Stochastic Models: Evaluation Exercise

Lecture 5 – M/G/2 Case Study

Sebastian Müller
Lecture 5

Goal of the Evaluation

- Work from a raw event log to a plausible stochastic queueing model.
- Infer arrival and service parameters from data, including uncertainty.
- Assess performance (waiting times, mean number in queue L_q, utilisation) of a two-server system.
- Practice telling a clear modelling story from noisy real-world data.

System Description

- Small ML-backed support platform handling “complex” tickets.
- Two identical human agents process tickets in parallel (two servers).
- All incoming tickets during the observation window enter this two-server system and are recorded in the log.
- Each complex ticket incurs a fixed overhead (reading context, loading tools) plus a random processing time.

Note

For modelling we treat this as an $M/G/2$ queue: Poisson arrivals (unknown rate), i.i.d. service times with unknown distribution G, and two identical servers.

What Data You Get

- One CSV log from a contiguous observation window (no gaps).
- One row per completed ticket, with fields:
  - arrival_time, start_service_time, completion_time
  - service_time, wait_time, system_time
  - queue_len_at_arrival (tickets in system just before arrival)
- Jobs that arrive before the end of the window are included even if they complete later.

Modelling Tasks

- Clean and validate the log (nonnegative waits, temporal ordering).
- Diagnose the arrival process: inter-arrival distribution, approximate stationarity, estimate $\hat{\lambda}$ with a CI.
- Explore the service-time distribution: evidence of a lower bound (fixed overhead) and an approximately memoryless tail.
- Propose a simple parametric family for G (e.g. constant offset + exponential) and fit its parameters.

Performance and Uncertainty

- Use your fitted model to estimate utilisation $\hat{\rho} = \hat{\lambda}\hat{m}_1/2$ for the two-server system.
- Estimate the mean number of waiting jobs L_q from the data and relate it to $\hat{\lambda}$ and $\overline{W}_q$ (Little's Law).
- Compare empirical mean waiting time from the log to a model-based prediction (via simulation or an $M/G/2$ approximation).
- Quantify uncertainty for at least one key metric (e.g. bootstrap or regenerative analysis over cycles).

Note

You may reuse any error-control or bootstrap techniques from earlier lectures (delta method, input bootstrap, regenerative bootstrap, simulation-based CIs, . . .).

Deliverables and Grading

- A clear description of your chosen model (arrival process, service family, number of servers).
- Parameter estimates with at least one uncertainty measure (CI or standard error).
- Comparison between empirical and model-based performance (focus on waiting time, mean number in queue L_q, and utilisation).
- Short discussion of diagnostics and model limitations (what the model misses, robustness of conclusions).

Note

Emphasis is on *reasoned modelling and diagnostics*, not on guessing the exact hidden parameters.