Project Report on queuing Systems (Summary)

Markov Mavericks

March 3, 2025

Phase-Type Distributions

- **Problem:** Real-world systems often have non-exponential job sizes or interarrival times (e.g., Uniform, Deterministic).
- **Solution:** Use mixtures of Exponential distributions (phases) to model these systems.
  - Allows conversion to CTMC (Markovian structure).
  - Two key tools:
    - **Hypoexponential (Erlang-k):** For low variability ($C^2 < 1$).
    - **Hyperexponential:** For high variability ($C^2 > 1$).

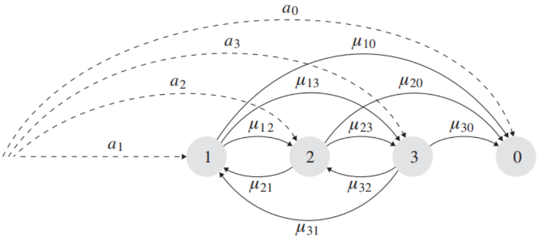


Figure: A 3-phase PH distribution is the time until absorption.

Squared Coefficient of Variation (SCV)

- **Definition:** For a random variable X:

$$C_X^2 = \frac{\mathrm{Var}(X)}{\mathbf{E}[X]^2} = \frac{\mathbf{E}[X^2]}{\mathbf{E}[X]^2} - 1.$$

- **Examples:**
  - Exponential distribution: $C^2 = 1$.
  - Deterministic: $C^2 = 0$.
  - High variability (e.g., web requests): $C^2 \gg 1$.

Erlang and Hypoexponential Distributions

- **Erlang-k Distribution:**
  - Modeled as the sum of k i.i.d. exponential stages: $T = T_1 + T_2 + \cdots + T_k$.
  - Each stage: $T_i \sim \text{Exp}(k\mu)$ so that $\mathbf{E}[T] = 1/\mu$.
  - Variance: $\text{Var}(T) = \frac{1}{k\mu^2}$, yielding $C_T^2 = 1/k$.
  - As $k \to \infty$, $C_T^2 \to 0$ (approaching a deterministic distribution).
- **Hypoexponential Distribution:** A generalization where the exponential stages have different rates.

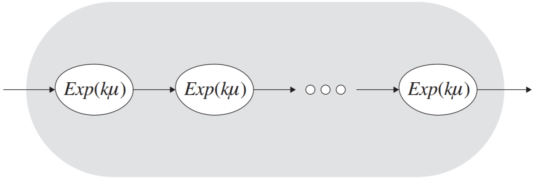


Figure: Illustration of Erlang-k distributions.

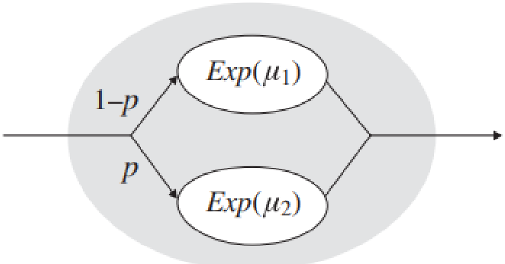
Erlang Distribution... Continued

- The Erlang distribution is a special case of the Gamma distribution with an integer shape parameter r.
- It is used to model service times in queuing systems with multiple stages of service.
- In an Erlang process, the service time is divided into r **sequential exponential phases**, each with a mean of $\frac{1}{\lambda}$.
- Its probability density function is

$$f(x; r, \lambda) = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{(r-1)!}, \quad x \geq 0.$$

Hyperexponential Distributions

- **High Variability Modeling:** Suitable for distributions with $C^2 > 1$.
- **Structure:** Rather than sequential stages, the process takes one of several exponential "paths" immediately.
- **Example:**
  - With probability p: $T \sim \text{Exp}(\mu_1)$.
  - With probability $1 - p$: $T \sim \text{Exp}(\mu_2)$.
- **Interpretation:** Represents systems (e.g., web request times) with a mix of fast and slow responses.



Definition of k-phase PH Distributions

- **Setup:** Consider a continuous-time Markov chain (CTMC) with $k+1$ states.
- **States:**
  - States $1, \ldots, k$: *Transient (phases)*.
  - State 0: *Absorbing*.
- **Parameters:**
  - $\mathbf{a} = (a_0, a_1, \ldots, a_k)$: Initial probability vector, with $\sum_{i=0}^{k} a_i = 1$.
  - T: A $k \times (k+1)$ rate transition matrix; entry $T_{ij} = \mu_{ij}$ is the rate from state i to j (for $i \neq j$).
- **Interpretation:** The PH distribution is the distribution of time until absorption (i.e., until the process reaches state 0).

Coxian Distributions

- **Definition:** A special subclass of PH distributions with a sequential structure.
- **Structure:** Similar to an Erlang-k but allows for *early absorption*:
  - At each stage i, there is a probability b_i of exiting to the absorbing state.
  - The remaining probability a_i (with $a_i + b_i = 1$) continues to the next phase.

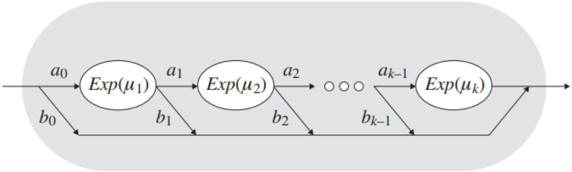


Figure: Illustration of a Coxian distribution.

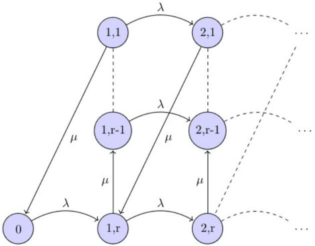
CTMC Modeling (State Description)

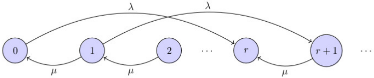
- **Option 1:** (k, l)
  - k: Number of customers in the system.
  - l: Remaining number of service phases of the customer currently being served
- **Option 2:** Total number of uncompleted phases of work in the system:

$$n = (k-1)r + l.$$

  We will follow this one.
- Define p_n as the steady-state probability of having n uncompleted phases.
- Obtained by solving the balance (equilibrium) equations.

State Diagrams: Option 1 and Option 2





Option 2: 1D State Diagram

Option 1: 2D State Diagram

Stationary Distribution

- Let us try to balance on each state for this...

$$\lambda\pi_0 = \mu\pi_1$$

$$\lambda\pi_n + \mu\pi_n \; = \; \mu\pi_{n+1}, \quad n = 1, 2, \ldots, r-1$$
$$\lambda\pi_n + \mu\pi_n \; = \; \lambda\pi_{n-r} + \mu\pi_{n+1}, \quad n = r, r+1, \ldots$$

- There cannot be a negative number of customers in the system, so by convention $\pi_n = 0$ for $n < 0$. Using this, we can combine the above equations to obtain:

$$\lambda\pi_n + \mu\pi_n \; = \; \lambda\pi_{n-r} + \mu\pi_{n+1}, \quad n \geq 1.$$

- Now we use the following method of solving linear equations in which we assume that:

$$\pi_n = x^n \;\; \forall \; n \in \{0, 1, \ldots\}$$

Stationary Distribution... continued

- Substituting and simplifying we get the follow r + 1 degree polynomial in x:

$$\mu x^{r+1} - (\lambda + \mu)\, x^r + \lambda \;=\; 0$$

- or equivalently, we have :

$$(\lambda + \mu) - \mu x \;=\; \frac{\lambda}{x^r}$$

- Next we proceed to prove the uniqueness of each of the r + 1 roots.

Proof:

- First it is trivial to note that x = 1 is a solution while x = 0 is not a solution. (Also, we will be using the proven fact that all the roots of the equation are s.t. $\mid x \mid < \ 1$.
- Now let us assume $\frac{1}{x}$ to be a root of the equation:

$$\mu \left(\frac{1}{x}\right)^{\ r+1} - (\lambda + \mu) \left(\frac{1}{x}\right)^{\ r} \ + \lambda = 0 \iff 1 - \frac{\lambda x}{\mu} \frac{1 - x^r}{1 - x} = 0$$

- f(x) defined below must be having r roots.

$$Say, f(x) = 1 - \frac{\lambda x}{\mu} \frac{1 - x^r}{1 - x} = 1 - \frac{\rho}{r} \ (x + x^2 + x^3 + \dots)$$

- For the distinctness of the roots, it must be true that the derivative of the function is not zero at any of the roots.

$$f'(x) = -\frac{\rho}{r} \ (1 + 2x + 3x^2 + \dots + rx^{r-1})$$

Proof:

- For the f'(x) to be zero, we need the following term to be zero:

$$1 + 2x + 3x^2 + \cdots + rx^{r-1} = 0$$

- Multiply (1 - x) on both sides, as x = 1 is not a root of the equation. We get:

$$1 + x + x^2 + x^3 + \cdots + x^{r-1} - rx^r = 0$$

- Now, from triangle inequalities (as $\mid x \mid > 1)$:

$$\mid 1 + x + x^2 + x^3 + \cdots + x^{r-1} \mid <= \ r \mid x^r \mid$$

- This shows a contradiction and hence f'(x) cannot be zero at any root. Hence, there exist unique roots.

Stationary Distribution... continued

- The generating function of the stationary distribution is defined as:

$$f(z) \;=\; \sum_{n=0}^{\infty} \pi_n z^n \;,\;\; |\; z \;| < 1$$

- By multiplying the balance equation by z^n and summing over all n $>=$ 1, we get the following:

$$(\lambda + \mu) \sum_{n=1}^{\infty} \pi_n z^n = \lambda \sum_{n=1}^{\infty} \pi_{n-r} z^n + \mu \sum_{n=1}^{\infty} \pi_{n+1} z^n .$$

$$(\lambda+\mu)\left(\sum_{n=0}^{\infty} \pi_n z^n - \pi_0\right) = \lambda z^r \sum_{n=0}^{\infty} \pi_n z^n + \mu z^{-1}\left(\sum_{n=0}^{\infty} \pi_n z^n - \pi_1 z - \pi_0\right)$$

Stationary Distribution... continued

- Substituting f(z) from above:

$$(\lambda + \mu)(f(z) - \pi_0) = \lambda z^r f(z) + \mu z^{-1}(f(z) - \pi_1 z - \pi_0),$$

- Now solving further for f(z):

$$f(z) = \frac{-z(\lambda+\mu)\pi_0 + z\lambda\pi_0 + \mu\pi_0}{-z\lambda - z\mu + \lambda z^{r+1} + \mu}.$$

- Finally, we get:

$$f(z) = \frac{1-\rho}{1-\rho\left(\frac{z+z^2+\cdots+z^r}{r}\right)}$$

Stationary Distribution... continued

- The denominator of $f(z)$ has r distinct roots z_i with $|z_i| > 1$. Note, that each root z_i corresponds to a root $\frac{1}{z_i} = x_i$. We can thus write $f(z)$ as

$$f(z) = \frac{1-\rho}{\left(1 - \frac{z}{z_1}\right) \cdots \left(1 - \frac{z}{z_r}\right)}.$$

- Using partial fraction decomposition this can be written as

$$f(z) = \frac{1-\rho}{\left(1 - \frac{z}{z_1}\right) \cdots \left(1 - \frac{z}{z_r}\right)} = (1-\rho)\left(\frac{A_1}{1 - \frac{z}{z_1}} + \cdots + \frac{A_r}{1 - \frac{z}{z_r}}\right),$$

- with

$$A_i = \left(\prod_{j \neq i} \left(1 - \frac{z_i}{z_j}\right)\right)^{-1},$$

for $i = 1, 2, \ldots, r$. Note that A_i is well-defined, since all z_i are distinct.

Stationary Distribution... continued

Furthermore, we can make use of the fact that

$$\sum_{n=1}^{\infty} z^n = \frac{1}{1-z},$$

for $z \in (0,1)$. This allows us to rewrite the last part of Eqn. as:

$$f(z) = (1-\rho) \sum_{n=0}^{\infty} \left(\sum_{i=1}^{r} A_i \cdot \left(\frac{1}{z_i} \right)^n \right) z^n.$$

From this and earlier equations, it follows that the stationary distribution of the Erlang queue is

$$\pi_n = (1-\rho) \sum_{i=1}^{r} A_i \left(\frac{1}{z_i} \right)^n,$$

for $n > 0$. As $\frac{1}{z_i} = x_i$, it follows that

$$\pi_n = (1-\rho) \sum_{i=1}^{r} A_i x_i^n.$$

Stability via Embedded Chain & State Update

- **Embedded Chain:** Now we will look at the embedded DTMC of our Markov chain. We sample the system at departure epochs.
- Let X_n denote the number of customers immediately after the nth departure.
- **Transition Mechanism:**
  - One customer departs (if the system is non-empty).
  - During the service of the departing customer, a random number A_n of customers arrive.
- **State Update Equation:**

$$X_{n+1} = \max\{X_n - 1, 0\} + A_n.$$

Derivation of $P(A_n = n)$ and $\mathbb{E}[A_n]$

- **Starting Point:** Condition on service time s:

$$P(A_n = n \mid S = s) = e^{-\lambda s} \frac{(\lambda s)^n}{n!}.$$

- **Unconditioning:** Integrate over s using the Erlang PDF:

$$f_S(s) = \frac{\mu^r \, s^{r-1} \, e^{-\mu s}}{(r-1)!}, \quad s \geq 0.$$

- **Final Result:** Recognizing the Gamma integral yields a Negative Binomial form:

$$P(A_n = n) = \binom{n+r-1}{n} \left(\frac{\mu}{\lambda+\mu}\right)^r \left(\frac{\lambda}{\lambda+\mu}\right)^n.$$

- **Mean of A_n:** Using generating function techniques, we obtain

$$\mathbb{E}[A_n] = \frac{r\lambda}{\mu}.$$

Lyapunov Function and Drift Definition

- **Lyapunov Function:** A Lyapunov function measures the "size" or "energy" of the system. In our analysis, we choose

$$V(x) = x,$$

where x denotes the number of customers in the system.

- **Drift Definition:** The drift is defined as the expected change in the Lyapunov function in one step:

$$\mathbb{E}\Big[V(X_{n+1}) - V(X_n) \mid X_n = x\Big]$$

Foster's Criterion and Application

- **Foster's Criterion:** A Markov chain is positive recurrent if there exists a function $V(x)$ and a finite set $\mathcal{C}$ such that for all $x \notin \mathcal{C}$,
$$\mathbb{E}[V(X_{n+1}) - V(X_n) \mid X_n = x] < 0.$$
- **Application:** Using $V(x) = x$, for $x \geq 1$ the drift is given by
$$\mathbb{E}[V(X_{n+1}) - V(X_n) \mid X_n = x] = -1 + \frac{\lambda r}{\mu}$$
This expression is less than 0 when $\frac{\lambda r}{\mu} < 1$.

Explanation and Stability Condition

- This drift expression arises because each departure reduces the customer count by 1, while during the service period an average of $\frac{\lambda r}{\mu}$ customers arrive.
- Thus, when $\frac{\lambda r}{\mu} < 1$, the overall expected change is negative for $x \geq 1$, ensuring the system tends to return to a smaller state.
- In our analysis, we take the finite set $\mathcal{C} = \{0\}$.
- **So the Stability Condition is $\frac{\lambda r}{\mu} < 1$.**

Components of Waiting Time

- **Waiting Behind Other Customers:**
  - Let $E(L_q)$ denote the average number of customers waiting.
  - Each customer requires r phases, with an average service time of $\frac{1}{\mu}$ per phase.
  - Thus, the waiting time due to customers ahead is:

$$\frac{r}{\mu} E(L_q).$$

- **Waiting Due to Residual Service Time:**
  - When the server is busy (which is by probability ρ), an arriving customer must wait for the remaining service time.
  - For an Erlang-r service process, the expected residual service time is given by

$$E(R) = \frac{1}{r} \sum_{k=1}^{r} \frac{k}{\mu} = \frac{r+1}{2\mu}.$$

Derivation of Mean Waiting Time

- **Total Waiting Time:** The overall waiting time is the sum of the waiting due to the customers ahead and the residual service time (incurred when the server is busy), hence:

$$E(W) = \frac{r}{\mu} E(L_q) + \rho\, E(R),$$

where $\rho = \frac{\lambda r}{\mu}$ is the server utilization.

- **Using Little's Law:** Since $E(L_q) = \lambda E(W)$, we substitute to get:

$$E(W) = \frac{r}{\mu} \lambda E(W) + \rho\, \frac{r+1}{2\mu}.$$

- **Final Expression:** Recognizing that $\frac{r}{\mu}\lambda = \rho$ and solving for $E(W)$ yields:

$$E(W) = \frac{\rho}{1-\rho}\, \frac{r+1}{2\mu}.$$

Mean Sojourn Time ($E[T]$)

- The mean sojourn time is the total time a customer spends in the system.
- It is the sum of the mean waiting time $E[W]$ and the mean service time $E[S]$.
- For an M/Er/1 queue, the mean service time is $\frac{r}{\mu}$.
- Thus, we have:

$$E[T] = E[W] + E[S] = \frac{\rho}{1-\rho}\frac{r+1}{2\mu} + \frac{r}{\mu},$$

where $\rho = \frac{\lambda r}{\mu}$ is the system utilization.

Mean Number of Customers ($E[N]$)

- According to Little's Law for the entire system:

  $$E[N] = \lambda\, E[T],$$

  where $E[T]$ is the mean sojourn time.
- Substituting the expression for $E[T]$ gives:

  $$E[N] = \lambda \left(\frac{\rho}{1-\rho} \, \frac{r+1}{2\mu} + \frac{r}{\mu} \right).$$

- Since $\rho = \frac{\lambda r}{\mu}$, we can express $\lambda = \frac{\rho\, \mu}{r}$. Substituting this, we obtain:

  $$E[N] = \frac{\rho\,\mu}{r} \left(\frac{\rho}{1-\rho} \, \frac{r+1}{2\mu} + \frac{r}{\mu} \right) = \frac{\rho^2 (r+1)}{2r(1-\rho)} + \rho.$$

Blocking Probability

- In an M/Er/1 queue with infinite buffer capacity, every arriving customer is eventually served.
- Thus, the blocking probability is:

$$P_{\text{block}} = 0.$$

- (Note: For finite-capacity models, such as M/Er/1/K, the blocking probability would be non-zero.)

Waiting Time Distribution

- The waiting time W in an $M/Er/1$ queue depends on the number of busy servers upon arrival. If the server is busy, the customer must wait until it becomes available.
- The system can be modeled as having r exponential service phases, where the arrival process interacts with the system state to create a mixture of exponentials.
- By and leveraging the balance equations and conditional probabilities, we obtain:

$$P(W > t) = \sum_{k=1}^{r} c_k \frac{x_k}{1 - x_k} e^{-\mu(1-x_k)t}, \quad t \geq 0.$$

Sojourn Time Distribution

- The sojourn time $T = W + S$ represents the total time a customer spends in the system, where S follows an Erlang-r distribution since it is the sum of r independent exponential service phases.
- To obtain the distribution of T, we integrate the waiting time distribution over the service time distribution. This involves computing the convolution of the waiting time density with the Erlang-r density.
- Applying this convolution, we solve for $P(T > t)$ by expanding the transformed expression using properties of exponential mixtures and Erlang distributions. After simplifications, we obtain:

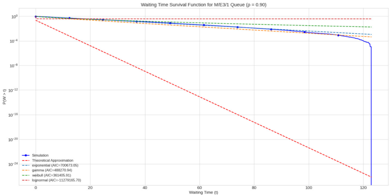
$$P(T > t) = \sum_{k=1}^{r} c_k \frac{(r-1)!(x_k)}{(1-x_k)(\mu x_k)^r} e^{-\mu(1-x_k)t}.$$

Simulation & Validation

- **Simulation Setup:**
  - Generate Poisson arrivals at rate λ.
  - Model service as r exponential phases (rate μ).
- **Validation:** Compare simulated averages of $E[W]$, $E(T)$, and $E(N)$ with theoretical results.

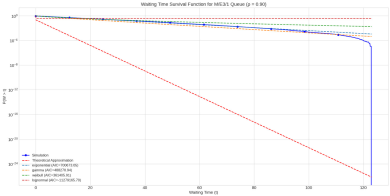
Empirical Survival Function

The empirical survival function $P(W > t)$ is computed from simulated data of the $\mathbf{M/E_3/1}$ queue with utilization $\rho = 0.90$. It represents the probability that a randomly chosen job experiences a waiting time greater than t.



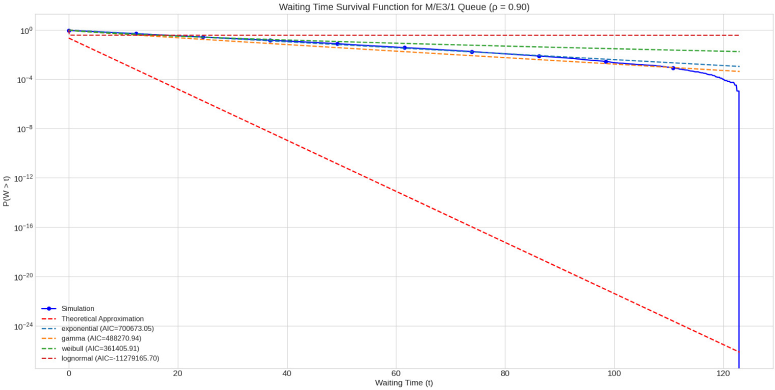
Theoretical Approximation

A theoretical approximation (red dashed line) estimates the tail behavior of the waiting time distribution. This curve serves as a reference to assess how closely fitted distributions approximate the simulated data.



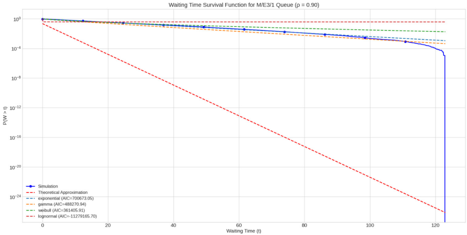
Fitted Distributions

- **Exponential**: Assumes a memoryless property but does not fit well.
- **Gamma**: More flexible and provides a better approximation.
- **Weibull**: Captures queuing system characteristics effectively.
- **Lognormal**: Provides the **best fit** based on AIC scores.



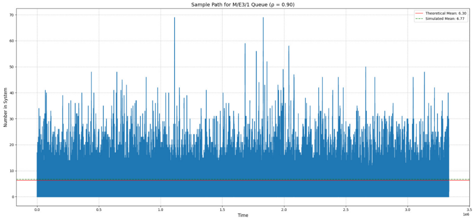
Observations and Conclusions

- The exponential distribution underestimates waiting time probabilities.
- The lognormal distribution provides the closest fit.
- The gamma and Weibull distributions also approximate well.
- The theoretical approximation slightly deviates in the tail region.



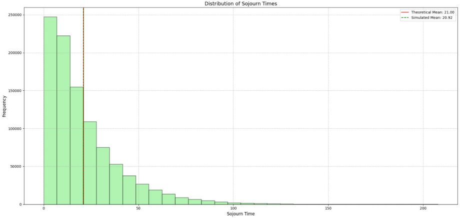
Sample Path Simulation

Sample paths show job arrivals, service start times, and departures. The number of jobs in the system over time is recorded to analyze system behavior.



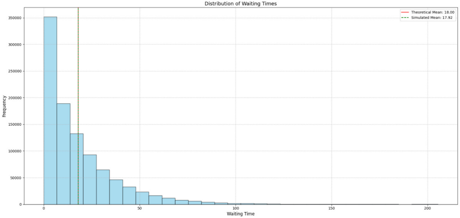
Performance Metric Validation

- **Mean Sojourn Time:** Estimated from simulation and compared to theoretical values.
- **Mean Number of Jobs:** Verified using Little's Law.



Waiting Time Distribution

The waiting time distribution provides insights into queuing behavior. Histograms show the simulated waiting times and their alignment with theoretical expectations.



Conclusion

- **Modeling Success:** Erlang and Phase-Type distributions provide a robust framework for queuing systems.
- **Key Results:** Tractable performance metrics and stability conditions for the M/Er/1 queue.
- **Future Work:** Extend to multi-server, finite-buffer, and priority systems.

References

- I. Adan and J. Resing, *queuing Systems*, 2015.
- Chapter 21 of *Performance Modeling*, various authors.
- Additional literature on Erlang and Phase-Type distributions.