Parallelization for Markov chains Monte Carlo with heterogeneous runtimes

Stanislas du Ché[1], Charles Margossian[2]

8th June 2023

[1]ISERP Columbia University USA; CEREMADE – Dauphine University France; [2]Center for Computational Mathematics, Flatiron Institute USA, formally Department of Statistics Columbia University USA

Context and set up

ODE-Based Model

- Likelihood obtained by solving an ODE [1].
- Pharmacokinetics – Michaelis-Menten model [2].

$$\begin{cases} y_0' &= -k_a y_0 \\ y_1' &= k_a y_0 - \frac{V_m C}{K_m + C} \end{cases}$$

Where $C = \frac{y_1}{V}$

- Non-Linear.
- Simplified.
- Sensitive to parameter values.

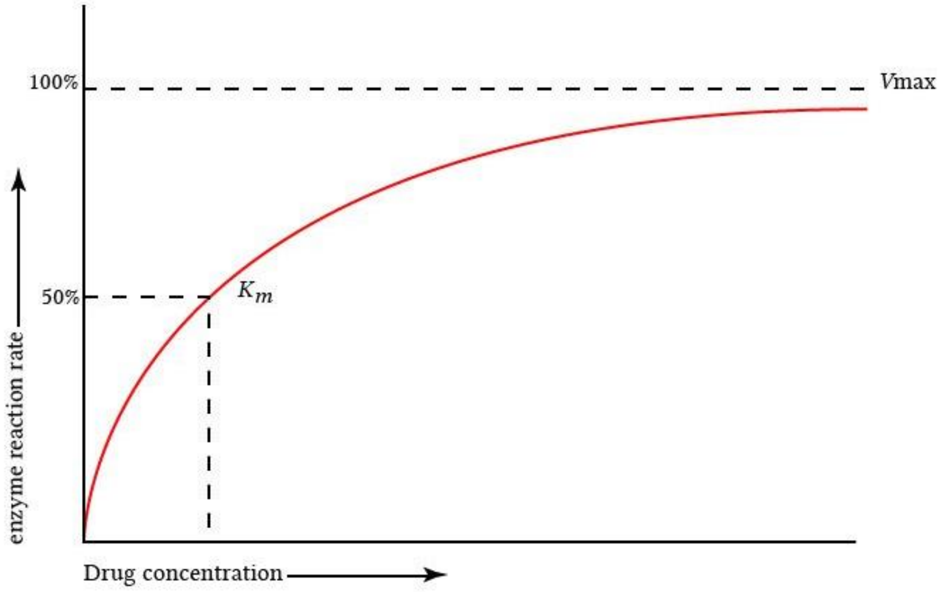


Figure 1: Michaelis-Menten Elimination Kinetics, relationship of concentration and clearance rate.

First Simulations

Experimental Setting

- Code in Stan.
- 8 chains on 4 cores.
- Wide priors (Naïve Computational Statistician).
- Runge-Kutta order 4/5 integrator (RK45).

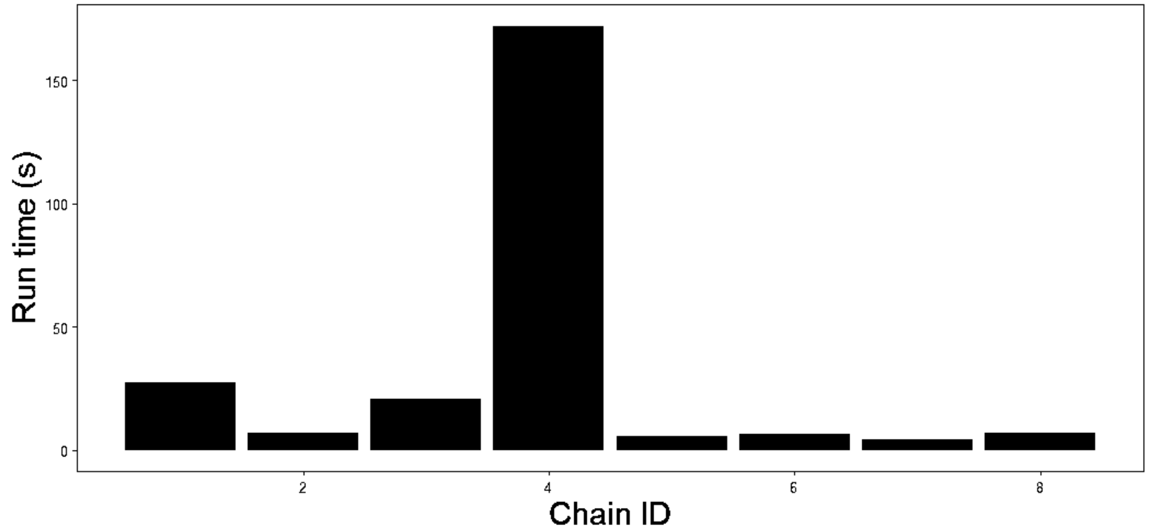


Figure 2: Running time of each chain ran.

Observations: Heterogeneous running time.
Solution: Let’s run more chains in parallel !

Second Simulation: More chains

Experimental Setting

- Code in Stan.
- 30 chains in parallel.
- Wide priors.
- RK45 integrator.

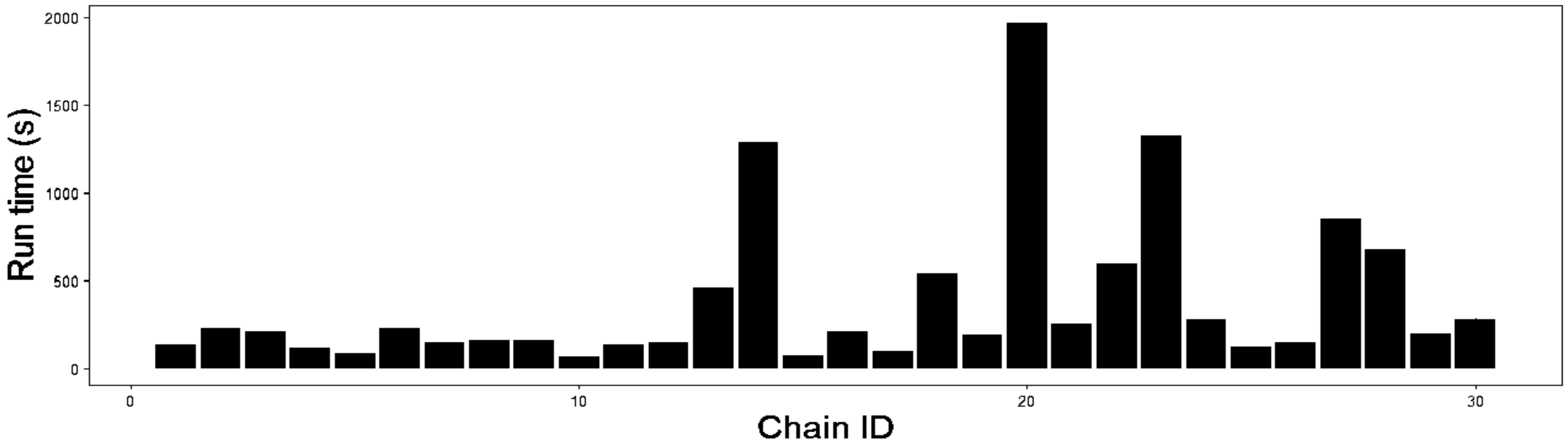


Figure 3: Heterogeneous computational time between chains.

- Wait for the slowest chains.
- 2000 sec !!

Second Simulation: More chains

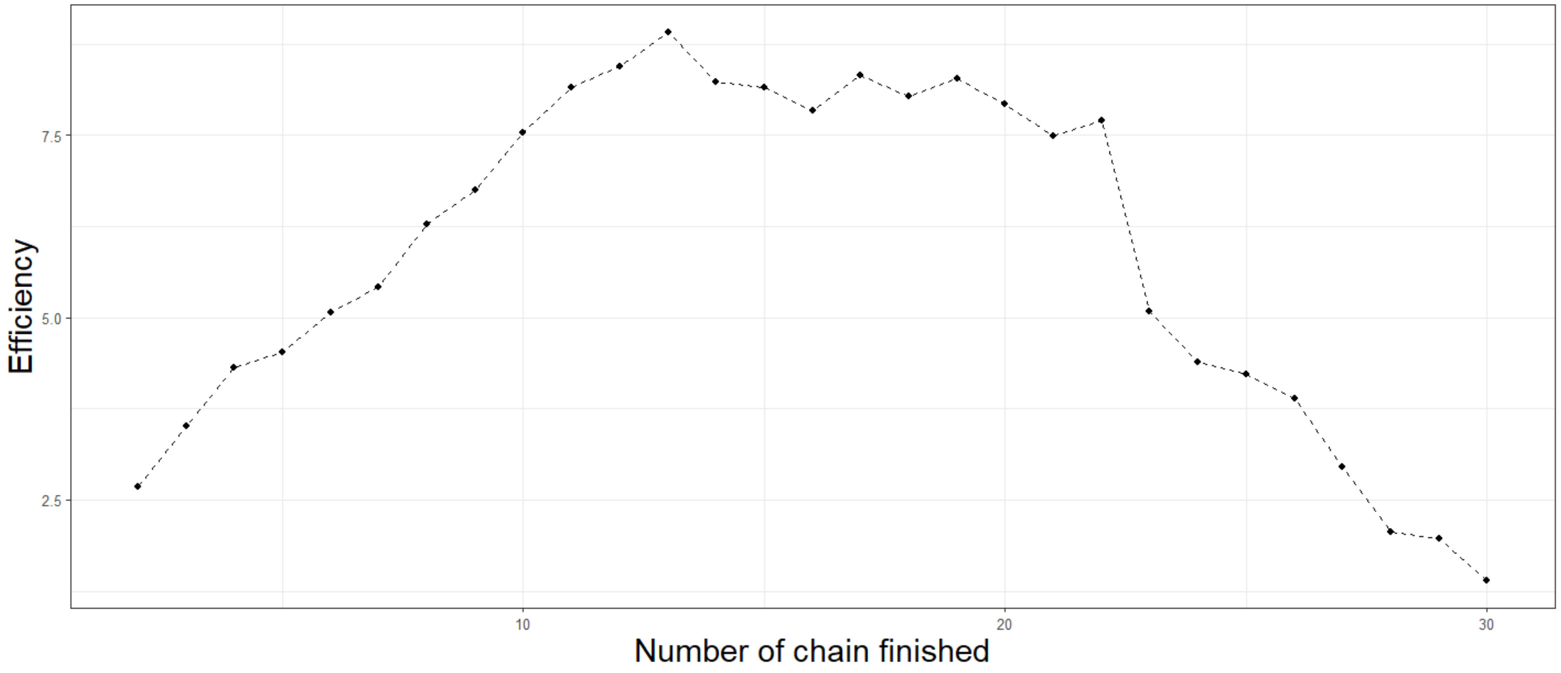


Figure 4: Efficiency (ESS/sec) by the number of chain finished.

- And it is **detrimental** !!

Reminder: Effective Sample Size (ESS)

$$ESS = \frac{N}{\sum_{t=-\infty}^{+\infty} \rho_t}$$

Understanding the Situation

Stiffness of an ODE: An ODE is considered stiff if for a “small” variation of its parameters, its behavior is very different.

Bad consequences for non-stiff ODE integrator (like RK45).

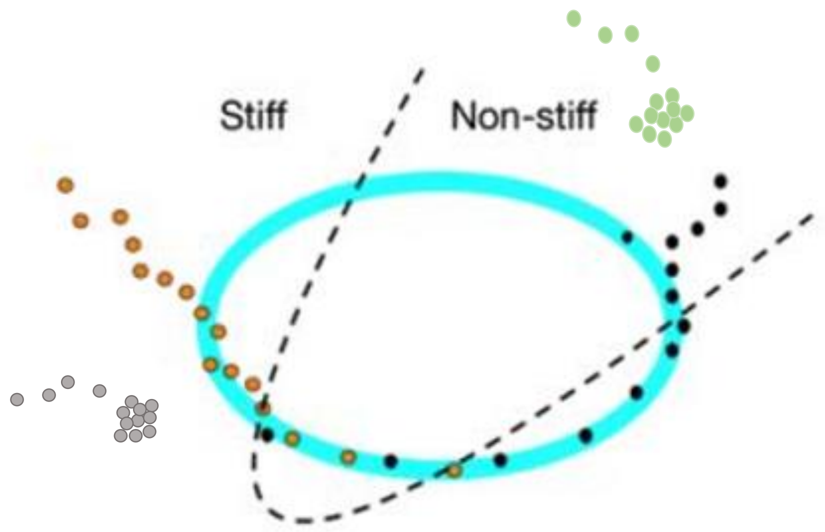
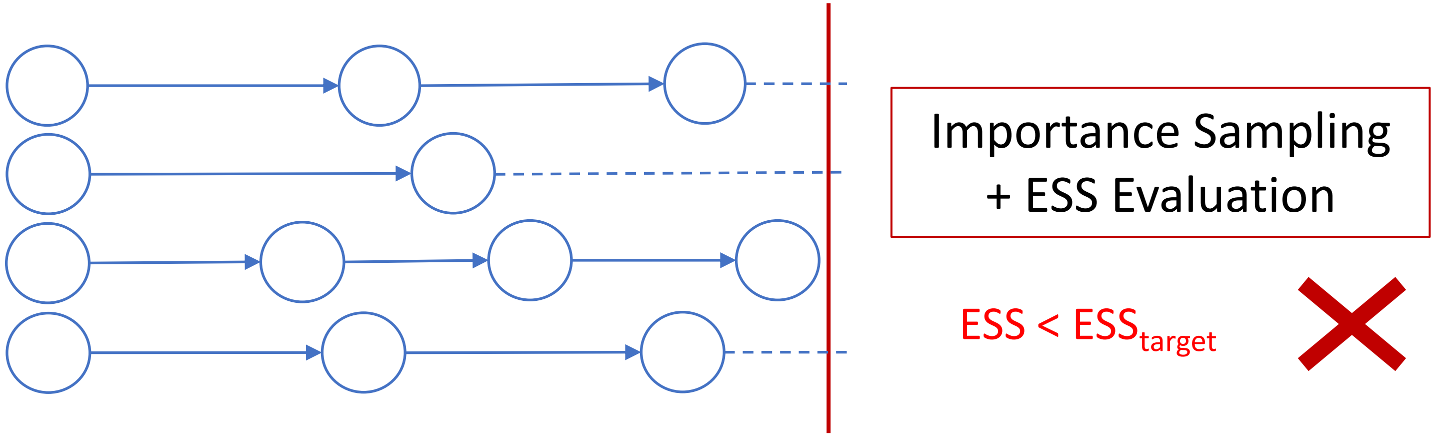


Figure 5: Intuition of the Parameter Space representation.

Chain Behaviors

- 4 possible scenarios for chain evolution in parameter space.
- So… what now ?

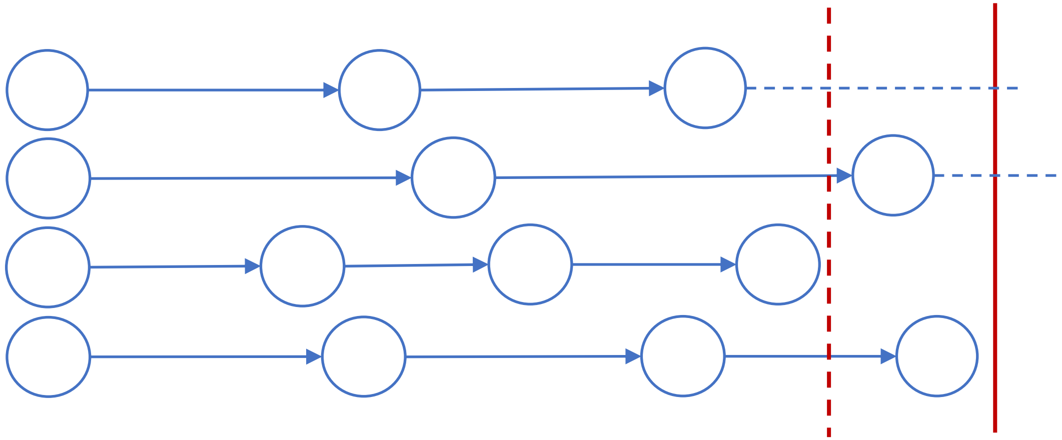
Waiting for the Fastest Chains – Experimental method



Importance Sampling

- Discard quick but unlikely samples.
- **Chain Stacking** by maximizing the leave-one-out log predictive density (loo$_{lpd}$) (Yao et al. 2022) [3].
- Stability of the loo$_{lpd}$ as convergence diagnostic.

Waiting for the Fastest Chains – Racing Chains Sampling



Importance Sampling
+ ESS Evaluation

$\mathrm{ESS} > \mathrm{ESS}_{\mathrm{target}}$

- **Keep** the finished chains.
- **Drop** the other ones.
- **Realize** Mixture Draw through Chain Stacking.

BIAS ?

Waiting for the Fastest Chains – Experimental Results

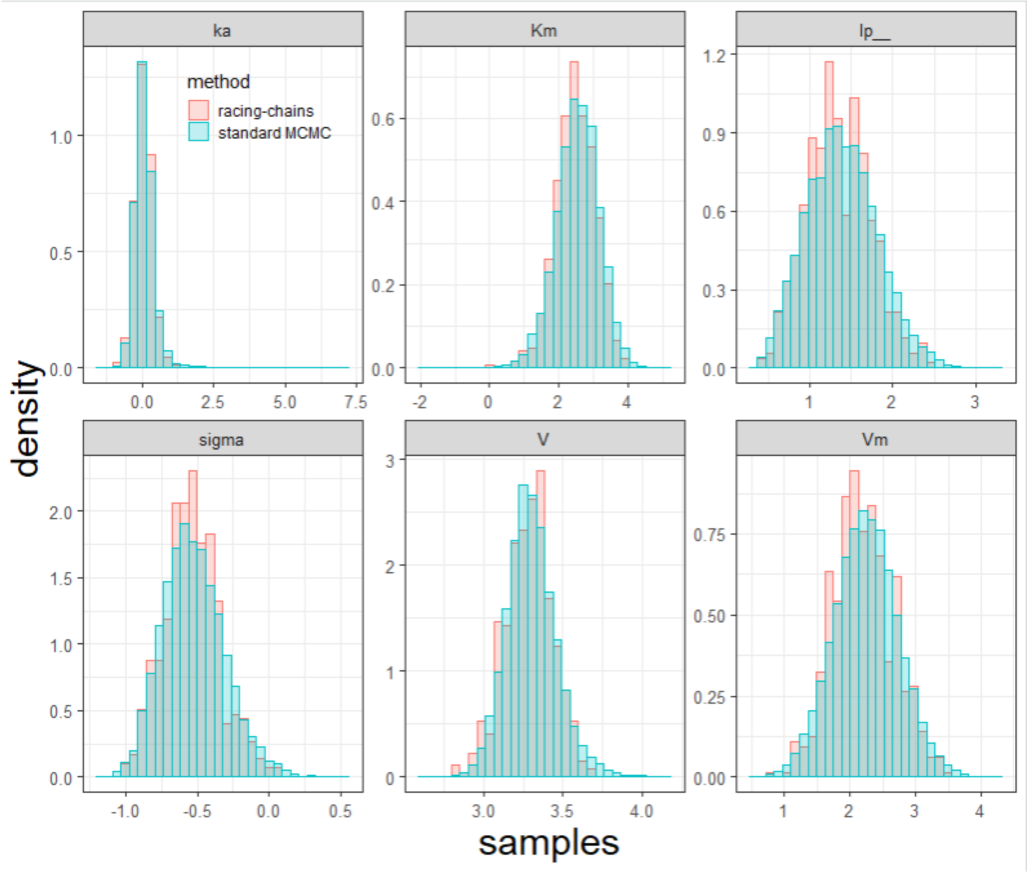
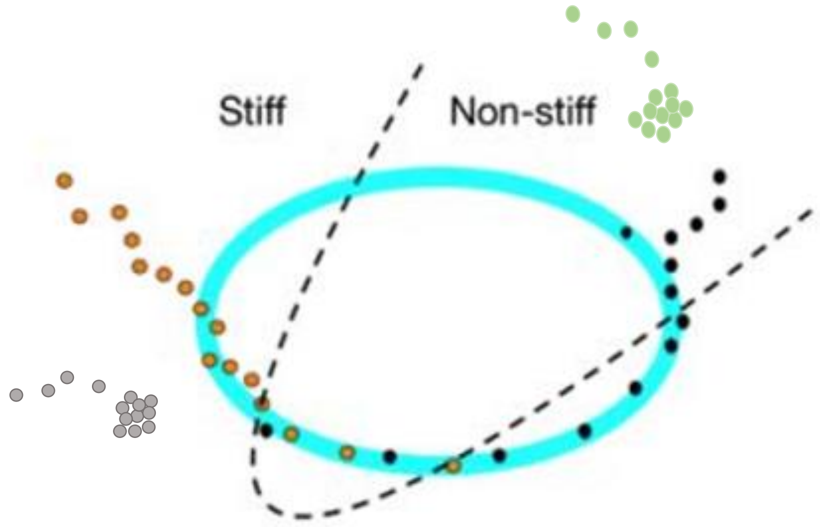


Figure 6: Posterior Distribution Racing Chain (9 Stacked chains) vs Standard MCMC (30 parallel chains).



Experimentally:

- No bias constated when dropping chains.
- Slowest Chains in Low Density regions. Planetary motion as other example [6].
- Stiff part of parameter space arguably sampled (Loo_{lpd} stability).

Waiting for the Fastest Chains – Experimental Results

	(a) Standard MCMC	**(b) Standard MCMC**	**(c) Racing-chains**	**(d) Racing-chains + stacking**
Chains	30	9	30 (9 kept)	9
Running time (s)	1918	775	**145**	**145**
Effective Sample Size	**2728**	1172	966	1107
Efficiency (ESS/s)	1.42	1.51	6.66	**7.63**

Observations:
- Gain of Computational Time.
- Smaller ESS but good enough (ESS_{target} = 1000).

Recap and Discussion

Recap:

- Launch more chains in parallel is **detrimental** in heterogeneous behavior situation.
- **Racing Chains + stacking** : heuristic method to gain computational time with reasonable risk.

Discussion:

- Bias introduced by dropping chains [5].
- **Wide** priors vs **thin** priors.
- Develop reliable convergence diagnostics; inspiration from unbiased Monte Carlo methods [7, 8, 9] ?

References

[1] Solving ODEs in a Bayesian context: challenges and opportunities. C. Margossian, L. Zhang, S. Weber and A. Gelman. Population Approach Group in Europe, 2021.

[2] Flexible and efficient Bayesian pharmacometrics modeling using Stan and Torsten, Part I. C. Margossian, Y. Zhang and W. Gillespie. *CPT: Pharmacometrics & Systems Pharmacology*, 2022.

[3] Stacking for non-mixing bayesian computations: The curse and blessing of multimodal posteriors. Yuling Yao, Aki Vehtari, and Andrew Gelman, 2020.

[4] Pareto Smoothed Importance Sampling. Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, Jonah Gabry,2022.

[5] Parallel computing and Monte Carlo algorithms. J. S. Rosenthal. Far East Journal of Theoretical Statistics, 4:207-236, 2000

[6] Bayesian model of planetary motion: exploring ideas for a modeling workflow when dealing with ordinary differential equations and multimodality. C. Margossian and A. Gelman. Stan Case Studies 7. 2020

[7] Annealed importance sampling. R. M. Neal. Statistics and Computing, 11:125 - 139, 2001

[8] Sequential Monte Carlo samplers. P. Del Moral, A. Doucet and A. Jasra. Journal of the Royal Statistical Society, Series B, 68:411-436, 2006

[9] Unbiased Markov chain Monte Carlo methods with couplings. P. E. Jacob, J. O'Leary and Y. F. Atchadé. Journal of the Royal Statistical Society, Series B, 82:543-600, 2020.