Markov chain Monte Carlo and Perfect Simulation

Lecture at Aristotle University of Thessaloniki

Wilfrid S Kendall

University of Warwick

15 May 2024

Warwick Statistics

Introduction

(a) Αριστοτέλης 384–322 BCE

Aristotle:

- “Think as the wise men think, but talk like the simple people do.”
- “The more you know, the more you know you don’t know.”

Handout available on the web: either use the QR-code

or visit https://wilfridskendall.github.io/talks/Thessaloniki-2024/.

Sketch of MCMC (I)

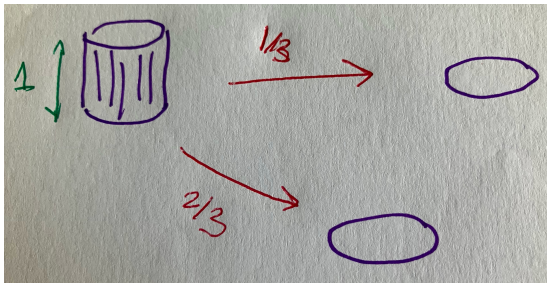
Figure 2: Edward Teller (1908-2003)

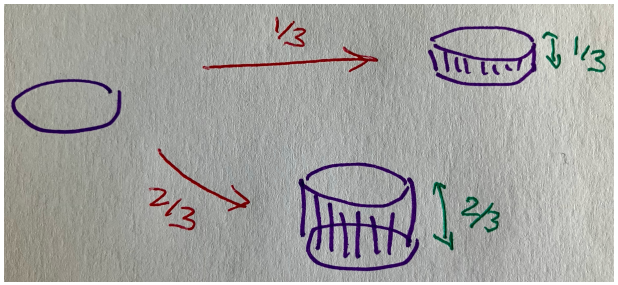
The original Markov chain Monte Carlo method (MCMC) was introduced by Metropolis *et al.* (1953). The senior author was Edward Teller ("father of the H-bomb").

[Fermi once remarked that] Teller was the only monomaniac he knew who had *several* manias (Brown & May, 2004).

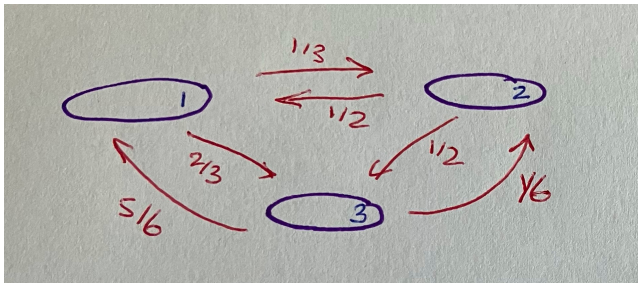
Sketch of MCMC (II): basic Markov chain theory

- Transition probabilities $p(a, b)$

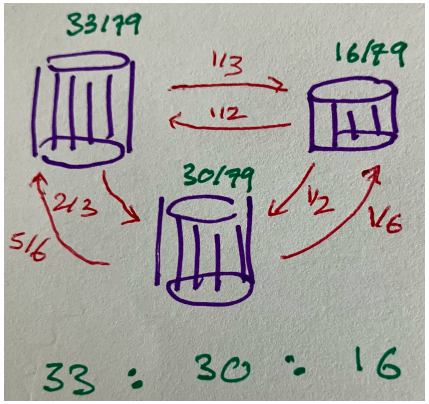




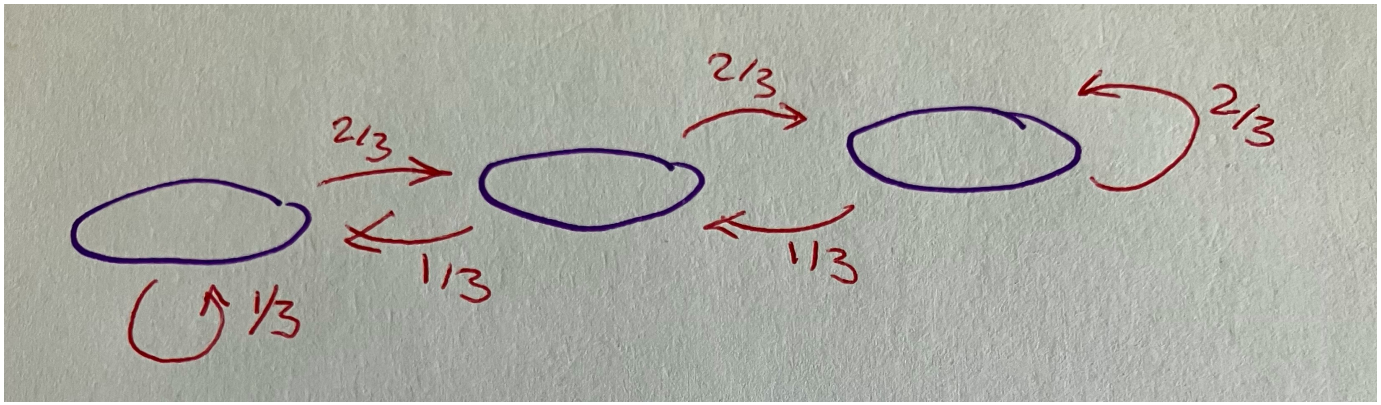
- Equilibrium probabilities $\pi(a)$ arise from balance equations;

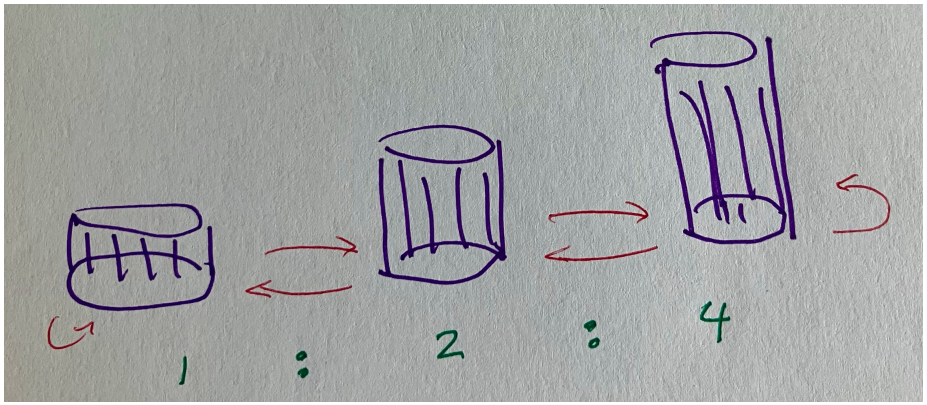


- Solve balance equations: equilibrium probabilities defined by the ratio

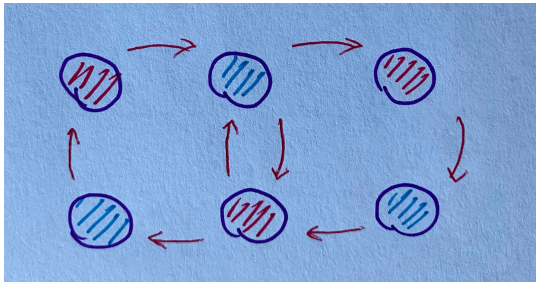


- detailed balance $\pi(a)p(a,b) = \pi(b)p(b,a)$, reversibility;

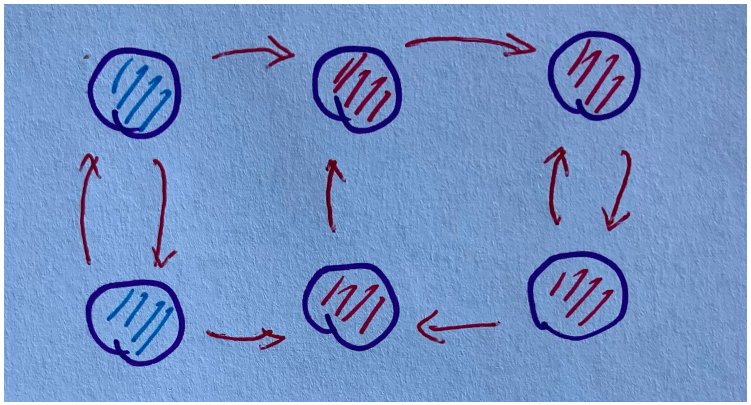




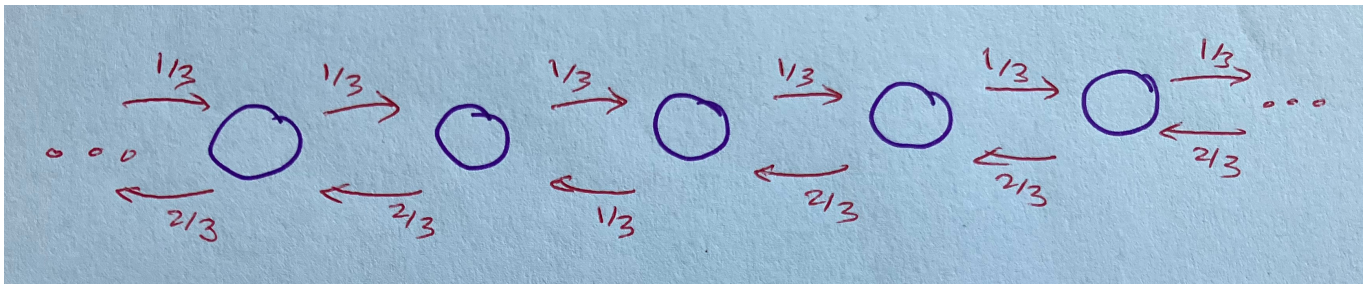
- Existence of equilibrium: we need
  - aperiodicity, not periodicity



  - irreducibility, not reducibility



  - recurrence (even uniform or geometric), not transience



Sketch of MCMC (III)

The following points are key to designing Markov chains which have the desired target equilibrium distribution.

- Under detailed balance we can condition on specific parts of the state-space by forbidding transitions;
- We can modify any chain, transition probabilities $p(a, b)$, to result in a chosen invariant target distribution $\pi(a)$. Simply censor each possible transition $a \to b$ with probability $\alpha(a, b) \in [0, 1]$ such that
$$\alpha(a, b)\pi(a)p(a, b) = \alpha(b, a)\pi(b)p(b, a);$$
- Common choice: Metropolis-Hastings
$$\alpha(a, b) = \min\{1, (\pi(b)p(b, a))/(\pi(a)p(a, b))\}.$$
- If still irreducible aperiodic recurrent, long-term equilibrium still $\pi(a)$.
- Markov chain Monte Carlo (MCMC), of great statistical importance.
- But, physicists always remind us, physicists got there fifty years earlier!

Warwick Statistics

Sketch of MCMC (IV)

Given the $\pi(a)$, how to design a Markov chain to have this as equilibrium?

(i) Independence sampler: propose a draw from a fixed probability distribution, apply Metropolis-Hastings;
(ii) Random walk Metropolis or RWM: propose a move using a random walk, apply Metropolis-Hastings;
(iii) Metropolis-adjusted Langevin or MALA: propose a Gaussian jump shifted using gradient of $\log \pi$, apply Metropolis-Hastings.

Can mix-and-match! RWM is often favourite: flexible, not too complicated.

Issues:

(A) Burn-in: *How long* till approximate equilibrium?
(B) Scaling: *How big* should be the RWM jump?

Question (B) is about how to get fast mixing. There is a beautiful and useful theory, but that is for another day.

Question (A) is what this lecture is all about.

Warwick Statistics

Sketch of MCMC (V)

- MCMC practicalities: Burn-in: what to do about it?
  - Theory tends to be much too pessimistic. Example: Zanella (2015a, 2015b) developed statistical methods for Anglo-Saxon history: a *simplified* model appeared to converge approximately in 10^5 steps (about 1 week on compute cluster), *versus* 10^9 steps in theory (around 2 centuries);
  - Is (a) one long run better or (b) many short runs? (Option (b) requires starts of short runs spread "evenly" over the sample space — almost as hard in high dimensions as the original problem!)
  - Diagnostics? (Meta-theorem: for any diagnostic technique there is a chain for which the technique is deceptive!)
  - Conclusion: effective MCMC requires very careful thought about appropriate length of run — think deeply about the problem!
- So we are looking for ways to address burn-in where possible.

Warwick Statistics

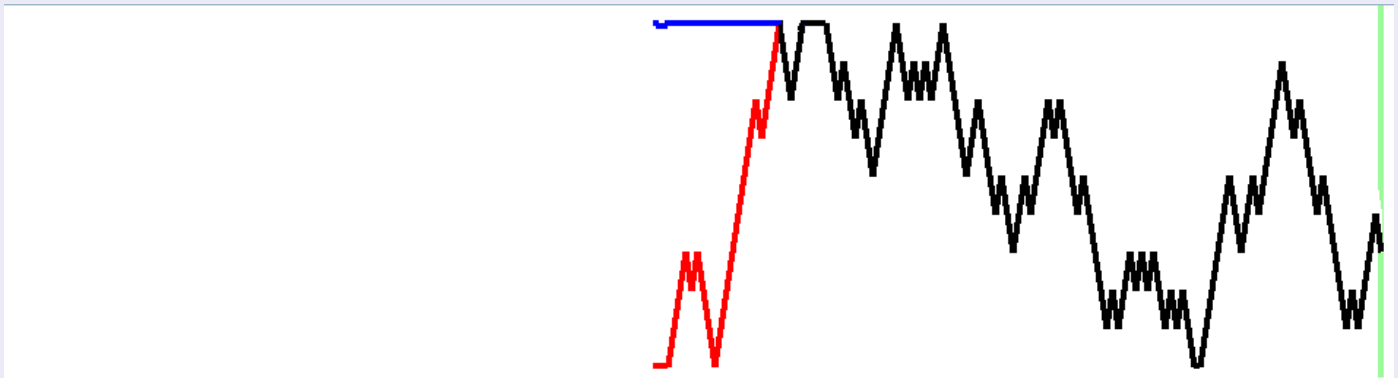
Perfect Simulation

- Propp & Wilson (1996) invented exact simulation / Coupling from the Past (CFTP) / perfect simulation;
- The key ideas of "*classic CFTP*":
  - extend simulation *backwards* through time,
  - exploit monotonicity by *coupling* maximal and minimal processes,
  - seek coalescence;
- We will study *random-walk-CFTP*, which can in fact be boosted to provide simple image reconstruction using Ising model;
- Everyone immediately suspects the term "perfect simulation" (WSK, 1998), because everyone knows it can't possibly be that good. That's exactly why the term was chosen!
- Aristotle: "Pleasure in the job puts perfection in the work."

Warwick Statistics

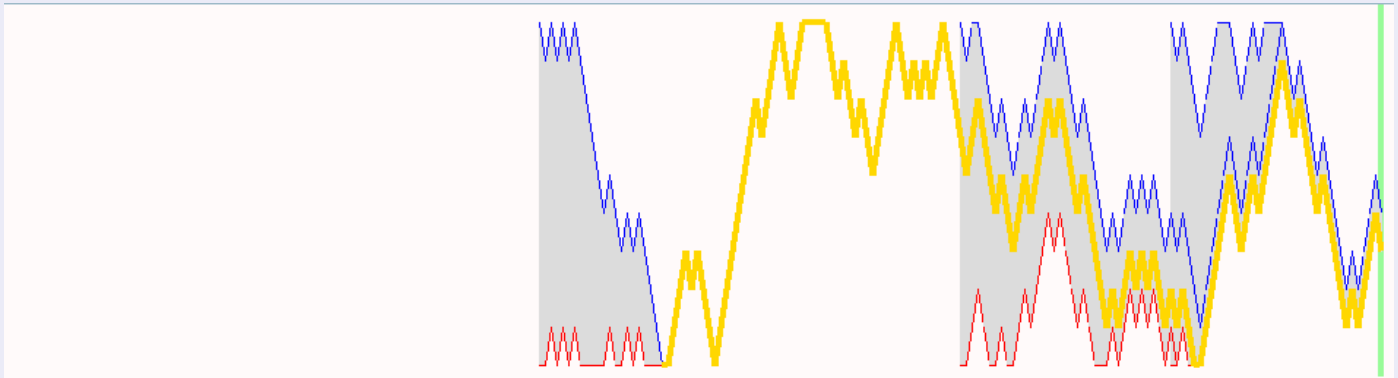
Classic CFTP for a simple random walk (I)

- Consider a simple random walk on $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.
  - probability of $+1$ jump is $p \in (0.1)$, of -1 jump is $1 - p$, **except that**
  - $+1$ jump replaced by staying still at state 9, **and**
  - -1 jump replaced by staying still at state 0.
- Conventional MCMC picks a starting point, then runs the simple random walk for long time till approximate equilibrium.
- How long? One way to *estimate* this is to run several coupled copies till they meet. If probability of meeting by time T is high, then deviation of X_T from equilibrium is statistically small,
- Generally not true that location at coupling is a draw from equilibrium.



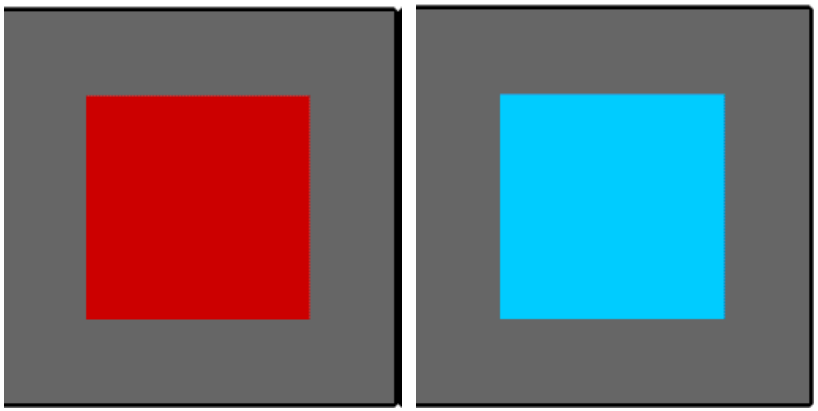
Classic CFTP for a simple random walk (I)

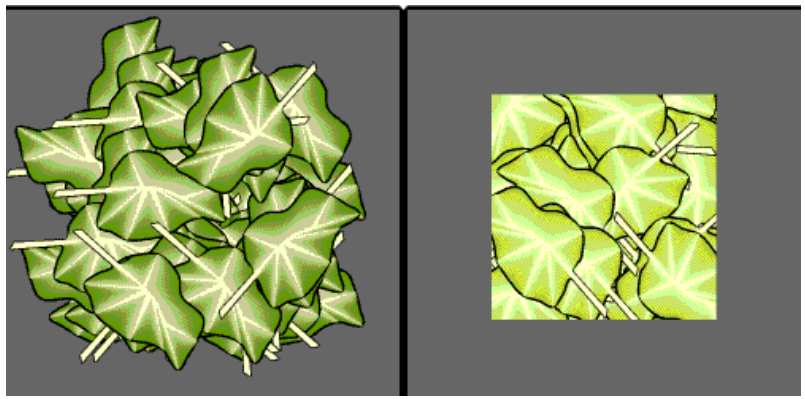
- So now start at a negative time $-T$, start at top (9) and bottom (0), and run to time 0.
- If not coupled, than back-off to time $-2T$ and repeat.
- We may need to repeat back-off several times.
- If coupled, then return the common value at time 0.
- The common value is an exact draw from equilibrium!



A very graphic example

- Many people find *DeadLeaves-CFTP* helpful (WSK & Thönnes, 1999);
- You are walking in a forest in Autumn.
- You view the leaves falling down on the red soil.
- Change your point of view! Consider a small mammal looking up from a hole in the ground to the blue sky.
- Once the hole in the ground is covered, the pattern for the mammal never changes, whereas for the walker the pattern changes ceaselessly.
- One may deduce, the animal's perceived pattern at coverage is distributed as the long-time equilibrium for the walker's pattern.





Some theory about CFTP

- What about cases where monotonicity fails? or there isn't a sensible "maximal" process? WSK (1998):
  - cross-couple upper and lower envelope processes,
  - dominate by amenable "dominating process" (time-reversible, can draw from equilibrium, can couple target processes below dominating process);
- Theoretical limits: *in principle*
  - *Classical CFTP* equivalent to uniform ergodicity (Foss & Tweedie, 1998).
  - *Dominated CFTP* is achievable under geometric ergodicity (WSK, 2004).
  - It is even possible to carry out Dominated CFTP in some **non**-geometrically ergodicity cases [Connor & WSK (2007); *nb* corrigendum];
- We can use *Dominated CFTP* to carry out perfect simulation for stable point processes (WSK & Møller, 2000);
- Detailed expositions are given by WSK (2005), Huber (2015). WSK (2015) shows how to implement CFTP in R.

Warwick Statistics

Applications to Queues and Epidemics

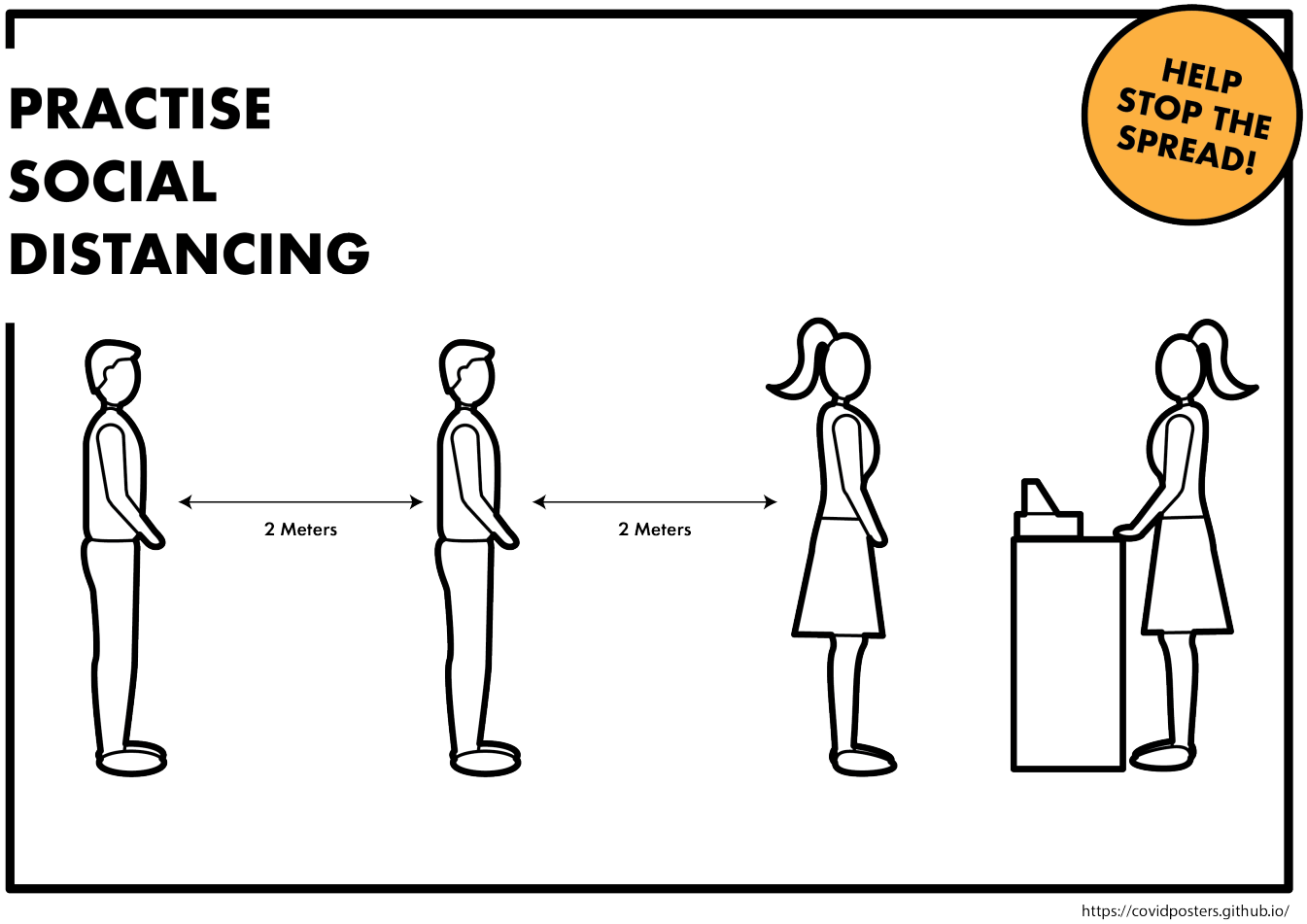


Figure 3: An illustration introducing *both* queues *and* epidemics!

Perfect Queues

The simplest queuing model ($M/M/1$: Poisson arrivals, exponential service times, single server) can be analyzed very thoroughly indeed! However:

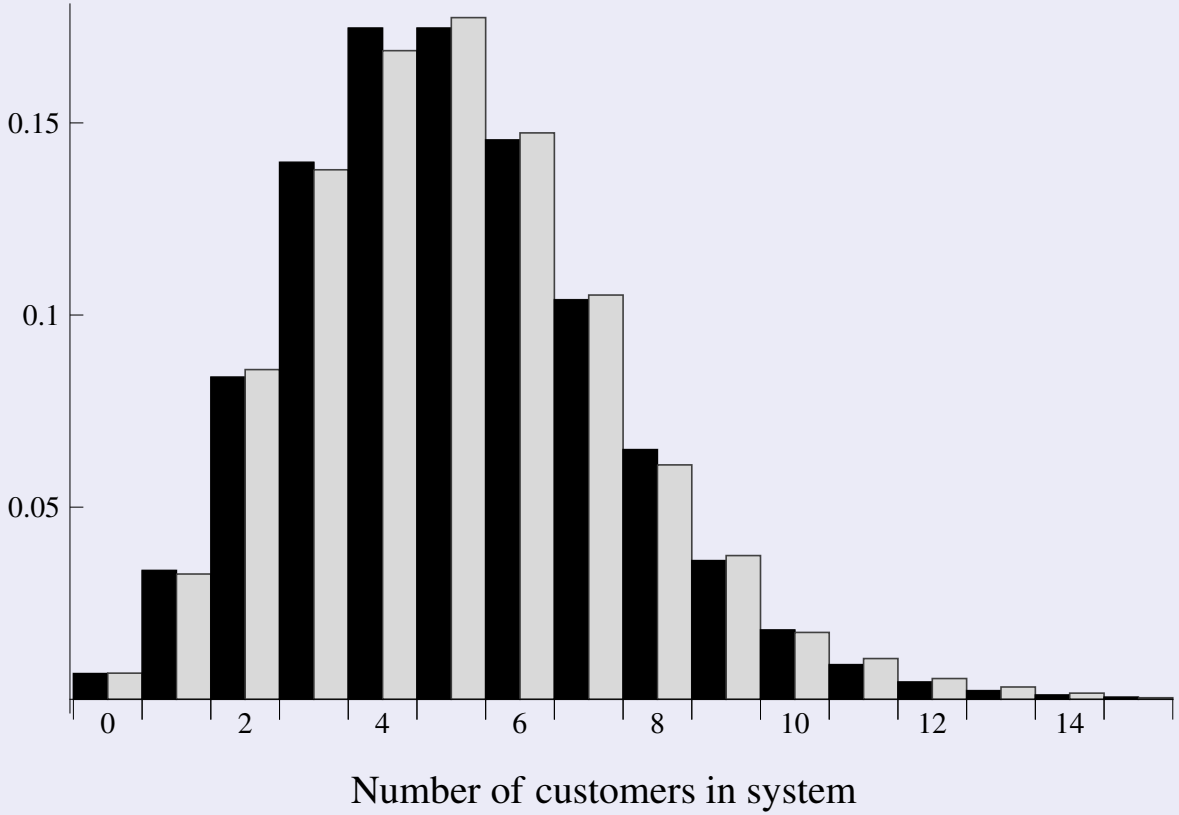
- Poisson arrivals are not unreasonable, but exponential service times can be ludicrous. The $M/G/1$ case (general service time *for just one server*) can be analyzed using the "embedded chain" (sample at each departure);
- Multi-server case: computation of *eg* waiting-time distribution is out of reach so use simulation (and insights from Kiefer & Wolfowitz, 1955);
- Sigman (2011) shows how to do CFTP in the "super-stable" case (traffic so low that it could have been handled by just one server), using Dominated CFTP and comparing to a "Processor-Sharing" discipline.

Warwick Statistics

- Connor & WSK (2015) extend Sigman (2011), showing how to apply Dominated CFTP to simulate (sub-critical!) queues perfectly; now generalized by others to the case of non-Poissonian inter-arrival times. (Technical point: pathwise domination needs service times to be assigned in order of commencement of service!) The idea is
  - dominate $M/G/c$ FCFS (FCFS means first come first served) by $M/G/c$ RA $= [M/G/1 \text{ RA}]^c$ (RA means assign to individual servers on arrival);
  - use fact that $M/G/1$ FCFS and $M/G/1$ PS *workloads* agree (PS means Processor Sharing: pool servers to serve everyone simultaneously) and $M/G/1$ PS can be simulated backwards in time;
  - so $[M/G/1 \text{ PS}]^c$ can be used to provide Dominated CFTP.
- Connor & WSK (2015) describe
  (a) CFTP coupling when dominating process empties,
  (b) and a faster CFTP coupling using upper and lower processes starting respectively at dominating process and at empty state.

Results (I)

Histogram of customer numbers for $M/M/c$ queue in equilibrium: arrival rate 10, service rate 2, and 10 servers, comparing theory (available for $M/M/c$ queue) with results of Connor & WSK (2015) algorithm.

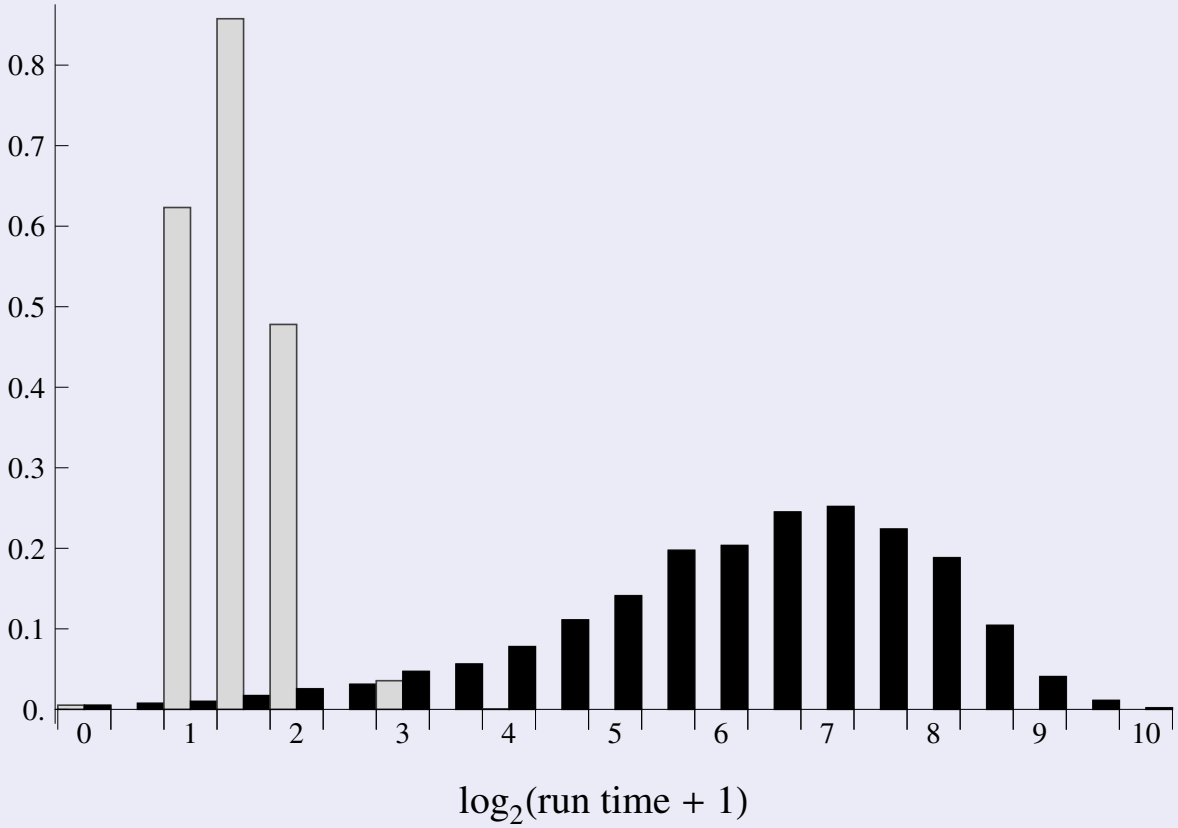


Results (II)

Comparison of log-run times for

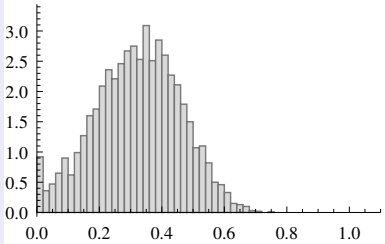
(a) CFTP coupling when dominating process empties (solid bars),

(b) a faster CFTP coupling using upper and lower processes (grey bars).

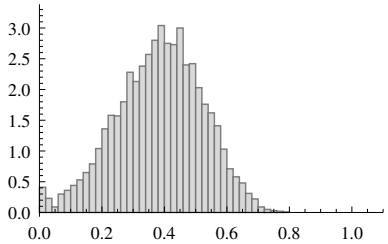


Results (III)

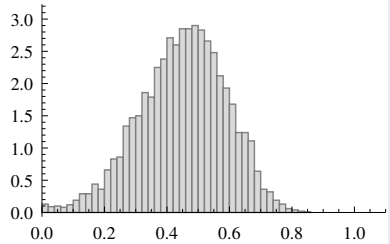
Marginal distributions of last six coordinates of the equilibrium Kiefer-Wolfowitz workload vector (anticipated work in system) for an $M/G/c = 25$ queue with Uniform$(0, 1)$ service times and arrival rate 25.



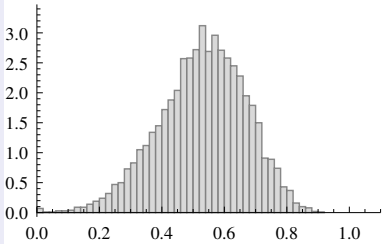
Workload at server 20



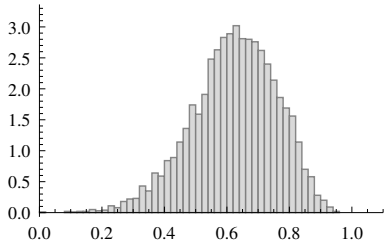
Workload at server 21



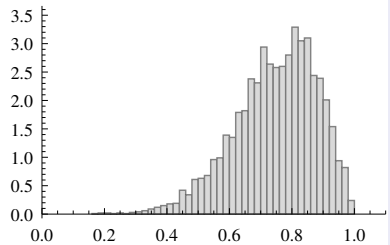
Workload at server 22



Workload at server 23



Workload at server 24



Workload at server 25

Perfect Epidemics

S-I-R deterministic epidemic: susceptibles s, infectives i, removals r:

$$\begin{aligned} s' &= -\alpha\, i\, s\,, \\ i' &= (\alpha\, s - \beta\,)\; i\,, \\ r' &= \beta\, i\,. \end{aligned}$$

(Total population $s + i + r = n$ is constant.)

S-I-R stochastic epidemic: a Markov chain (S, I, R) with transitions

Infection:	$S \to S - 1\,,$	$I \to I + 1$	at rate	$\alpha\, I\, S\,,$
Removal:	$I \to I - 1\,,$	$R \to R + 1$	at rate	$\beta\, I\,.$

Both models assume homogeneous mixing.

The first question asked about a new epidemic

"What is the R-number?"
The R-number is $\alpha\, n/\beta$: mean number of new infectives produced per infective at *start* of epidemic.
Whittle's threshold theorem: R-number $\gg 1$ means positive chance of epidemic infecting significant proportion of the population.

Wikipedia: "The British-registered *Diamond Princess* was the first cruise ship to have a major [COVID-19] outbreak on board, with the ship quarantined at Yokohama from 4 February 2020 for about a month. Of 3711 passengers and crew, around 700 people became infected and 9 people died."
Evidently $\alpha\, n/\beta \gg 1$ – as was sadly later confirmed, a sorrow for us all.

Inference on the R-number

Important, because the R-number controls severity of epidemic. However:

1. It's tough. *Either* massive assumptions (homogeneous mixing) *or* far too many parameters;
2. It's really tough. It's hard to get good information about infection times;
3. It's especially tough early on. You most need to know the answer when there is hardly any information available (I devised a simplified exercise for a Warwick second-year statistics module to show how tough this is);
4. Markov chain Monte Carlo (MCMC) can be used (O'Neill & Roberts, 1999) but what about burn-in?
5. Can we use perfect simulation?

Warwick Statistics

An easier question

1. Suppose we know n, α, β, observe removals, but *don't* observe infections which must be inferred.
2. We need a new chain: namely, the whole S-I-R trajectory evolving in *algorithmic time* using varying pattern of potential infections and removals.
3. Visualize n time lines, along which are scattered incidents:
   - potential removals, activated if time line is infected;
   - potential infections, activated if time line is infected *and* if designated target time line is lowest uninfected time line.
4. Using Poisson point processes of appropriate rates to scatter these incidents, we obtain an S-I-R epidemic.
5. Now let the point patterns evolve *in algorithmic time*, adding and removing incidents according to spatial immigration-death processes.
6. Result is a trajectory-valued chain which has unconditioned S-I-R as equilibrium.

Warwick Statistics

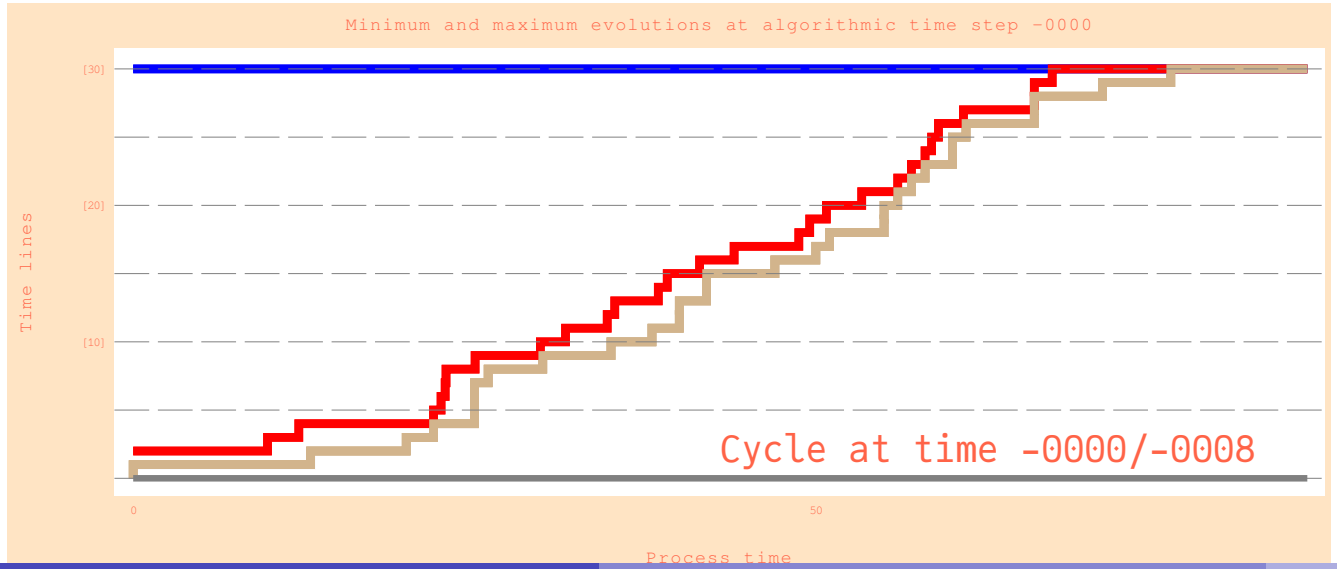
Conditioning on observed removals

- The trajectory-valued chain is *reversible*.
- So if we forbid the evolution to get rid of observed removals, and forbid it to introduce new activated removals, then the modified chain has invariant probability measure which conditions on observed pattern of removals. Implications:
  - conditioned removals can change time line (if still activated) but not time of occurrence;
  - removals can be introduced only if they don't activate;
  - sometimes infections cannot be removed (because that would result in losing a conditioned removal).
- More housekeeping details required to define algorithm precisely and make sure reversibility and mootonicity still work.
- Need to ensure irreducibility (or otherwise equilibrium will depend on starting point).
- Does this produce a feasible algorithm?

Warwick Statistics

Example

- Smallpox outbreak in a closed community of 120 individuals in Abakaliki, Nigeria (Bailey, 1975; O'Neill & Roberts, 1999).
- **Assume**
  - first observed removal is also the first removal: under a plausible improper prior we can then deduce what is the distribution of infectives I_0 at time 0;
  - all removals are recorded;
  - there are no further removals after the last observed removal.
- Coding in *julia* (Bezanson *et al.*, 2017), we can construct a perfect simulation GIF resulting in a draw from the conditional distribution of pattern of infections.



So what?

- You may be wondering, why this emphasis on unobserved infections given fixed α and β, when what we really want is inference on R-number $\alpha\, n/\beta$ for *unknown* α and β?
- Good question. But a re-weighting argument allows us to get (unbiased) estimates based on *different* α and β. Essentially the perfect simulation provides an exact simulation-based computation which permits us to integrate out the pattern of unobserved infections.
- This means we can (**work in progress**, Connor & WSK, 2024)
  - construct a steepest ascent algorithm (in effect, a variant on a Robbins-Monro stochastic optimization algorithm) to find *maximum a posterior* estimates of α and β;
  - or even, with some computational effort, compute the entire posterior joint density for α and β!

Warwick Statistics

Conclusion

- Are you worried about burn-in issues when doing MCMC. Consider whether perfect simulation can be applied!
- CFTP works even for significantly complex and relevant models of real-life phenomena.
- *Of course* really detailed models will resist perfect simulation: but it can be helpful to compare with a simpler model (especially, using fewer parameters).
- CFTP is clearly an important tool to be considered by the investigator seeking to implement accurate and informative MCMC.
- Thank you for your attention! QUESTIONS?

Warwick Statistics

References I

Bailey, N.T.J. (1975) *The mathematical theory of infectious diseases and its applications*, 2nd Ed. ed. Griffin.

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V.B. (2017) Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, **59**, 65–98.

Brown, H. & May, M. (2004) Edward Teller in the Public Arena. *Physics Today*, **57**, 51–53.

Connor, S.B. & WSK (2007) Perfect simulation for a class of positive recurrent Markov chains. *Annals of Applied Probability*, **17**, 781–808.

Connor, S.B. & WSK (2015) Perfect simulation of M/G/c queues. *Advances in Applied Probability*, **47**, 1039–1063.

Connor, S.B. & WSK (2024) Perfect Epidemics.

Foss, S.G. & Tweedie, R.L. (1998) Perfect simulation and backward coupling. *Stochastic Models*, **14**, 187–203.

Fraser, C. & Others (2023) OpenABM-Covid19: Agent-based model for modelling the Covid-19 and Contact-Tracing.

Huber, M.L. (2015) *Perfect Simulation*. Boca Raton: Chapman; Hall/CRC.

Kiefer, J. & Wolfowitz, J. (1955) On the Theory of Queues With Many Servers. *Transactions of the American Mathematical Society*, **78**, 1.

Warwick Statistics

References II

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., & Teller, E. (1953) Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, **21**, 1087.

Mizumoto, K., Kagaya, K., Zarebski, A., & Chowell, G. (2020) Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. *Eurosurveillance*, **25**, 6pp.

O'Neill, P.D. & Roberts, G.O. (1999) Bayesian Inference for Partially Observed Stochastic Epidemics. *Journal of the Royal Statistical Society Series A: Statistics in Society*, **162**, 121–129.

Propp, J.G. & Wilson, D.B. (1996) Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, **9**, 223–252.

R Development Core Team (2010) R: A Language and Environment for Statistical Computing.

Sigman, K. (2011) Exact simulation of the stationary distribution of the FIFO M/G/c queue. *Journal of Applied Probability*, **48**, 209–213.

WSK (1998) Perfect Simulation for the Area-Interaction Point Process. *Probability towards 2000* (Accardi, L. & Heyde, C.C. eds). Springer-Verlag, pp. 218–234.

WSK (2004) Geometric ergodicity and perfect simulation. *Electronic Communications in Probability*, **9**, 140–151.

WSK (2005) Notes on Perfect Simulation. Singapore: World Scientific, pp. 93–146.

Warwick Statistics

References III

WSK (2015) Introduction to CFTP using R. *Stochastic geometry, spatial statistics and random fields*, *Lecture notes in mathematics*. Springer, pp. 405–439.

WSK & Møller, J. (2000) Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes. *Advances in Applied Probability*, **32**, 844–865.

WSK & Thönnes, E. (1999) Perfect simulation in stochastic geometry. *Pattern Recognition*, **32**, 1569–1586.

Zanella, G. (2015b) Bayesian Complementary Clustering, MCMC, and Anglo-Saxon Placenames (PhD Thesis).

Zanella, G. (2015a) Random partition models and complementary clustering of Anglo-Saxon place-names. *Annals of Applied Statistics*, **9**, 1792–1822.

Warwick Statistics

Technical information

Image	Attribution	
Aristotle	After Lysippos	*Public domain* via Wikimedia Commons
Edward Teller	Lawrence Livermore National Laboratory restored by w:User:Greg L, Papa Lima Whiskey	*CC BY-SA 3.0* via Wikimedia Commons
Perfect Ising	Result of code written by WSK	
Classic CFTP and Dead leaves	Result of code written by WSK	
Queues	https://covidposters.github.io/	*Open source*
$M/M/c$ customers	Result of code written by Stephen Connor	
$M/M/c$ runtimes	Result of code written by Stephen Connor	
$M/M/c$ loads	Result of code written by Stephen Connor	
Diamond Princess	Alpsdake	*CC BY-SA 4.0*
Epidemic	Result of code written by WSK	

These notes were produced from `Thessaloniki-2024.qmd`:

Author:	Wilfrid Kendall W.S.Kendall@warwick.ac.uk
Date:	Sat May 18 15:12:32 2024 +0100
Summary:	Final version of Thessaoniki talk. Added “Notes for next time” in `TODO.md`.

Warwick Statistics