ORIE 4580/5580: Simulation Modeling and Analysis

ORIE 5581: Monte Carlo Simulation

Unit 14: Ranking, Selection, and Optimization

Sid Banerjee

School of ORIE, Cornell University

comparison of alternate systems

till now we have seen how to:

– simulate complex discrete-event systems

– compute statistics about these models.

we now want to use these to compare different system configs

comparing systems: main ideas

- use of simultaneous confidence intervals
- practical significance and indifference zones
- use of common random numbers

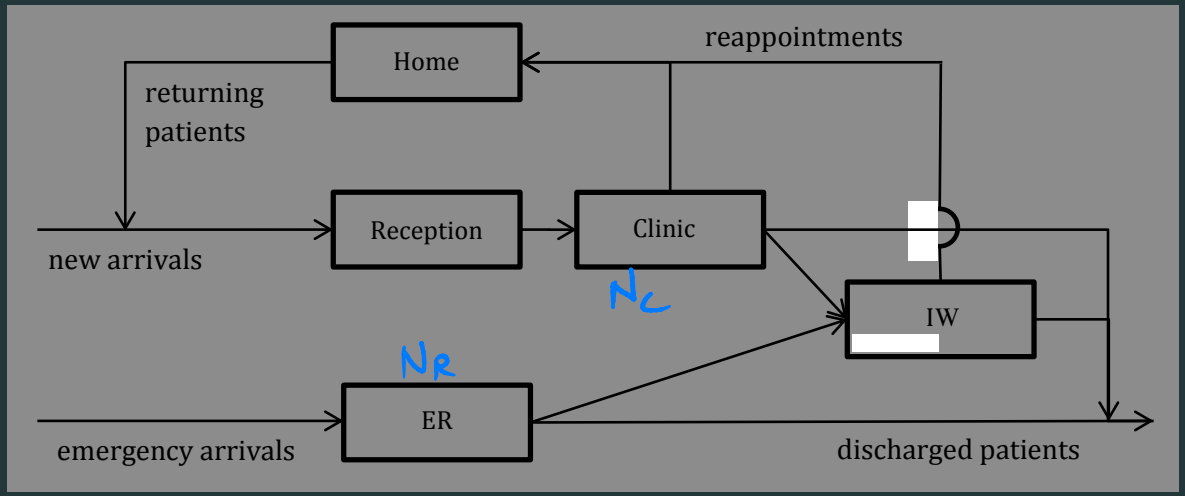
'statistics'

(statistical significance) 'philosophy'

'Simulation'!

Problem - all our data is random

example: staffing the Fingerlakes hospital



hospital employs 15 doctors $N_R + N_C = 15$

Q: how should we allocate doctors to optimize service?

questions and models

Sim opt / ranking + selection

- how do we divide the doctors between ER and clinic?
- what is the added benefit of hiring another doctor?
- is it useful to have 'floaters' who can go to the clinic/ER as needed?

systems - (N_R, N_C)

$(6,9)$ vs $(7,8)$ vs $\boxed{(8,7)}$ vs $(9,6)$ ...

how much improvement

$(6,10)$ vs $(7,9)$ vs $(8,8)$ vs $\boxed{(9,7)}$ ...

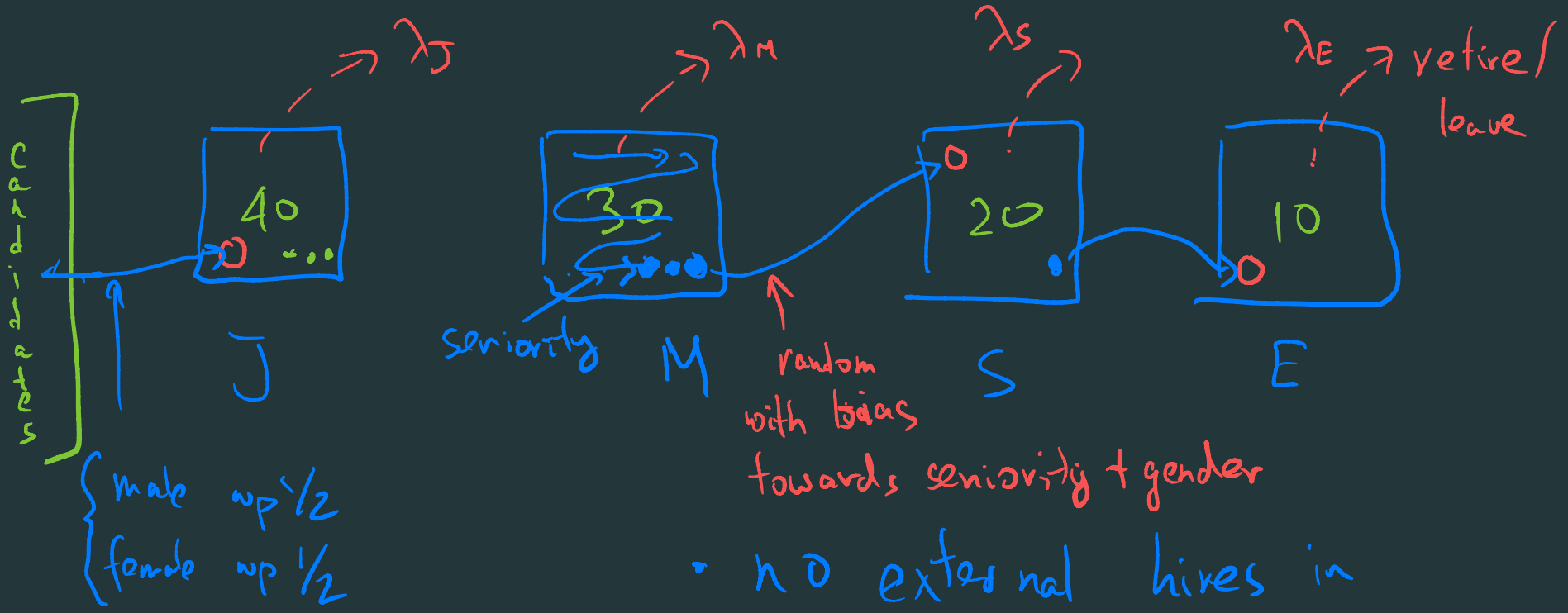
Control / Markov decision process

$15 = N_R + N_C + N_F$

$6 \quad 6 \quad \boxed{3}$ ← can move around based on instantaneous demand

systems ≡ policies

example: combating lack of diversity in companies



{ male up 1/2
female up 1/2

- no external hires in 'senior' posns

- adapted from Schelling segregation

example: combating lack of diversity in companies

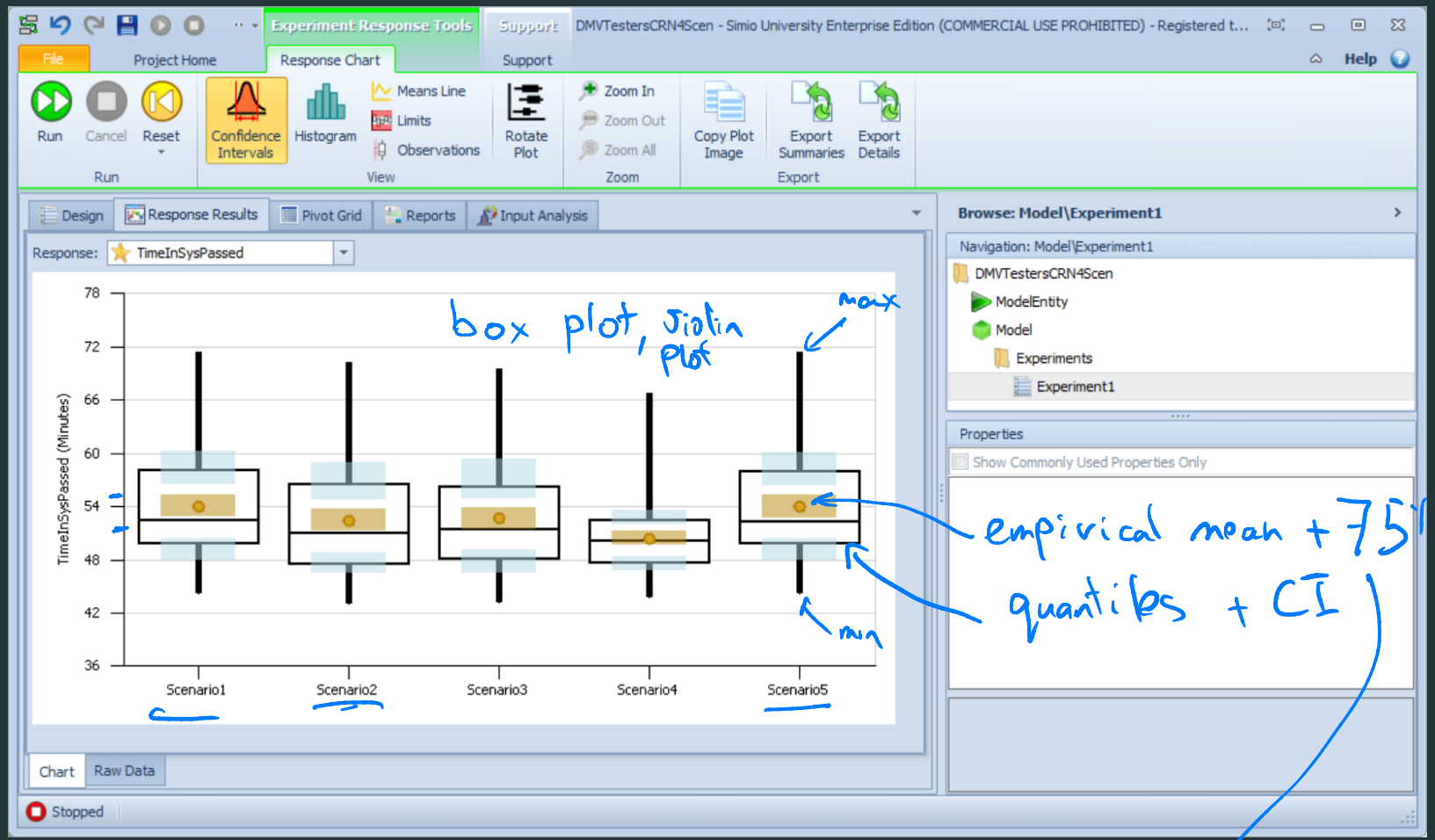
'Rooney rule': for every position, interview top male and female candidate

Status quo (interview top 3) vs 'Rooney rule' (interview top male and top female candidate)

- what is the right objective?
- what do numbers mean?

simultaneous confidence intervals

how do we use CIs to compare different scenarios?



does this mean that with prob 0.75, scenario 4 is the best?

simultaneous confidence intervals: the union bound

the union bound

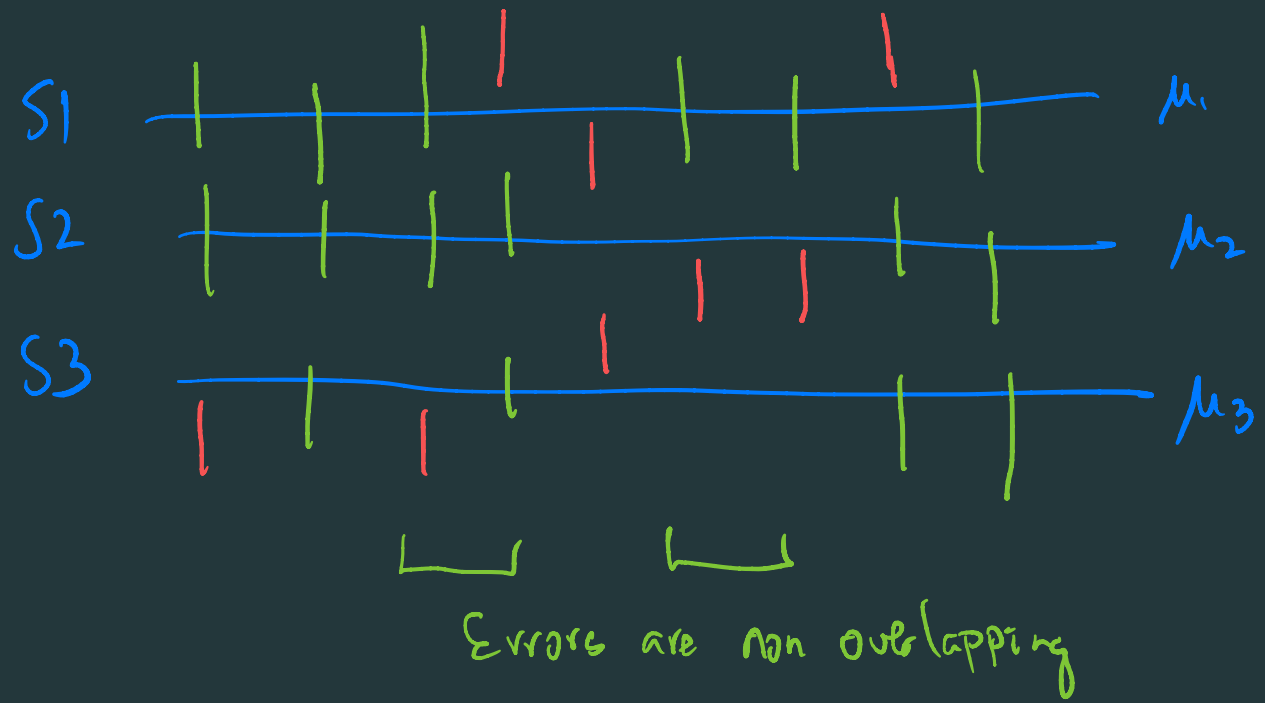
let $A_1, A_2, \ldots, A_k$ be events. then

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_k) = 1 - \mathbb{P}(A_1^c \cup A_2^c \cup \cdots \cup A_k^c)$$

$$\geq 1 - (\mathbb{P}(A_1^c) + \mathbb{P}(A_2^c) + \cdots + \mathbb{P}(A_k^c))$$

All A_i hold simultaneously

let A_i = event that the ith ci contains its true mean...



Eg - 75% CI

- Can compare systems only when ALL CIs are 'correct'

practical and statistical significance

a practically meaningful difference depends on the problem at hand:

- $10,000 on a portfolios return
- 5 minutes in waiting time for COVID test
- 20 people being unable to connect to a Zoom meeting

statistical significance depends on sampling variability in estimates:

- a 95% confidence interval for the difference in expected time between scenarios is 4 ± 7 minutes. what can we conclude?
- what if it was 4 ± 1 minute?

controlling significance

- we use **statistical procedures** to tell us whether we can believe the difference we see in the results from two or more scenarios.
- we use the **number of replications** to control the size of the difference that is detectable; that is, to control the error in our estimates.
- **you** have to decide what difference is practically significant.

ranking and selection

- given a set of systems, simulates each for a random amount of time and returns a single system i that is estimated to be the best
- to keep this from running for ages in the event of ties or near ties, we specify an indifference zone δ – smallest difference worth detecting (practical significance)
- *"with probability ≥ 0.95, system i is the best, or is within δ of the best"*
- be careful! run time increases as $\delta \to 0$.

CS

statistical significance

practical significance

• PAC guarantee - Probably Approximately Correct

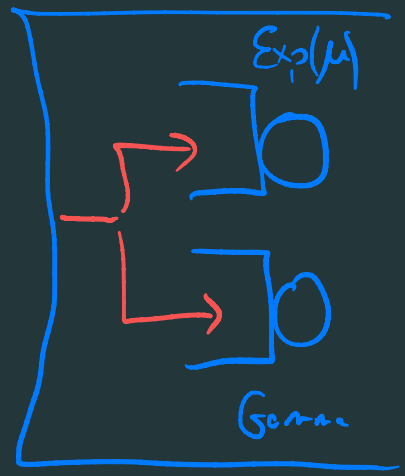
is comparing different simulations fair?

Eg - System 1 - M/M/1, arrival rate λ, service rate μ

M/GI/1, arrival rate λ, service time ~ Gamma with mean μ

↑ general service

Want them to be same



- Replication 1 - $\bar{W}_1(M/M/1) = 101$

$\bar{W}_1(M/GI/1) = 99$

$\Rightarrow \mathbb{E}[D] = \mathbb{E}\left[\bar{W}_1(M/M/1) - \bar{W}_1(M/GI/1)\right] = 2$

$Var(D) = ?$ (can not happen if same arrivals)

Eg - 2nd sim had more arrivals than first

common random numbers

Want to check if $D = X - Y \gtrless 0$?

Refined version: $X - Y > \delta$, $X - Y < -\delta$, $|X - Y| < \delta$

- let X and Y be rvs giving output from two different scenarios.
- want $\mu_X - \mu_Y = \mathbb{E}[X] - \mathbb{E}[Y] = \mathbb{E}[X - Y] = \mathbb{E}[D]$

$\delta \equiv$ significance level

- if X, Y independent (different streams) then

  $Var(D) = Var(X - Y) = Var(X) + Var(Y)$

Alternate: $Z = \mathbb{1}_{\{D > 0\}}$ if $Z \sim Ber(1/2)$ then $Var(Z) = \frac{1}{4}$

- in general (whether independent or not)

  $Var(D) = Var[X - Y] = Var(X) + Var(Y) - 2Cov(X, Y)$

- use CRN to try to make $Cov(X, Y) > 0$.

opposite of variance reduction

$\bar{X} = \frac{X+Y}{2}$, $Var(\bar{X}) = \frac{1}{4}\left(Var(X) + Var(Y) + 2Cov(X, Y)\right)$

If negative $\Rightarrow$ var reduction

clicker question: comparing queueing disciplines

Can be GI/GI/s

consider an M/M/1 queue with arrival rate λ and service rate $\mu > \lambda$

suppose you build two simulation models

- in the first, you serve jobs in a First-In First-Out (FIFO) order
- in the second in a Last-In First-Out (LIFO) order

15 (a) the average queue length in FIFO is smaller than that in LIFO

15 (b) the average queue length in LIFO is smaller than that in FIFO

45 (c) the averages are same, but LIFO has higher variance in queue lengths

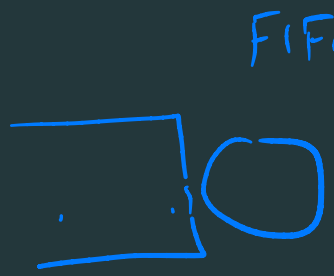
25 (d) ✓ the queue length distributions are identical in the two

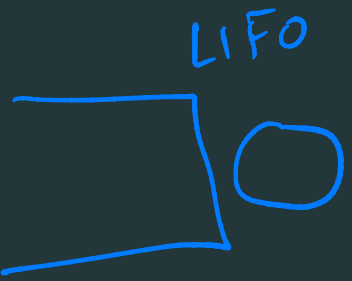
clicker question: comparing queueing disciplines

consider an M/M/1 queue with arrival rate λ and service rate $\mu > \lambda$

suppose you build two simulation models

- in the first, you serve jobs in a First-In First-Out (FIFO) order
- in the second in a Last-In First-Out (LIFO) order





Idea - make arrival ad service times the 'same' in the 2

clicker question: comparing queueing disciplines part 2

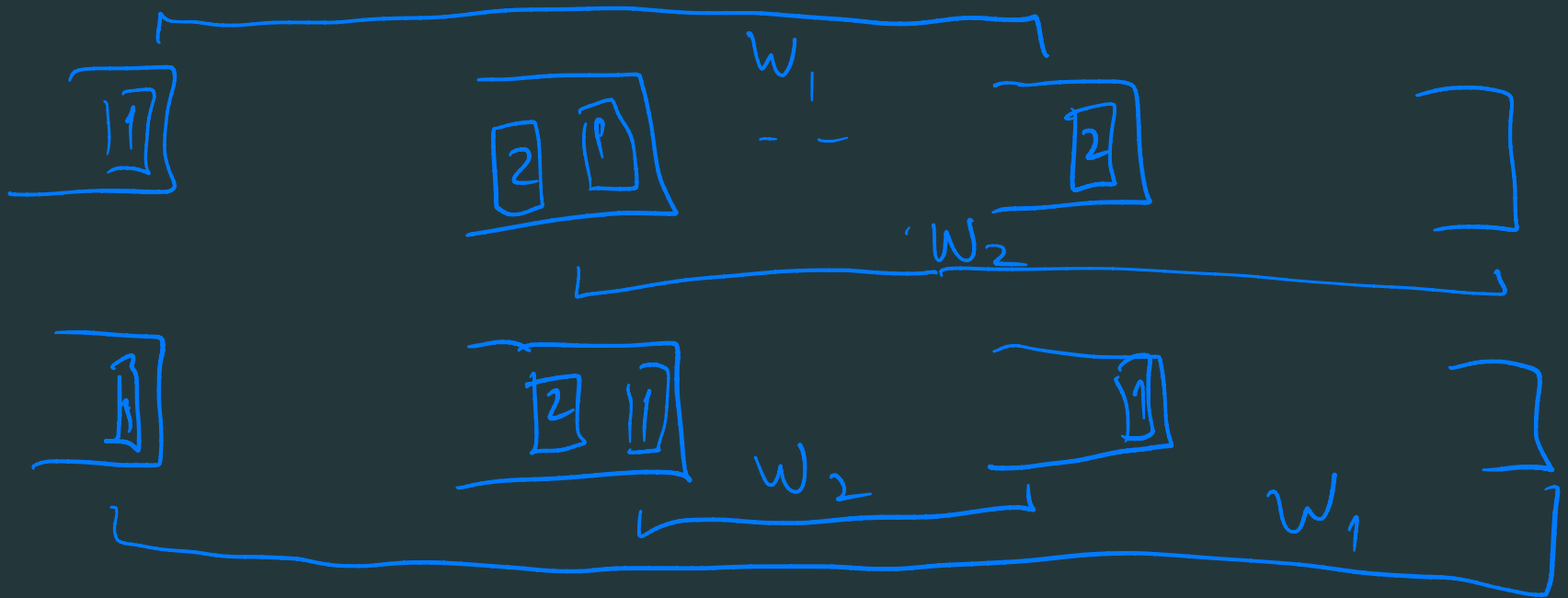
in the previous setting (M/M/1 queue with arrival rate λ, service rate $\mu > \lambda$ under FIFO and LIFO service), what can we say about the time in system?

17 (a) the average time in system in FIFO is smaller than in LIFO

17 (b) the average time in system in LIFO is smaller than in FIFO

56 (c) the averages are same, but LIFO has higher variance in time in system

10 (d) the time in system distributions are identical in the two



RNG streams

(Most useful in general discrete event simulation)

- original model used a single stream (stream 0) for everything
  – scrambles the sequence
- fix using streams

streams in python

- create different rng objects for each stream
- eg. in numpy:

```
arrival_stream = np.random.randomstate(seed=0)
service_stream = np.random.randomstate(seed=1)
t = arrival_stream.exponential(1.0/arrival_rate)
```

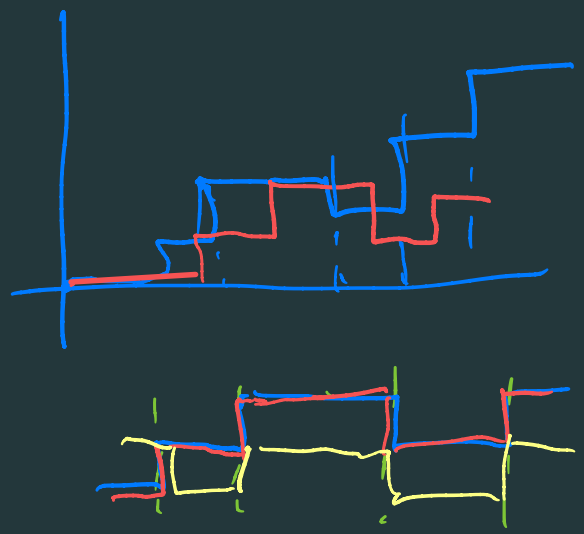
different PRNG objects

≈ independent

np.random.rand( )

using CRN in Markovian simulations

Eg - diversity sim



Difficult to avg replicates because times are different

Idea - Generate the retirement times and identities using CRN

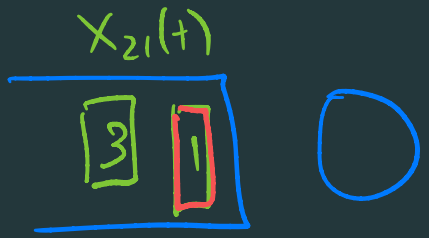
using CRN in Markovian simulations

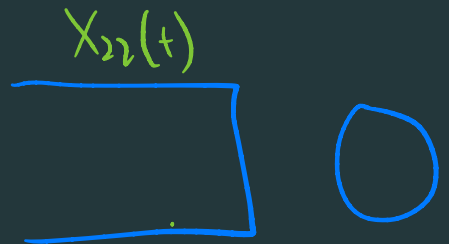
Eg - ① M/M/2 queue, arrival rate λ (common line for both servers)

② 2 M/M/1 queues, arrivals join shortest queue

(All servers work at rate μ)

$X_2(t) = X_{21}(t) + X_{22}(t)$





- If one of the queues is starved, service rate $= \mu$
- Common arrivals to both at rate λ
- Common 'service arrivals' at rate 2μ if $X_1(t) \geq 2$, $X_{21}(t)$ AND $X_{22}(t) \geq 1$
  - when service arrives - remove random agent at HOL

using CRN in Markovian simulations (coupling)

- common Arrivals to both queues at rate λ
- Common 'service arrivals' to both at rate 2μ. At this time
  - In system 1, if $X_1(t) \geq 2$, serve customer
    if $X_1(t) = 1$, serve customer w.p. $\frac{1}{2}$
  - In system 2, pick random server, and serve if customer is present

(working through all cases)

$$\Rightarrow X_1(t) \leq X_{21}(t) + X_{22}(t)$$

↑ sample-path stochastic dominance

using CRN in Markovian simulations

simulation optimization

ranking and selection:

– comparing small number of systems

– need to simulate each system at least a bit

what can we do for bigger problems?

simulation optimization

- simulation optimization: **search** over different systems
- Markov decision processes: **optimize** over decisions (controls)

simulation optimization is hard

- local vs global optima
- many decision variables means huge decision space. e.g., shifts start on the hour, up to 11 agents can start each hour, 11^{24} possible solutions (systems)
- **estimation error** means we can never be certain that one solution x is better than another y.
- **simulation noise** (estimation error) can swamp the signal.

optimization bias

the estimated objective value for a minimization problem is always lower than it should be

tools and techniques

sample average approximation use a fixed run-length and common random numbers. minimize estimated function with deterministic optimization software

metamodeling fit a simpler function, e.g., polynomial, to simulation output, minimize the simpler function

stochastic approximation somehow estimate the slope (gradient) of the objective function at current point, and take a step in the opposite direction; repeat

random search given a current point or set of points, randomly choose new ones and simulate

- easily adapted to broad classes of problems
- no guarantees
- not much good with lots of variables