Paul Vanetti and **Arnaud Doucet** (*University of Oxford*)

ScaLE is an interesting Monte Carlo scheme preserving exactness under subsampling and is thus relevant to large data applications. Piecewise deterministic Markov chain Monte Carlo (MCMC) schemes also demonstrate similar features [BFR19, BCVD18]; however, these continuous-time algorithms can be difficult to understand and implement. A discrete-time alternative is the Scalable Metropolis-Hastings (SMH) algorithm [CVBC+19], a subsampling MCMC method whose invariant distribution is also the true posterior.

Like ScaLE, SMH makes use of control variate ideas, leveraging a Taylor series approximation to the log-posterior density at the MLE. Like Metropolis-Hastings (MH), SMH relies on a proposal q but uses the following acceptance probability

$$\alpha(\mathbf{x},\mathbf{x}') = \min\left\{1, \frac{\widehat{\pi}(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')}{\widehat{\pi}(\mathbf{x})q(\mathbf{x}'|\mathbf{x})}\right\}\prod_{i=1}^{n}\min\left\{1, \frac{f_i(\mathbf{x}')/\hat{f}_i(\mathbf{x}')}{f_i(\mathbf{x})/\hat{f}_i(\mathbf{x})}\right\},$$

where $f_i(\mathbf{x})$ is the likelihood from observation i, $\hat{f}_i(\mathbf{x})$ its control variate approximation and $\widehat{\pi}(\mathbf{x}) = f_0(\mathbf{x})\prod_{i=1}^{N}\hat{f}_i(\mathbf{x})$.

Naively simulating an event of probability $\alpha(\mathbf{x},\mathbf{x}')$ admits a $O(n)$ cost. However, as with ScaLE, this can be improved if the derivatives of the per-datum terms are uniformly bounded and if the posterior concentrates. When using a first-order Taylor approximation for $\widehat{\pi}$, SMH accesses $O(1)$ data at each iteration and $O(n^{-1/2})$ when using a second-order approximation [CVBC+19].

We tested SMH on the *airline* and *heterogeneous* datasets, using a second-order Taylor approximation centered at the MLE given in the text; see Table 1. We used independent proposals drawn from the Gaussian distribution corresponding to this Taylor approximation of the log-posterior.

For *airline*, SMH demonstrates a notable improvement over MH, visiting only 0.34 data on average per iteration. As for *heterogeneous*, it was necessary to use more finely-grained bounds than those described in [CVBC+19] by considering each of the d^3 partial derivatives; see [Van20] for details. Our implementation generated $\sim 3.0 \times 10^5$ effective samples per second and it visited only 0.015 data per iteration on average.

Although SMH outperforms significantly MH on these datasets, Monte Carlo methods are here of limited interest. Over 99.9% of the proposals sampled from our independent normal proposal distribution were accepted, suggesting that the posterior distribution is very close to its Laplace approximation in both cases.

Despite their current limitations, it is our hope that subsampling Monte Carlo methods such as ScaLE and SMH might form a basis for new algorithms to perform Bayesian inference in more challenging settings, such as random effect models.

Dataset	Size (n)	Algorithm	Data/sample	ESS/s
Airline	1.21×10^8	MH	1.21×10^8	0.28
		SMH	0.34	4.7×10^4
Heterogeneous	10^7	MH	10^7	3.3
		SMH	0.015	3.0×10^5

Table 1: Performance of MH and SMH in terms of average number of data points accessed per iteration (Data/sample) and Effective Sample Size per second (ESS/s).

References

[BCVD18] Alexandre Bouchard-Côté, Sebastian Vollmer, and Arnaud Doucet. The bouncy particle sampler: A nonreversible rejection-free Markov chain Monte Carlo method. *Journal of the American Statistical Association*, 113(522):855–867, 2018.

[BFR19] Joris Bierkens, Paul Fearnhead, and Gareth Roberts. The zig-zag process and super-efficient sampling for Bayesian analysis of big data. *The Annals of Statistics*, 47(3):1288–1320, 2019.

[CVBC+19] Rob Cornish, Paul Vanetti, Alexandre Bouchard-Côté, George Deligiannidis, and Arnaud Doucet. Scalable Metropolis–Hastings for exact Bayesian inference with large datasets. In *International Conference on Machine Learning*, pages 1351–1360, 2019.

[Van20] Paul Vanetti. *Piecewise-deterministic Markov chain Monte Carlo*. PhD thesis, University of Oxford, 2020.