A Telescopic Introduction to Multilevel Monte Carlo for Simulation and Inference

Data Science Under the Hood

David J. Warne

School of Mathematical Sciences, Queensland University of Technology

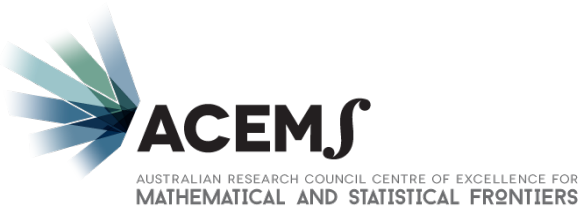
Centre for Data Science, Queensland University of Technology

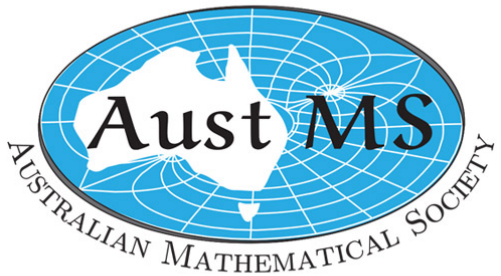
ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS)

13 May 2021

Acknowledgments

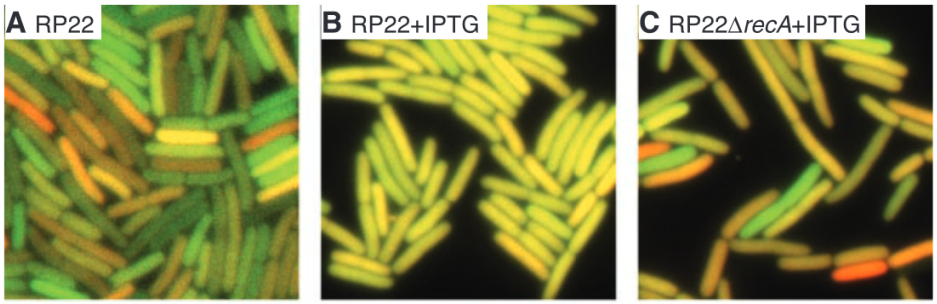
- **People:**
  Matthew Simpson (QUT), Ruth Baker (Oxford), Tom Prescott (Oxford), Mike Giles (Oxford), Chris Lester (Oxford).
- **Funding:**
  - ACEMS
  - Centre for Data Science at QUT
  - Australian Mathematical Society (Lift-off Fellowship)



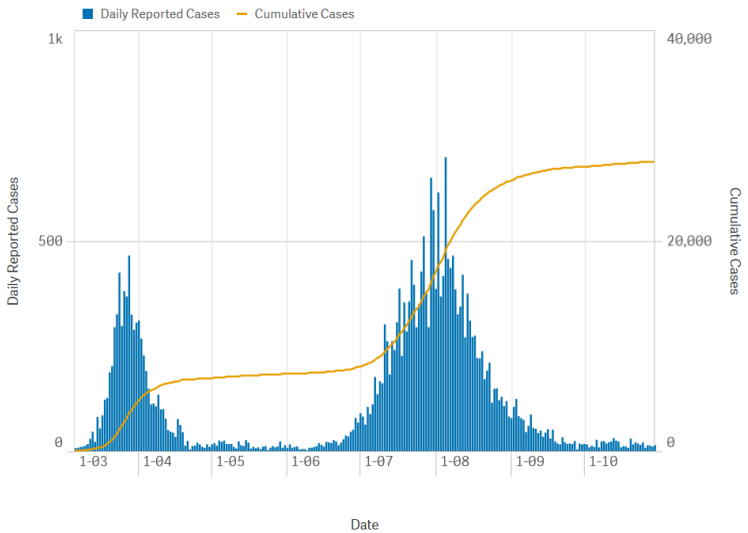


Motivation: Stochastic Systems in the Real World

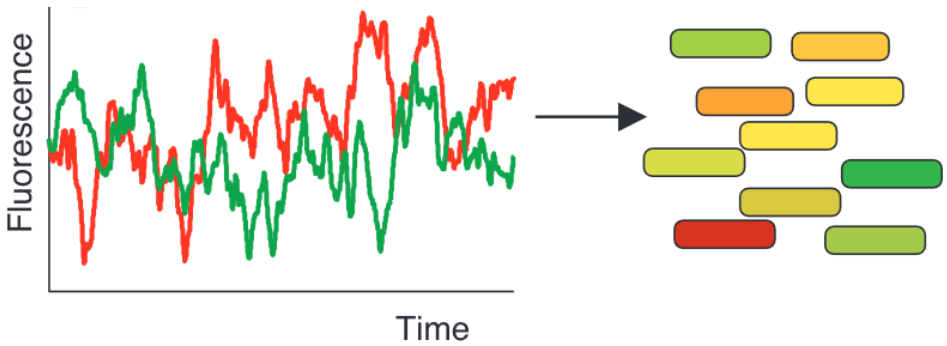
Finance



Cellular processes



Epidemics



Gene expression Elowitz et al. (2002) *Science*, v297

Example: Biochemical Reaction Networks

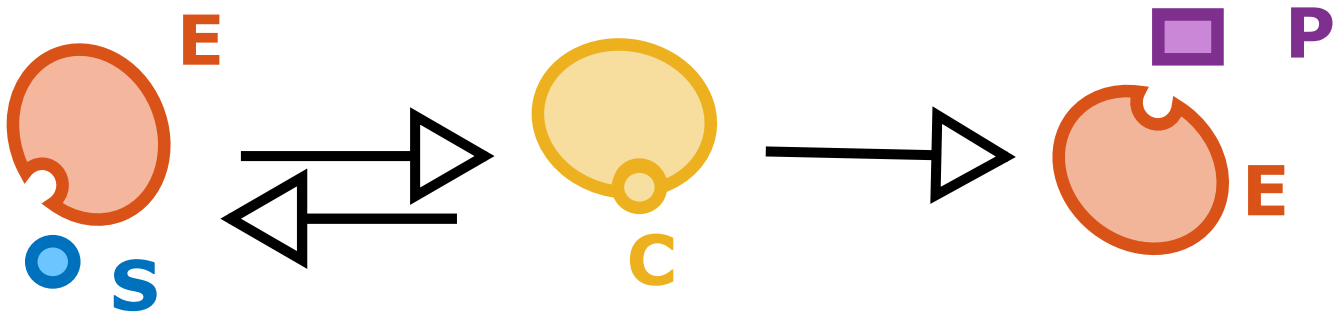
Let $X \in \mathbb{N}^{1\times N}$ be the vector of populations of N chemical species, interacting according to the M reactions,

$$\sum_{i=1}^{N} X_i \nu_{i,j}^{-} \xrightarrow{k_j} \sum_{i=1}^{N} X_i \nu_{i,j}^{+}, \quad j = 1, 2, \ldots M.$$

where ν^{-}, $\nu^{+} \in \mathbb{N}^{N\times M}$ are called stoichiometries. k_j is the rate parameter and $\nu_j = (\nu_{*,j}^{+} - \nu_{*,j}^{-})^{\mathsf{T}}$ is the state change for reaction j.

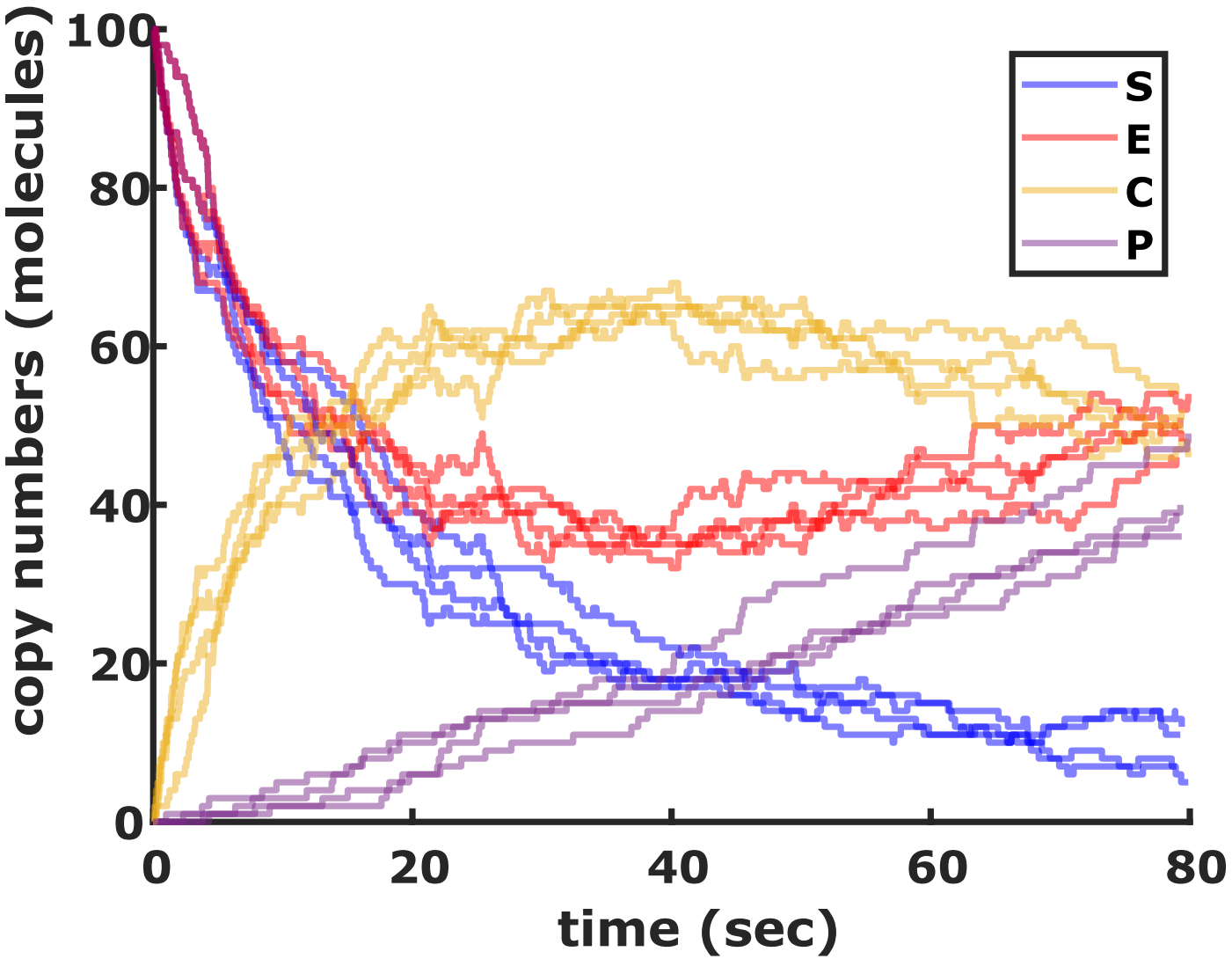
Example: Michaelis-Menten enzyme kinetics

- $X = [S, E, C, P]$; substrate (S), enzyme (E), complex (C), and product (P);
- Reaction 1: $S + E \xrightarrow{k_1} C$, with $\nu_1 = [-1, -1, 1, 0]$, propensity $k_1 SE$;
- Reaction 2: $C \xrightarrow{k_2} S + E$, with $\nu_2 = [1, 1, -1, 0]$, propensity $k_2 C$;
- Reaction 3: $C \xrightarrow{k_3} P + E$, with $\nu_3 = [0, 1, -1, 1]$, propensity $k_3 C$;



Example Realisations

Initially, $S = E = 100$, and $C = P = 0$. Rates: $k_1 = 0.001$, $k_2 = 0.005$ and $k_3 = 0.01$.



Exact Stochastic Simulation

- We can simulate exact sample paths from these systems.

Gillespie's method (omitting some details)

1. Start with system at time t with state X;
2. Draw a random variable, $\Delta t > 0$, for the next reaction time $t + \Delta t$;
3. Randomly select a reaction $j \in [1, M]$ to occur;
4. Update state and time base on reaction event $X \leftarrow X + \nu_j$ and $t \leftarrow t + \Delta t$;
5. Repeat until $t > T$.

- Often we are interested in estimating some averaged behaviour,

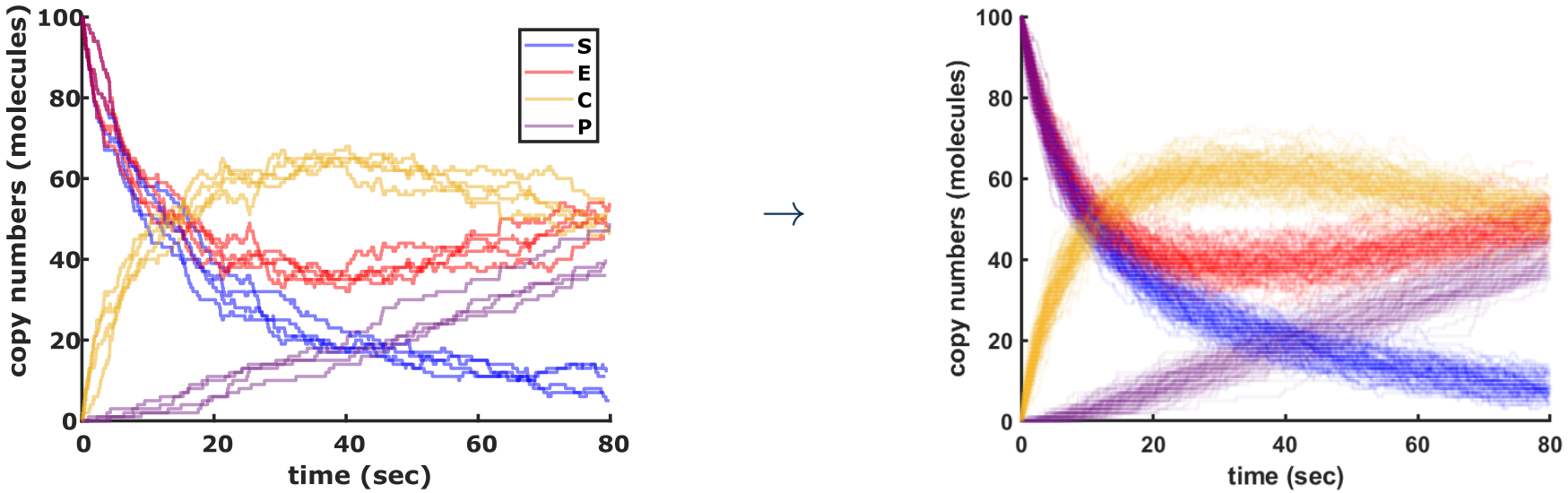
$$\mathbb{E}\left[f(X_T)\right] = \int f(X_T)p(X_T)\mathrm{d}X_T,$$

where f is some "well behaved function" and $p(X_T)$ is the state probability density at time T. Examples:

- $f(X_T) = X_T$, leads to the average state $\mu = \mathbb{E}\left[X_T\right]$,
- $f(X_T) = (X_T - \mu)^2$, leads to the variance of the state $v = \mathbb{V}\left[X_T\right]$
- $f(X_T) = 1$ if $X_T < x$ and $f(X_T) = 0$ otherwise, leads to $\mathbb{P}(X_T < x)$ (i.e. cumulative distribution function $F(x)$).

We don't typically have access to $p(X_T)$, so we use repeated simulations.

$N = 4$ $N = 100$



Then we estimate the expectation using realisations, $X_T^1, \ldots, X_T^N$,

$$\mathbb{E}\left[f(X_T)\right] \approx \hat{f} = \frac{1}{N}\sum_{i=1}^{N} f(X_T^i).$$

Note, $\mathbb{E}[\hat{f}] \to \mathbb{E}\left[f(X_T)\right]$ and $\mathbb{V}[\hat{f}] \to \mathbb{V}\left[f(X_T)\right]/N$ as $N \to \infty$.

For high precision estimates we need large N, which can be prohibitive.

Approximate Stochastic Simulation

Aim: to reduce simulation cost of each realisation.

Assume propensities constant over time interval of length $\tau > 0$.

Tau-leaping method (omitting some details)

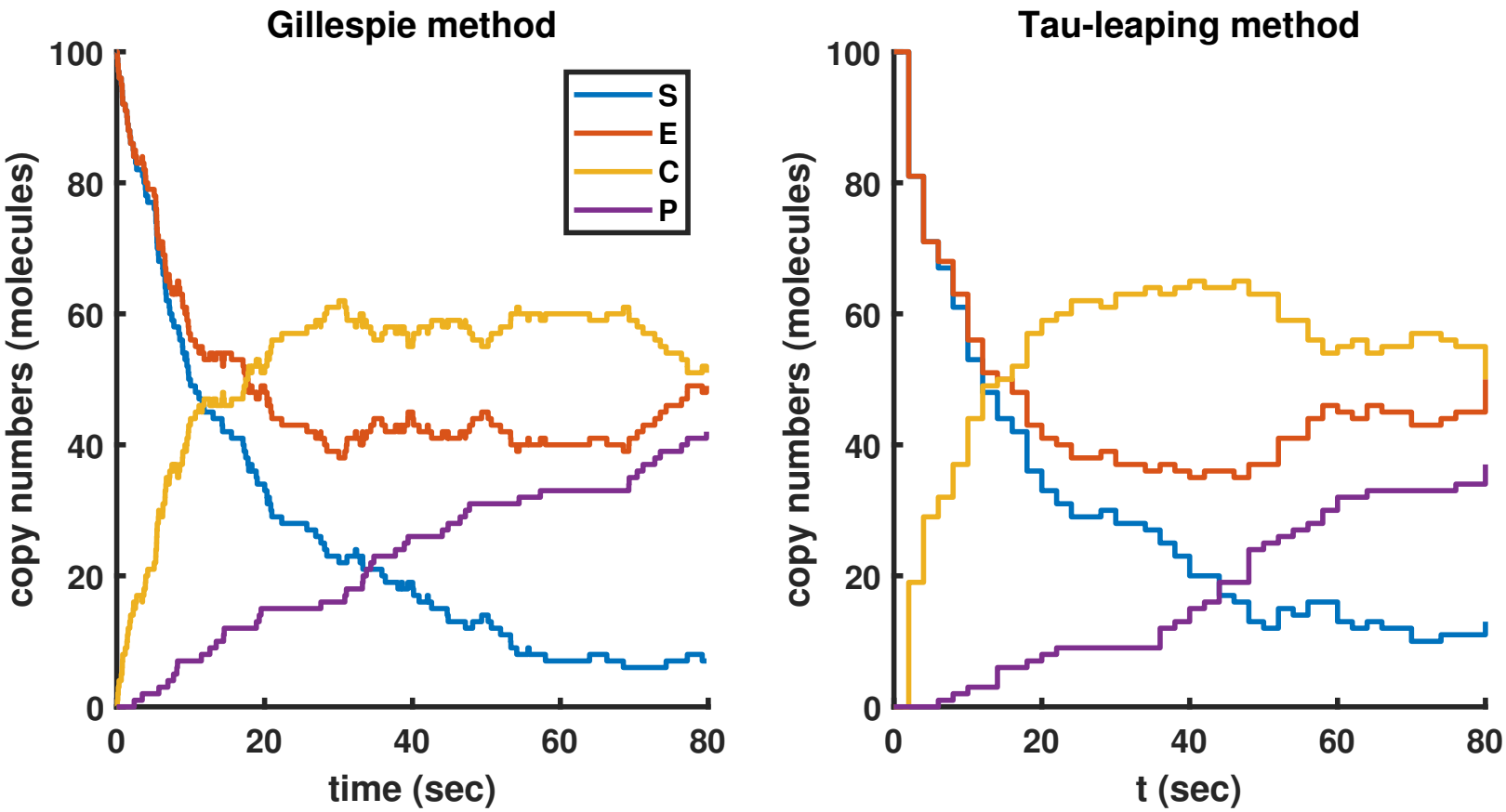
1. Start with system state Z at time t;
2. Draw random variables $Y_1, \ldots, Y_M$ that count reaction events over $[t, t+\tau)$;
3. Update state and time $Z \leftarrow Z + \sum_{j=1}^{M} Y_j \nu_j$ and $t \leftarrow t + \tau$;
4. Repeat until $t > T$.

Now we have a fixed compute cost of T/τ steps per realisation;

BUT! Our simulations are not exact any more, so $\mathbb{E}[f(Z_T)] \neq \mathbb{E}[f(X_T)]$ in general.

Exact vs Approximate Stochastic Simulation

Initially, $S = E = 100$, and $C = P = 0$. Rates: $k_1 = 0.001$, $k_2 = 0.005$ and $k_3 = 0.01$. For approximation, $\tau = 2$.



Computational Challenge

Suppose we use Z_T for Monte Carlo estimate of $\mathbb{E}[f(X_T)]$.

That is, choose τ small enough so

$$\mathbb{E}[f(X_T)] \approx \mathbb{E}[f(Z_T)] \approx \hat{f}_Z = \frac{1}{N}\sum_{i=1}^{N} f(Z_T^i).$$

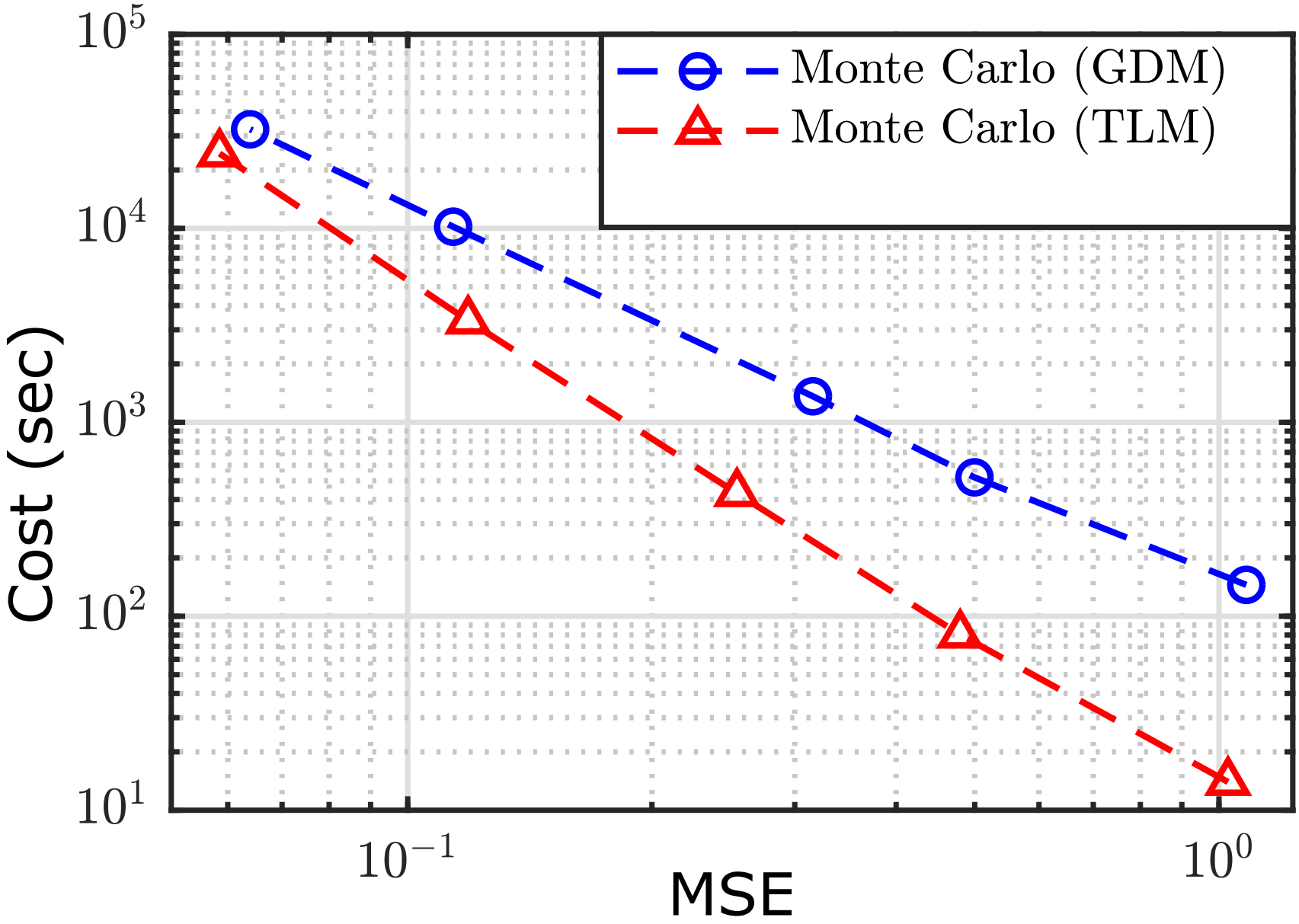
What is small enough? Think in terms of mean squared error (MSE),

$$\text{MSE} = \text{bias}^2 + \text{variance}.$$

Informally, we have bias $\propto \tau$, variance $\propto 1/N$ and cost $\propto 1/\tau$.

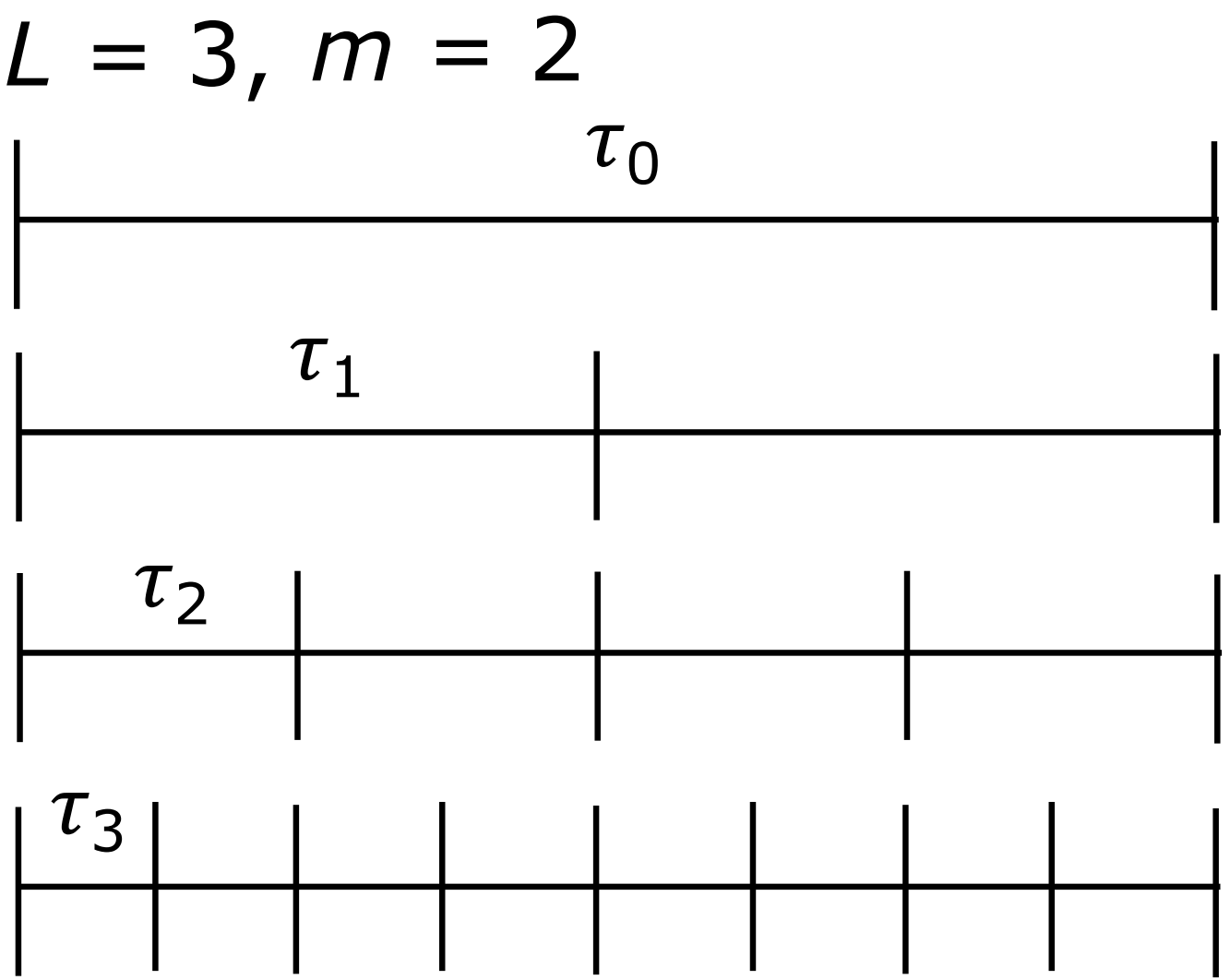
Therefore, achieving MSE $\propto h^2$ requires cost $\propto N/h \propto 1/h^3$. I.e., computational cost scales poorly as $h \to 0$ and will eventually be more costly than exact simulation.

Monte Carlo Performance (Exact vs Approximate)



Key Idea: Multilevel Telescoping Sum (Giles, 2008)

For $\ell = 0, 1, \ldots, L$, denote $Z_{\ell,T}$ as an approximation to X_T using $\tau_\ell \propto m^{-\ell}$.



Key Idea: Multilevel Telescoping Sum (Giles, 2008)

For $\ell = 0, 1, \ldots, L$, denote $Z_{\ell,T}$ as an approximation to X_T using $\tau_\ell \propto m^{-\ell}$.

$$\mathbb{E}\left[f(X_T)\right] \approx \underbrace{\mathbb{E}\left[f(Z_{L,T})\right]}_{\substack{\text{low bias} \\ \text{approximation}}}$$

$$= \underbrace{\mathbb{E}\left[f(Z_{L-1,T})\right]}_{\substack{\text{slightly biased} \\ \text{approximation}}} + \underbrace{\mathbb{E}\left[f(Z_{L,T}) - f(Z_{L-1,T})\right]}_{\text{bias correction}}$$

$$= \underbrace{\mathbb{E}\left[f(Z_{L-2,T})\right]}_{\substack{\text{slightly more biased} \\ \text{approximation}}} + \underbrace{\mathbb{E}\left[f(Z_{L-1,T}) - f(Z_{L-2,T})\right] + \mathbb{E}\left[f(Z_{L,T}) - f(Z_{L-1,T})\right]}_{\text{two bias corrections}}$$

$$\vdots$$

$$= \underbrace{\mathbb{E}\left[f(Z_{0,T})\right]}_{\substack{\text{very biased} \\ \text{approximation}}} + \underbrace{\sum_{\ell=1}^{L} \mathbb{E}\left[f(Z_{\ell,T}) - f(Z_{\ell-1,T})\right]}_{L \text{ bias corrections}}.$$

Seems like a bad idea? (recall for independent r.v., $\mathbb{V}\left[X - Y\right] = \mathbb{V}\left[X\right] + \mathbb{V}\left[Y\right]$)

Coupling Approximate Simulation Paths

In the tau-leaping method, $Y_1, \ldots, Y_M$ are Poisson random variables representing the number of events over the interval $[t, t+\tau)$.

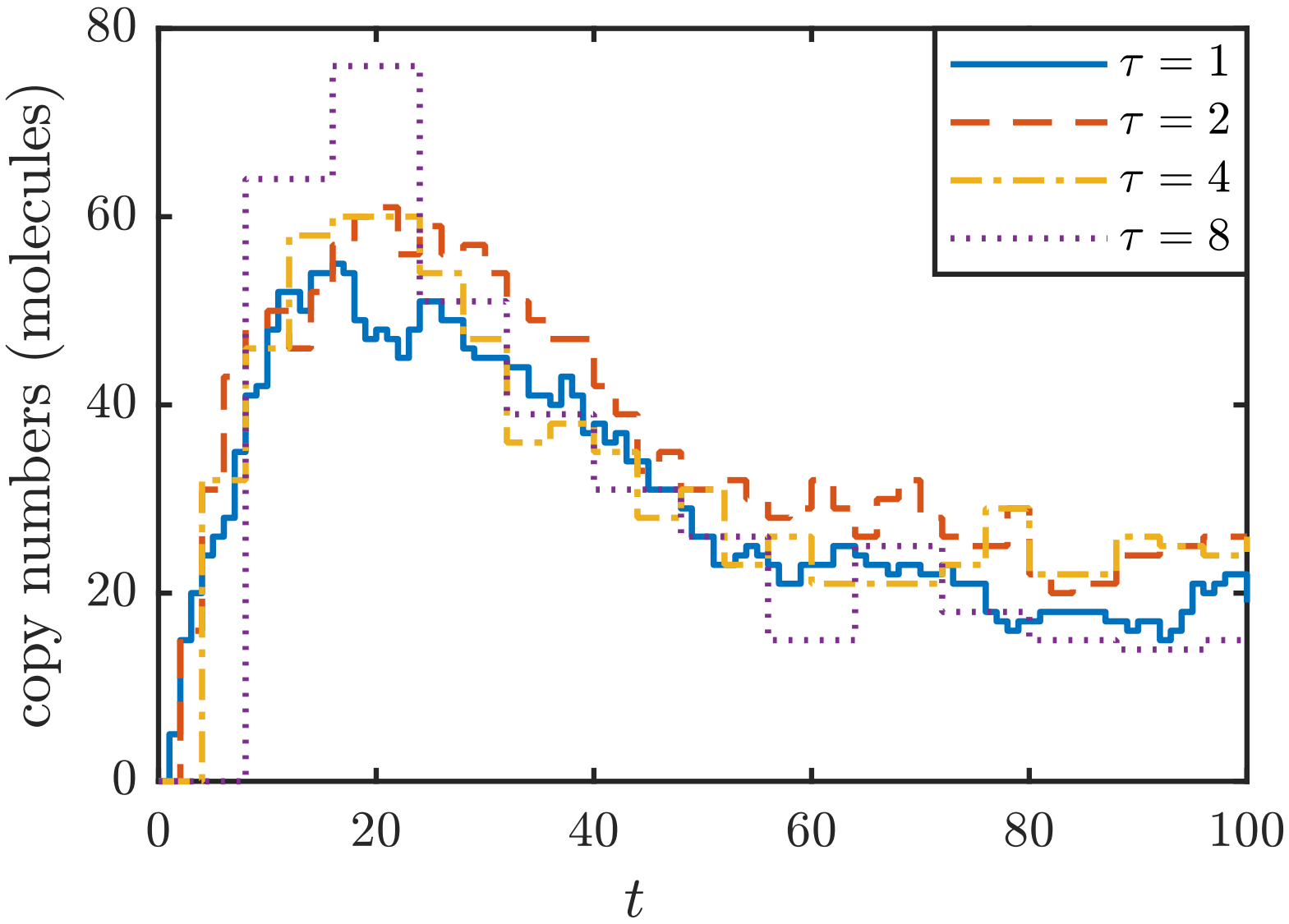
Thickening property: $\text{Poisson}(a) + \text{Poisson}(b) = \text{Poisson}(a+b)$.

That is, can use m steps of length τ_ℓ to generate one step of length $\tau_{\ell-1}$.

The result is a coupled pair of paths $(Z_{\ell,t}, Z_{\ell-1,t})$, that represent two approximations of the *same* exact sample path X_t.

This does not violate the telescoping sum.

Sample pairs



Variance Reduction and Optimal Sample Sizes

This scheme induces a positive correlation between $(Z_{\ell,t}, Z_{\ell-1,t})$ pairs.

Recall: for correlated r.v. $\mathbb{V}[X - Y] = \mathbb{V}[X] + \mathbb{V}[Y] - 2\mathbb{C}[X, Y]$.

That is we get a variance reduction in the correction estimator

$$\hat{B}_\ell = \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} \left[f(Z^i_{\ell,t}) - f(Z^i_{\ell-1,t}) \right].$$

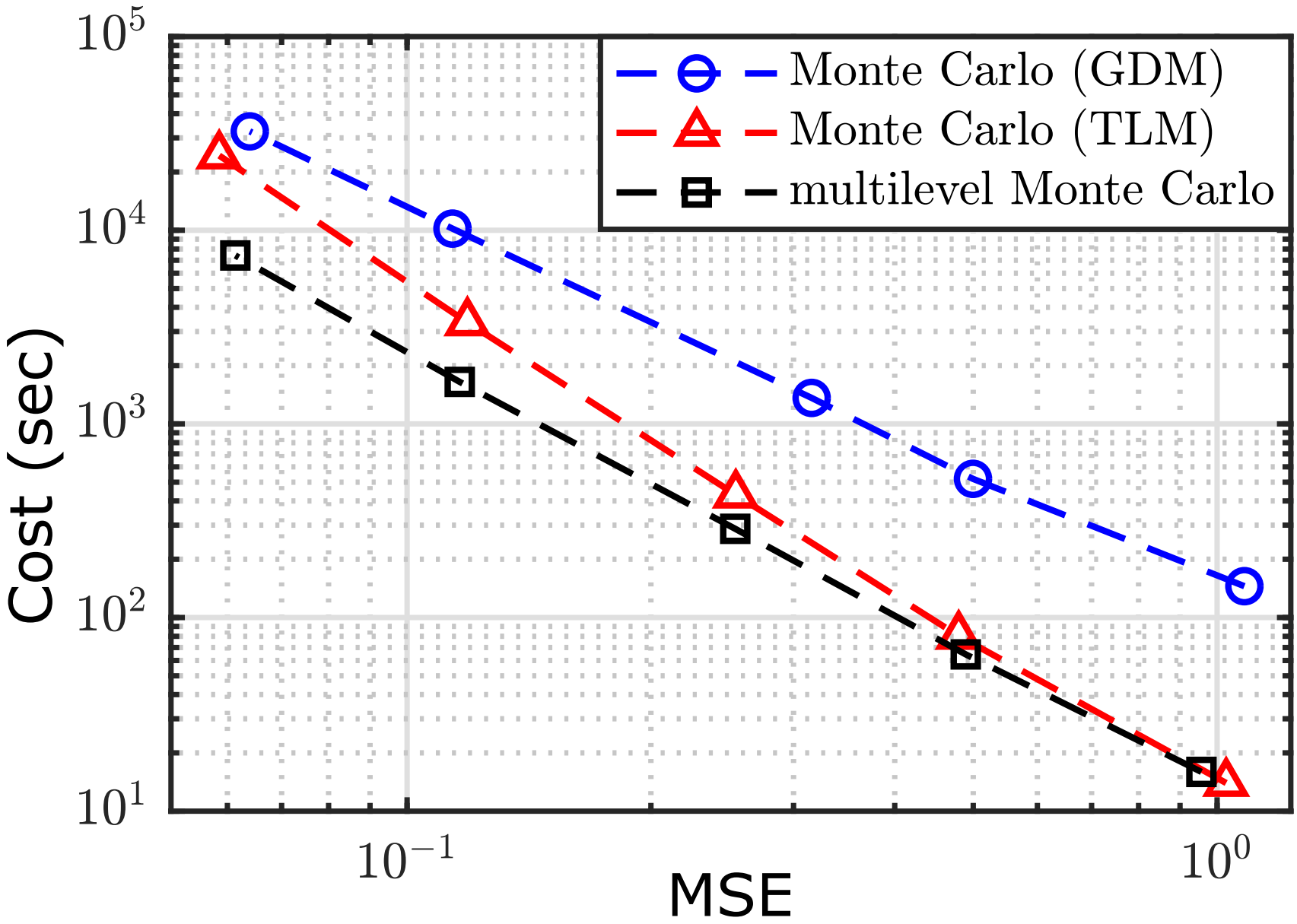
We can use path-wise convergence properties to show that $\mathbb{V}\left[\hat{B}_\ell\right] \propto \tau_\ell / N_\ell$.

For target MSE $\propto h^2$ we can optimise the choice of L and N_ℓ for $\ell = 1, \ldots, L$.

$$L \propto \frac{-\log h}{\log m}, \quad N_\ell \propto \frac{1}{h^2} \sqrt{\frac{v_\ell}{c_\ell}} \sum_{n=0}^{L} \sqrt{c_n v_n}$$

where v_ℓ and c_ℓ are the variance an cost of each level.

Expected cost (one of three cases): $1/h^2$; $(\log h)^2/h^2$; $1/h^{2+(1-\delta)}$ for $\delta \in (0, 1)$


Monte Carlo (GDM)
Monte Carlo (TLM)
multilevel Monte Carlo
Cost (sec)
MSE
10^5
10^4
10^3
10^2
10^1
10^{-1}
10^0

Further Extensions for MLMC Simulation

- Exact coupling between Gillespie and tau-leap paths. I.e. MLMC is unbiased regardless of L;
- Adaptive time-stepping, higher-order schemes, implicit schemes;
- Analysis and extensions for functions f that are not "nice";
- Multi-index Monte Carlo (for stochastic PDEs);
- Randomised bias corrections to enable unbiased estimators when exact simulation is unavailable.

MLMC for Inference?

In a Bayesian context, we want to estimate expectation with respect to the posterior distribution of parameters, $\boldsymbol{\theta}$, given data, $\mathcal{D}$.

$$\mathbb{E}\left[f(\boldsymbol{\theta}) \mid \mathcal{D}\right] = \int f(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D}) \mathrm{d}\boldsymbol{\theta},$$

where $p(\boldsymbol{\theta} \mid \mathcal{D}) \propto p(\mathcal{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$. Sure, we can write down the telescoping sum:

$$\mathbb{E}\left[f(\boldsymbol{\theta}) \mid \mathcal{D}\right] \approx \mathbb{E}\left[f(\boldsymbol{\theta}_0) \mid \mathcal{D}\right] + \sum_{\ell=1}^{L} \mathbb{E}\left[f(\boldsymbol{\theta}_\ell) - f(\boldsymbol{\theta}_{\ell-1}) \mid \mathcal{D}\right],$$

but what does it really mean here? What are our levels? How do we sample each level? Coupling mechanisms?

This can get really tricky.

Application to Inference of Rate Parameters

In real biological studies:

- Don't known kinetic rates, $\boldsymbol{\theta} = [k_1, \ldots, k_M]$;
- Observation error, $\mathrm{Y}_t = g(\mathrm{X}_t)$;
- Few observations, $\mathrm{Y}_{obs} = [\mathrm{Y}_{t_1}, \mathrm{Y}_{t_2}, \ldots, \mathrm{Y}_{t_n}]$, and n is small.

- $p(\mathrm{Y}_{obs} \mid \boldsymbol{\theta})$ intractable (sort of);
- MLMC has been very successful in the forwards problem;
- How can we use MLMC for the inverse problem?

Approximate Bayesian Computation (ABC)

The simplest way to implement ABC methods; generates n i.i.d samples from $p(\boldsymbol{\theta} \mid \rho(\mathrm{Y}^S, \mathrm{Y}_{obs}) \leq \epsilon)$.

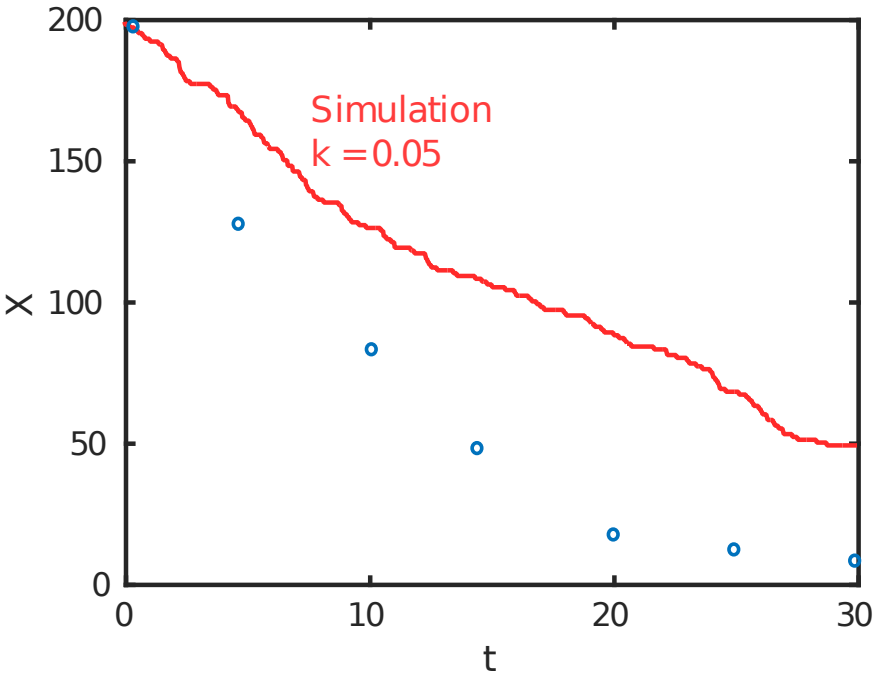
ABC Rejection Sampling

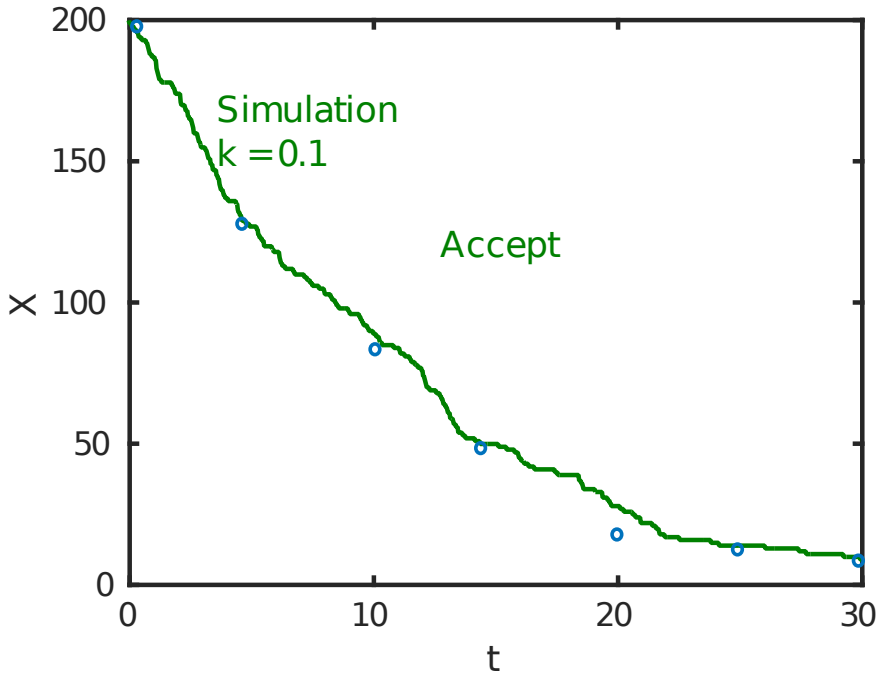
1: **for** $i = 1, \ldots, n$ **do**
2: **repeat**
3: Sample prior $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta})$
4: Generate simulated data $\mathrm{Y}^S \sim s(\mathrm{Y} \mid \boldsymbol{\theta}^*)$
5: **until** $\rho(\mathrm{Y}^S, \mathrm{Y}_{obs}) \leq \epsilon$
6: Set $\boldsymbol{\theta}^i = \boldsymbol{\theta}^*$
7: **end for**

We want ϵ small due to bias, but cost per sample ϵ^{-q} where q is the data dimensionality.

ABC Rejection Sampling Example

$$X \xrightarrow{k} Y$$





Multilevel ABC Estimator (Warne, Baker, Simpson, 2018)

We introduce MLMC by:

- Sequence of thresholds, $\{\epsilon_\ell\}_{L \geq \ell \geq 0}$, with $\epsilon_\ell > \epsilon_{\ell+1}$;
- Yields ABC approximations $\boldsymbol{\theta}_\ell \sim p(\boldsymbol{\theta} \mid \rho(\mathrm{Y}^S_{obs}, \mathrm{Y}_{obs}) \leq \epsilon_\ell)$.

The Monte Carlo estimator, $\hat{F}_\ell(\mathrm{s}) = \sum_{\ell=0}^{L} \hat{Y}_\ell(\mathrm{s})$, for $\mathrm{s} \in \mathbb{R}^M$,

$$\hat{Y}_\ell(\mathrm{s}) = \begin{cases} \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} G_{D(\mathrm{s})}(\boldsymbol{\theta}_\ell^i) & \ell = 0 \\ \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} G_{D(\mathrm{s})}(\boldsymbol{\theta}_\ell^i) - G_{D(\mathrm{s})}(\boldsymbol{\theta}_{\ell-1}^i) & \ell > 0 \end{cases}$$

where $G_{D(\mathrm{s})}(\boldsymbol{\theta})$ is a Lipschitz continuous approximation to $\mathbb{1}_{D(\mathrm{s})}(\boldsymbol{\theta})$ with $D(\mathrm{s}) = (-\infty, s_1] \times (-\infty, s_2] \times \cdots \times (-\infty, s_M]$.

Coupling Strategy

We update the MLMC estimate iteratively, i.e., when computing $\hat{Y}_\ell(\mathrm{s})$ we have $\hat{F}_{\ell-1}(\mathrm{s})$.

1. Sample $\boldsymbol{\theta}_\ell^1, \ldots, \boldsymbol{\theta}_\ell^{n_\ell}$ using ABC with ϵ_ℓ;
2. Let $w_j^i = \frac{1}{n_\ell} \sum_{k=1}^{n_\ell} \mathbb{1}_{(-\infty,0]} \left(\theta_{\ell,j}^k - \theta_{\ell,j}^i\right)$;
3. Generate $\boldsymbol{\theta}_{\ell-1}^1, \ldots, \boldsymbol{\theta}_{\ell-1}^{n_\ell}$ where $\boldsymbol{\theta}_{\ell-1}^i = \left[\hat{F}_{\ell-1,1}^{-1}\left(w_1^i\right), \ldots, \hat{F}_{\ell-1,M}^{-1}\left(w_M^i\right)\right]$.

DISCLAIMER: This approach is an approximation, so technically the coupling does not satisfy the telescoping summation (aside from the univariate case).

A Tractable Example

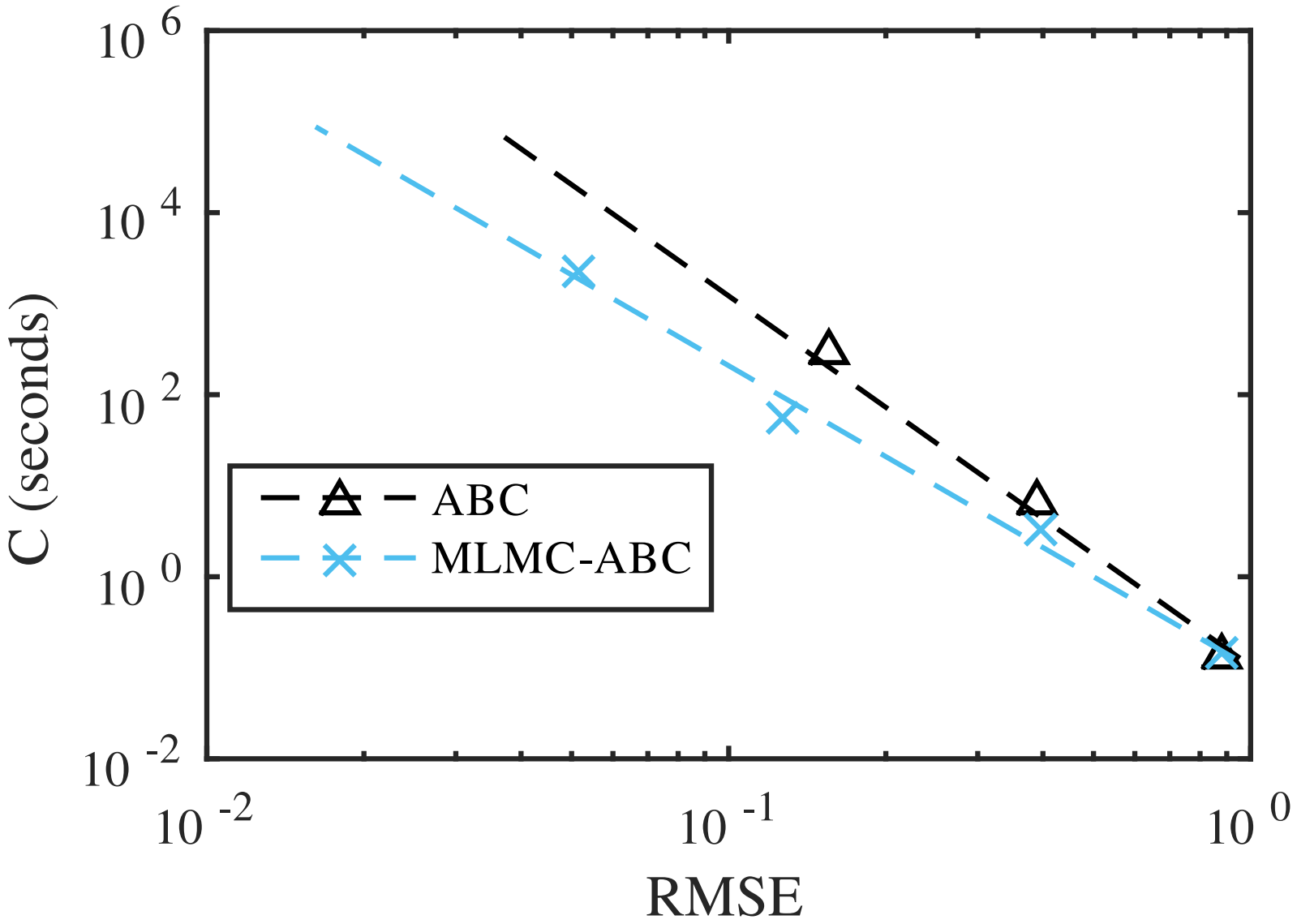
Susceptible-Infected-Susceptible Model

Spread of disease from infected population, I, to susceptible population, S, with no immunity.

$$I + S \xrightarrow{k_1} 2I, \quad I \xrightarrow{k_2} S$$

For $S_0 = 100$, $I_0 = 1$, the master equation can be computed exactly, so we can evaluate convergence in RMSE.

Numerical Results: Susceptible-Infected-Susceptible



Further Extensions for MLMC Inference

- Really active area of research;
- Various extensions to other sampling techniques:

  e.g. Markov chain Monte Carlo, sequential Monte Carlo, particle filters;
- Likelihood-base and likelihood-free context;
- Multifidelity Monte Carlo (a bit like the randomised MLMC idea);
- Various applications;
- My current work (Warne, Prescott, Baker, Simpson) combing both multilevel and multifidelity for ABC; This is supported by AustMS and CDS;

References

Giles, 2008. Multilevel Monte Carlo path simulation. *Operations Research* v56.

Anderson and Higham, 2012. Multilevel Monte Carlo for continuous time Markov chains, with applications in biochemical kinetics. *Multiscale Modeling and Simulation* v10

Rhee and Glynn, 2015. Unbiased estimation with square root convergence for SDE models. *Operations Research* v63

Lester, Baker, Giles, and Yates, 2016. Extending the multi-level method for the simulation of stochastic biological systems. *Bulletin of Mathematical Biology* v78

Warne, Baker, and Simpson, 2018. Multilevel rejection sampling for approximate Bayesian computation. *Computational Statistics & Data Analysis* v124

Warne, Baker, and Simpson, 2019. Simulation and inference algorithms for stochastic biochemical reaction networks: form basic concepts to state-of-the-art. *Journal of the Royal Society Interface* v16

Jasra, Jo, Nott, Shoemaker, and Tempone, 2019. Multilevel Monte Carlo in approximate Bayesian computation. *Stochastic Analysis and Applications* v37

Dodwell, Ketelsen, Scheichl, and Teckentrup, 2019. Multilevel Markov chain Monte Carlo. *SIAM Review* v61

Prescott and Baker, 2020. Multifidelity approximate Bayesian computation. *SIAM/ASA Journal of Uncertainty Quantification* v8

Thank You!