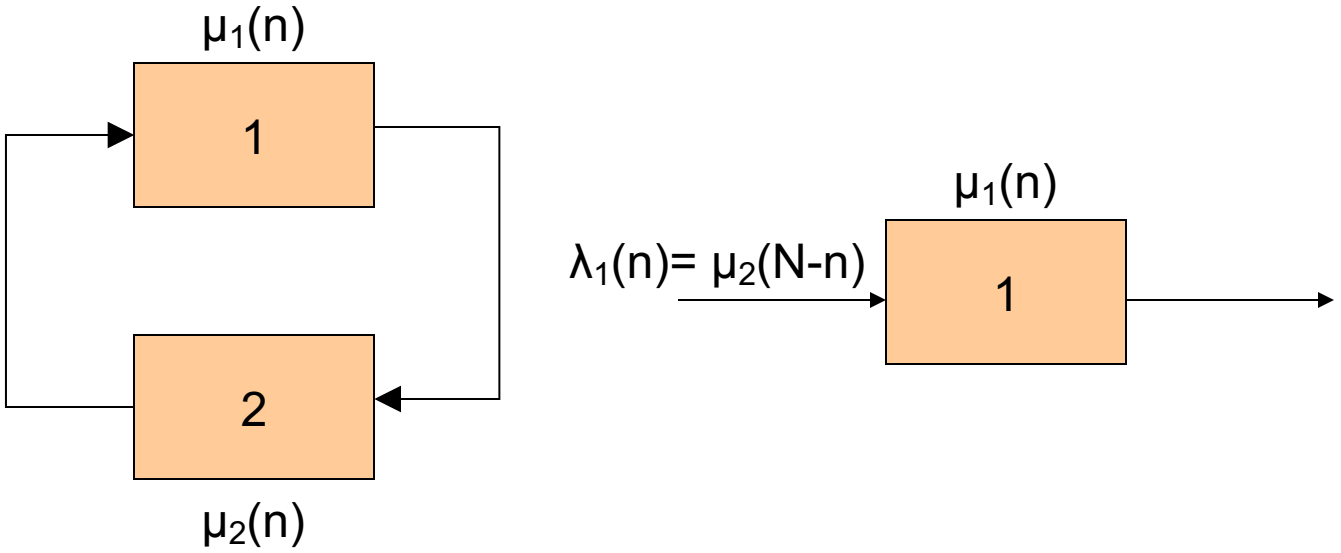
- As an example, consider M/M/K/SI/K/N station with arrival model (a).
  - $\mu(n)=\mu_0$

$$\lambda(n)=\begin{cases}(N-n)\lambda_0 & \text{for } n<\min(N,K)\\ 0 & \text{for } n\geq\min(N,K)\end{cases}$$

  - The state probabilities and performance parameters could be obtained as before.
  - Note that system suffers from blocking if K<N

Finite Population Systems (Con.)

- An isolated station with finite population can be view as a closed network of two stations :



Example

- A telephone exchange is to be set up for a small community of 50 customers, each of whom independently attempts to call at the average rate of one per hour and talks for 12 minutes on the average. The call attempts (including reattempts by customers who get a busy tone) can be adequately described by Poisson process. System has a capacity of K voice channel. Find the minimum value of K such that the probability of service denial is 2% or less.

Example (Con.)

- **Solution** : The system can be modeled as a M/G/K/SI/K/N, with N=50
  - $\lambda(n) = \begin{cases} (N-n)\lambda_0 & \text{for } n < \min(N,K) \\ 0 & \text{for } n \geq \min(N,K) \end{cases}$,λ_0=1/60
  - $\mu(n)=n\mu_0$, μ_0= 1/12
  - $\rho=\lambda_0/\mu_0=0.2$
  - So we get

$$P(n) = \rho^n \binom{50}{n} P(0)$$

Example (Con.)

- The probability reject call P(K) is given by

$$P(K) = \rho^K \binom{50}{K} \left[\sum_{n=0}^{K} \rho^n \binom{50}{n} \right]^{-1}$$

- Setting $P(K) \leq 0.02$ and solving the equation we get K=14.

Response-Time Distribution

- We briefly discuss how to determine response-time distribution for M/M systems
- The general approach is to pick a tagged customer and account for all the delays it encounters.

Response-Time Distribution (Con.)

- Suppose M/M/1 system
  - Let R denote the response time for a tagged customer
  - Let X denote the number of customers that it finds on its arrival
  - So the distribution function R is:

$$F_R(t) = \Pr(R \le t \mid X = n) = \sum_{n=0}^{\infty} \Pr(R \le t \mid X = n) \Pr(X = n)$$

Response-Time Distribution (Con.)

- Let R(n) denote response time of the tagged customer given that if finds n customers ahead of iteself.
  - $R(n)= R_1 + S_2 +\ldots+ S_n + S_{n+1}$
    - R_1 is the remaining service time of the customer currently receiving service
    - Each S_i has exponential distribution with mean $1/\mu$
    - Distribution of R_1 is the same as S_i (memoryless property of exponential distribution)
    - R(n) is sum of n+1 independent, identically distributed random variables

Response-Time Distribution (Con.)

- Let f denote the density function of R:

$$f_{R(t)} = \sum f_{R(t)}(t)\Pr(X = n) = \sum_{n=0}^{\infty}\left[\frac{\mu(\mu t)^n}{n!}e^{-\mu t}\right](1-\rho)\rho^n =$$

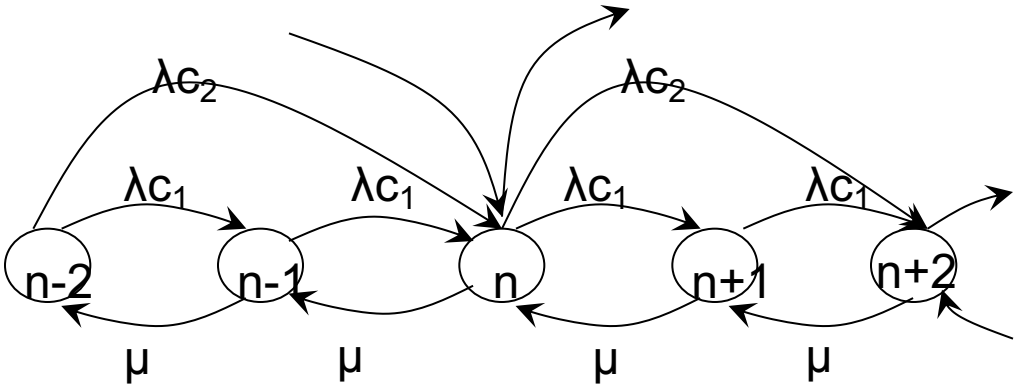
$$= \mu(1-\rho)e^{-\mu t}\sum_{n=0}^{\infty}\frac{(\rho\mu t)^n}{n!} = \mu(1-\rho)e^{-\mu(1-\rho)t}$$

  - The response-time distribution is also exponetial
  - The mean average response time is $1/\mu(1-\rho)$

Batch Systems and Method of Stages

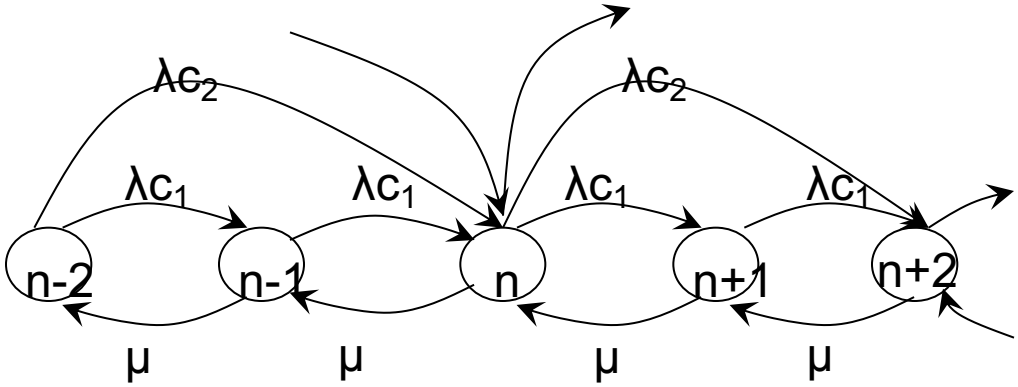
- We examine M/M systems with batch arrivals and services
  - Customers may arrive or get served as a group with a number of K, where K can be a random variable : M^k/M or $M^{(x)}/M$.
  - Server may have K stages, each has an exponential service time distribution with rate kμ. So the overall service time distribution is Erlang with the rate equal to μ : M/M^k
  - The behavior of $M^k/M/1$ is similar to $M/M^k/1$

Analysis of Batch Systems



- The batch size is a random variable denoted C.
  - Let Prob(C=k)=c_k
  - Services occur singly, backward transitions occur only to adjacent states and have rate μ
  - Forward transitions from state n can occur to any n+k, k>0 state with rate $c_k\lambda$

Analysis of Batch Systems (Con.)



- The global balance equation are given by:

$$\lambda \sum_{k=1}^{n} P(n-k)c_k + \mu P(n+1) = (\lambda + \mu)P(n)$$

where for n=0, λP(0)= μP(1)

Analysis of Batch Systems (Con.)

- To solve the equation we use Z-transform:
  - Let $\Phi_C(z)=\Sigma\, z^n c$ denote the z-transform of the sequence $c_1, c_2, \ldots$
  - Let $\Phi(z)$ denote the z-transform of state probabilities.
  - Multiply both sides of equation by z^n and then sum over n, we get :

$$\rho\,\Phi_C(z)\Phi(z)+\frac{1}{z}\left[\Phi(z)-P(0)\right]=(1+\rho)\left[\Phi(z)-P(0)\right]$$

$$\Phi(z)=\frac{(1-z)P(0)}{1-z-\rho z(1-\Phi_C(z))}$$

Analysis of Batch Systems (Con.)

- We use Utilization law to compute P(0)
  - $1-P(0)=U=\lambda_{total}/\mu$, where $\lambda_{total}=\bar{C}\lambda$
  - $P(0)=(1-\bar{C}\rho)$
  - To ensure ergodicity, $P(0)>0$, $\rho<1/$
- We assume C has a geometric distribution
  - $c_k = (1-a)a^{k-1}$ for k= 1.. ∞. 0<a<1
  - $\Phi_C(z)=z(1-a)/(1-az)$ and $\bar{C} = \Phi'_C(1)=1/(1-a)$

Analysis of Batch Systems (Con.)

- The expression for Φ (z) simplifies to

$$\Phi(z) = \frac{(1-a-\rho)(1-az)}{(1-a)(1-z(a+\rho))}$$

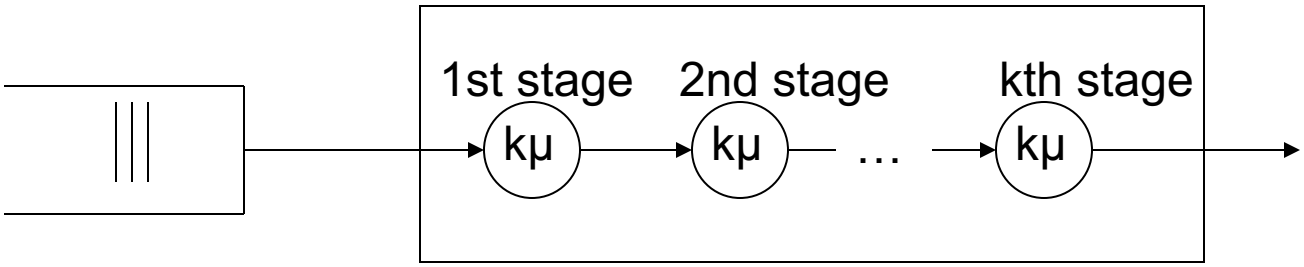
- The average queue length is

$$Q = \frac{\overline{C}}{(1-\overline{C}\rho)}\overline{C}$$

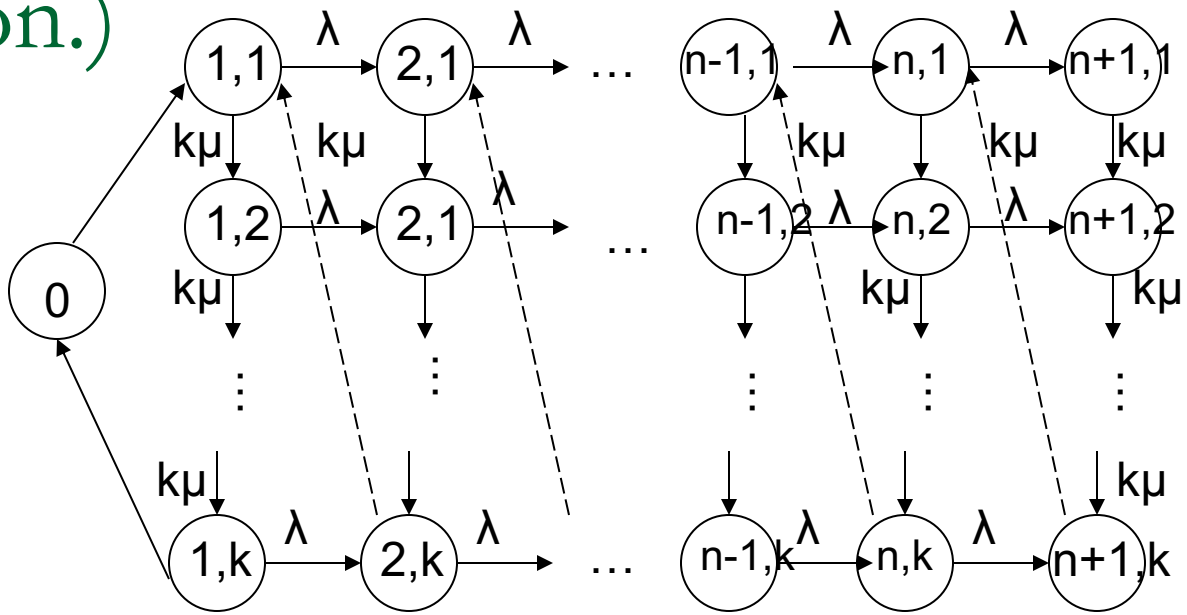
- It can be noted that $Q_{batch}=Q_{expo}\,\overline{C}$

Staged Service with FCFS Scheduling

- The k-stage Erlang system is shown below.
  - At most one customer is allowed to enter the "service box"
  - State system is pair (n,j) where n gives the total number of customers at station and j=1..k gives the total number of customers at the station



Staged Service with FCFS Scheduling (Con.)



- The balance equations are as follows:
  - $(\lambda+\mu)P(n,1) = \lambda P(n-1,1)+k\mu P(n+1,k)$ for $j=1$
  - $(\lambda+\mu)P(n,j) = \lambda P(n-1,j)+k\mu P(n,j-1)$ for $j>1$
  - $(\lambda)P(0) = k\mu P(1,k)$

Staged Service with FCFS Scheduling (Con.)

- Let $\rho=\lambda/(k\mu)$, P(0) can be found by utilization law
  - $U=1-P(0)= \lambda/\mu=k\rho$ ➔ $P(0)=1-k\rho$
- Let P(n) denote the probability of finding n customers in the system :

$$P(n) = \sum_{j=1}^{k} P(n, j)$$

Staged Service with FCFS Scheduling (Con.)

- Use two dimensional z-transform of P(n,j):

$$\Phi^*(y,z) = \sum_{j=1}^{k}\sum_{n=1}^{\infty} y^j z^n P(n,j)$$

- Let $\Phi(z)$ denote the z-transform of P(n):

$$\Phi(z) = \sum_{j=1}^{k} z^n P(n,j) = P(0) + \Phi^*(1,z)$$

- By determining $\Phi^*(1,z)$, we can determine $\Phi(z)$:

$$\Phi(z) = \frac{(1-k\rho)(1-z)}{1-z[1+\rho(1-z)]^k}$$

Staged Service with FCFS Scheduling (Con.)

- Differentiating $\Phi(z)$ and evaluating z=1, we get the average queue length:

$$Q = \frac{k\rho(2-\rho(k-1))}{2(1-k\rho)}$$

  - Note that the queue length is smaller for a comparable M/M queue and decrease monotonically with k.
  - Let $k \to \infty$ with the utilization $U=\lambda/\mu=k\rho$, the Avg. queue length is :

$$Q_{M/D/1} = \frac{U(1-U/2)}{1-U}$$