Consistency of Markov chain quasi Monte Carlo with multiple proposals

Tobias Schwedes
Department of Mathematics
Imperial College London
London, UK
t.schwedes14@imperial.ac.uk

Ben Calderhead
Department of Mathematics
Imperial College London
London, UK
b.calderhead@imperial.ac.uk

Abstract

In [1] a Markov chain Monte Carlo (MCMC) method was introduced that allows for parallelisation by proposing multiple samples in each iteration, and that has as stationary distribution the correct target density. Here, we prove that this property still holds true when the driving sequence of pseudo-random numbers is replaced by completely uniformly distributed (c.u.d.) numbers. For the proof a coupling argument similar to [3] is applied. Initial numerical simulations confirm our theoretical result, and suggest a scaling of order n^{-1} as we increase parallelisation instead of the usual $n^{-1/2}$.

1 Introduction

In many probabilistic problems in science, MCMC is applied due to its ability to sample from arbitrary distributions known up to a constant. Comprehensive introductions on MCMC methods can be found in [17], [16], [21] and [12]. Estimators resulting from MCMC scale independently of dimensionality, however they have the fairly slow universal convergence rate of $n^{-1/2}$, similar to classic Monte Carlo methods. For the latter, convergence rates can be achieved of order close to n^{-1} when samples are generated by a suitable low-discrepancy sequence, i.e. points which are highly homogeneously distributed in space, instead of the usual pseudo-random numbers ([7]). These so called quasi-Monte Carlo (QMC) methods, despite their generally deteriorating performance with increasing (effective) dimension [25], can therefore lead to significant computational savings compared to standard Monte Carlo.

The first applications of QMC in the context of MCMC go back to [5] and [22], which assume a discrete state space. In [5], the driving sequence of uniformly distributed independent and identically distributed (i.i.d.) random numbers is replaced by a c.u.d. sequence. The same approach is used in [19], [3] and in this work. In [15], a Gibbs sampler that runs on randomly shuffled QMC points is introduced. Later, [2] uses a weighting of rejected samples to generate balanced proposals. Both successfully applied QMC in MCMC, however without providing any theoretical support. [6] uses QMC in multiple-try Metropolis-Hastings (M-H), and [14] within an exact sampling method introduced by [20]. In [13] the so called array-randomised QMC (RQMC) was introduced, which uses quasi-Monte Carlo to update multiple chains that run in parallel. Further, the roter-router model, which is a deterministic analogue to a random walk on a graph, was applied in [8] on a number of problems.

Based on the coupling argument by Chentsov from [5], it has been proven in [19] that an MCMC method defined on a finite state space still has the correct target as its stationary distribution when the driving sequence of i.i.d. numbers is replaced by weakly c.u.d. (w.c.u.d.) numbers. Following this, [24] proves theoretical properties of w.c.u.d. sequences. They also run a Gibbs sampler by w.c.u.d. numbers and achieve significant performance improvements compared to using i.i.d. inputs. Recently, the result from [19] was generalised to c.u.d. numbers and continuous state spaces by Chen ([3]).

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

In this work, we generalise Chen's result to an MCMC method that uses multiple proposals in each iteration, introduced in [1]. Due to the fact that the state space is covered evenly by proposals in each iteration using c.u.d. numbers, which thus allows the sampling mechanism a larger spectrum of moves, one might expect that using c.u.d. numbers in multiple proposal MCMC should harvest the low-discrepancy benefits more effectively than in the single proposal case. Here, we perform numerical simulations that suggest a scaling of n^{-1}, which is known from QMC methods.

This work is organised as follows. In Section 2 the multiple proposal MCMC from [1] is presented. Section 3 introduces the concept and some constructions of c.u.d. numbers from the literature. Regularity conditions are given in Section 4, which are necessary for the consistency theorem formulated in Section 5. In Section 6, numerical experiments for multiple proposal MCMC driven by c.u.d. numbers are presented for a simple estimation problem.

2 MCMC with multiple proposals

In [1], a natural generalisation of the well-known M-H algorithm ([10]) was suggested that allows for parallelising a single chain by proposing multiple points in parallel. In every MCMC iteration, a finite state Markov chain on the proposed points is constructed and samples are drawn, in such a way that the overall procedure has the correct target density π as its stationary distribution. Mathematically, this approach corresponds to a M-H algorithm over a product space. In order to see this note first that any joint probability distribution $p(x_{1:N+1})$ can be factorised in $N+1$ different ways, using conditional probabilities, of the form, $p(x_{1:N+1}) = p(x_i)p(x_{\setminus i}|x_i)$, where we use the notation $x_{\setminus i} = x_{[1:i-1,i+1:N+1]}$. Let us set $p(x_i) = \pi(x_i)$ and $p(x_{\setminus i}|x_i) = \kappa(x_i, x_{\setminus i})$ for a proposal kernel κ. Hence, in the ith factorisation, $x_i \in \Omega \subset \mathbb{R}^d$ is distributed according to the target density while the other $x_{\setminus i}$ are distributed according to the proposal kernel conditioned on x_i. Referring to [1, 23], a discrete auxiliary and uniformly distributed random variable $I \in \{1, \dots, N+1\}$ is introduced that determines which factorisation of the joint probability is used, i.e. $p(x_{1:N+1}, I = i) = \frac{1}{N+1}\pi(x_i)\kappa(x_i, x_{\setminus i})$. The multiple proposal M-H can be considered as a single Markov chain over the product space of variables $(x_{1:N+1}, I)$ by applying a combination of two transition kernels, each of which preserves the underlying joint stationary distribution. First, the states of a finite state Markov chain are created by updating the variables $x_{\setminus i}$ conditioned on x_i and $I = i$. Secondly, we sample I conditioned on $x_{1:N+1}$ using the stationary distribution of $I|x_{1:N+1}$. When $I = i$, we observe that x_i is allocated the target density $\pi(x_i)$. These samples will be the ones collected in every iteration. In practice, generating proposals $x_{\setminus I}$ from $\kappa(x_I, \cdot)$ is typically performed via applying its generalised inverse Ψ_{x_I} to a set of pseudo-random numbers $u_1, ..., u_{Nd} \in [0, 1]$ such that $x_{\setminus I} = \Psi_{x_I}(u_1, ..., u_{Nd})$. This procedure is described in Algorithm 1.

Algorithm 1: Multiple proposal Metropolis-Hastings

Input: Initialise starting point $\tilde{x}_1 \in \Omega \subset \mathbb{R}^d$, number of proposals N, auxiliary variable $I = 1$ and counter $n = 0$, generate an pseudo-random sequence $u_1, u_2, ... \in (0, 1]$;

for each MCMC iteration $i = 1, 2, ...$ **do**
- Set $u^i = (u_{(i-1)Nd+1}, ..., u_{iNd}) \in (0, 1]^{Nd}$, and sample $\tilde{x}_{\setminus I}$ conditioned on I, i.e., draw N new points from $\kappa(\tilde{x}_I, \cdot) = p(\tilde{x}_{\setminus I}|\tilde{x}_I)$ by the inverse $\Psi_{\tilde{x}_I}(u^i)$;
- Compute the stationary distribution of $I|\tilde{x}_{1:N+1}$, i.e. $p(I = j|\tilde{x}_{1:N+1}) = \pi(\tilde{x}_j)\kappa(\tilde{x}_j, \tilde{x}_{\setminus j}) / \sum_k \pi(\tilde{x}_k)\kappa(\tilde{x}_k, \tilde{x}_{\setminus k})$, which can be done in parallel;
- **for** $m = 1, ..., N$ **do**
  - Sample new I via $p(\cdot|\tilde{x}_{1:N+1})$;
  - Set new sample $x_{n+m} = \tilde{x}_I$;
- **end**
- Update counter $n = n + N$

end

The detailed balance condition for Algorithm 1 can be formulated as

$$p(I = i)\pi(x_i)\kappa(x_i, x_{\setminus i})p(I = j|x_{[1:N+1]}) = p(I = j)\pi(x_j)\kappa(x_j, x_{\setminus j})p(I = i|x_{[1:N+1]}), \quad (1)$$

which is easily verified to hold. Assuming that the chain is also ergodic, the sequence $(x_i)_i$ of the multiple proposal M-H consistently samples π, that is, the empirical distribution function $F_n(x) :=$

$\frac{1}{n}\sum_{m=1}^{n} \mathbb{1}_{[x_m,\infty)}(x)$ satisfies $F_n(x) \to F(x)$ as $n \to \infty$ for any continuity point x of F, where $F(x) := \int_{-\infty}^{x} \pi(y)\mathrm{d}y$. Referring to [1], multiple proposal M-H has two main advantages in efficiency compared to a single proposal M-H approach: First, the likelihoods of multiple proposals can be computed in parallel. Secondly, the acceptance rate for proposals is increased. Our idea is to increase efficiency further by replacing the pseudo-random numbers in Algorithm 1 by more evenly distributed numbers, for which consistency can be proven, and thereby exploring the state space more thoroughly.

3 Completely uniformly distributed numbers

Completely uniformly distributed numbers belong to the class of low-discrepancy sequences and were introduced in [11]. The star discrepancy D_n^{*d} for a set of points $P = \{x_1, ..., x_n\} \subset (0,1]^d$ is defined by

$$D_n^*(P) = \sup_{b\in(0,1]^d} \left| \frac{\#\{x_i \in (0,b]\}}{n} - \mathrm{vol}\left((0,b]\right) \right|. \tag{2}$$

A sequence $(u_i)_{i\geq 1} \subset (0,1]$ is called c.u.d. if for any $d \geq 1$ the points $x_i^{(d)} = (u_i, \ldots, u_{i+d-1}) \in (0,1]^d$ fulfill $D_n^{*d}(x_1^{(d)}, \ldots, x_n^{(d)}) \to 0$ as $n \to \infty$. In other words, any sequence of overlapping blocks of u_i of size d yield the desirable uniformity property $D_n^{*d} \to 0$.

There are a number of techniques to construct c.u.d. sequences in the literature. In [19] a c.u.d. sequence is used that is based on a linear congruential generator developed in [9]. The lattice construction from [18] and the shuffling strategy for QMC points from [15] are both shown to produce c.u.d. points in [24]. Further, [4] presents constructions of c.u.d. points based on fully equidistributed LFSR, and antithetic and round trip sampling.

4 Regularity conditions

Similarly to [3], the consistency proof given below relies on two regularity conditions. The first one defines coupling properties of the sampling method, and the second one suitable integrability over the sample space.

Coupling: For any MCMC iteration in Algorithm 1, let $\phi(x_I, (u_1, ..., u_{Nd})) = (x_1, ..., x_N)$ denote the innovation operator that produces the N new samples. Let $\mathcal{C} \subset (0,1]^{Nd}$ have positive Jordan measure. If for any $u \in \mathcal{C}$ it holds $\phi(x,u) = \phi(x',u)\ \forall\, x, x' \in \Omega$, then $\mathcal{C}$ is called a *coupling region*.

Integrability: For $k \geq 1$, let $x_k = x_j^i = x_j^i(u^1, ..., u^i)$ with $i \geq 1, 1 \leq j \leq N$ and $k = (i-1)N + j$ denote the kth N-proposal M-H update, i.e. the jth sample in the ith iteration, according to Algorithm 1. The method is called regular if $g : [0,1]^{iNd} \to \mathbb{R}$, defined by $g(u^1, \ldots, u^i) = f(x_k(u^1, \ldots, u^i))$, where $f : \mathbb{R}^d \to \mathbb{R}$ is bounded and continuous, is Riemann integrable.

5 Consistency

The following theorem states that if the driving sequence $(u_i)_{i\geq 1}$ of pseudo-random numbers for the multiple proposal M-H is replaced by a sequence of non-overlapping tuples of c.u.d. numbers, then consistency still holds true for the resulting algorithm.

Theorem. Let $x_0 \in \Omega$. For $k \geq 1$, let $x_k = x_j^i$ with $i \geq 1, 1 \leq j \leq N$ and $k = (i-1)N + j$ denote, as before, the kth N-proposal M-H update, from Algorithm 1, which is assumed to be irreducible and Harris recurrent with stationary distribution π. It is also assumed to be regular and to have a coupling region $\mathcal{C}$. Further, let $u^i = (v_{iNd+1}, ..., v_{(i+1)Nd})$ for a c.u.d. sequence $(v_i)_{i\geq 1}$. Then, $(x_k)_{k\geq 1}$ consistently samples π.

For the proof, we use a coupling argument similar to [3] to guarantee ergodicity of the sequence of multiple proposal M-H samples, and show consistency in terms of $1/n \sum_k f(x_k) \to \int f(x)\pi(x)\mathrm{d}x$ for bounded continuous f. The proof further relies on a technical lemma, stating that overlapping dk-tuples of c.u.d. numbers, with successive tuples overlapping on d entries, are uniformly distributed. A similar argument is used in the consistency proofs in [5], [19] and [3].

The fairly strong assumption of the existence of a coupling region suggests finding appropriate classes of proposal and acceptance mechanisms where this holds true might be hard. A non-trivial example uses independent proposal sampling under soft boundedness conditions.

6 Numerical simulations

Here, we analyse the convergence properties of Algorithm 1 driven by c.u.d. numbers numerically on a simple example problem: Let the target π be given by the standard Normal distribution on $\mathbb{R}^d$. The construction of the c.u.d. sequence used here was introduced in [4] and relies on LFSR with a transition mechanism based on primitive polynomials over the Galois field $GF(2)$. In a first experiment, proposals are independently generated by inverting the kernel $\kappa(x_{1:N}) = \mathcal{N}(0, \varepsilon^2 \, \mathrm{I}_{Nd})$ for $\varepsilon > 0$, and the posterior mean is estimated by the arithmetic mean. This method is consistent due to the consistency theorem. Further, Figure 1 suggest that the standard deviation of this estimator scales with order n^{-1} as we increase the proposed number of samples in each MCMC iteration instead of the usual $n^{-1/2}$.

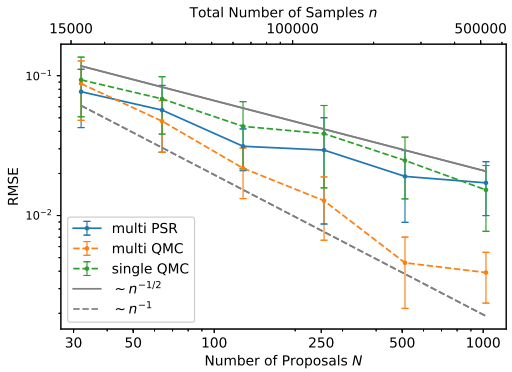

Figure 1: RMSE of the sample mean based on M-H with independent proposals using pseudo-random (PSR) vs. c.u.d. (QMC) numbers as driving sequences, resp., for increasing proposal numbers and sample sizes. The results are based on 25 MCMC simulations.

In a second experiment, a random walk proposal kernel $\kappa(x_I, \cdot) = \mathcal{N}(x_I, \varepsilon^2 \, \mathrm{I}_{Nd})$ is applied. Although not analytically proven, our simulations again suggest consistency in this case: Figure 2 presents the convergence of the empirical distribution function in terms of the $\|\cdot\|_\infty$-norm for a range of different proposal numbers. Since uniform convergence implies pointwise convergence, which is equivalent to consistency for continuous distributions, the results suggest that consistent samples are generated.

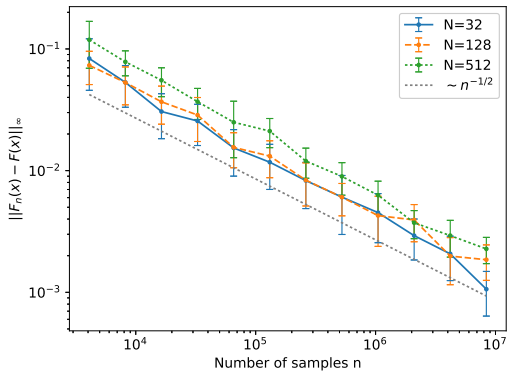

Figure 2: Maximum-norm error for the empirical distribution function of samples from the N proposal M-H with random walk proposals, using c.u.d. numbers as driving sequence, for increasing sample sizes. The results are based on 25 MCMC simulations.

References

[1] Ben Calderhead. A general construction for parallelizing Metropolis- Hastings algorithms. *Proceedings of the National Academy of Sciences*, 111(49):17408–17413, 2014.

[2] Suneal K Chaudhary. *Acceleration of Monte Carlo methods using low discrepancy sequences*. PhD thesis, UCLA, 2004.

[3] S Chen, Josef Dick, and Art B Owen. Consistency of Markov chain quasi-Monte Carlo on continuous state spaces. *The Annals of Statistics*, pages 673–701, 2011.

[4] Su Chen, Makoto Matsumoto, Takuji Nishimura, and Art B Owen. New inputs and methods for Markov chain quasi-Monte Carlo. In *Monte Carlo and quasi-Monte Carlo methods 2010*, pages 313–327. Springer, 2012.

[5] NN Chentsov. Pseudorandom numbers for modelling markov chains. *USSR Computational Mathematics and Mathematical Physics*, 7(3):218–233, 1967.

[6] Radu V Craiu and Christiane Lemieux. Acceleration of the multiple-try Metropolis algorithm using antithetic and stratified sampling. *Statistics and computing*, 17(2):109, 2007.

[7] Josef Dick, Frances Y Kuo, and Ian H Sloan. High-dimensional integration: the quasi-Monte Carlo way. *Acta Numerica*, 22:133–288, 2013.

[8] Benjamin Doerr and Tobias Friedrich. Deterministic random walks on the two-dimensional grid. *Combinatorics, Probability and Computing*, 18(1-2):123–144, 2009.

[9] Karl Entacher, Peter Hellekalek, and Pierre L'Ecuyer. Quasi-Monte Carlo node sets from linear congruential generators. *Monte Carlo and Quasi-Monte Carlo Methods*, pages 188–198, 1998.

[10] W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[11] Nikolai Mikhailovich Korobov. Concerning some questions of uniform distribution. *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, 14(3):215–238, 1950.

[12] David P Landau and Kurt Binder. *A guide to Monte Carlo simulations in statistical physics*. Cambridge university press, 2014.

[13] Pierre L'Ecuyer, Christian Lécot, and Bruno Tuffin. A randomized quasi-Monte Carlo simulation method for Markov chains. *Operations Research*, 56(4):958–975, 2008.

[14] Christiane Lemieux and Paul Sidorsky. Exact sampling with highly uniform point sets. *Mathematical and Computer Modelling*, 43(3):339–349, 2006.

[15] JG Liao. Variance reduction in Gibbs sampler using quasi random numbers. *Journal of Computational and Graphical Statistics*, 7(3):253–266, 1998.

[16] Jun S Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

[17] MEJ Newman and GT Barkema. *Monte Carlo Methods in Statistical Physics chapter 1-4*. Oxford University Press: New York, USA, 1999.

[18] Harald Niederreiter. Pseudo-random numbers and optimal coefficients. *Advances in Mathematics*, 26(2):99–181, 1977.

[19] Art B Owen and Seth D Tribble. A quasi-monte carlo metropolis algorithm. *Proceedings of the National Academy of Sciences of the United States of America*, 102(25):8844–8849, 2005.

[20] James Gary Propp and David Bruce Wilson. Exact sampling with coupled markov chains and applications to statistical mechanics. *Random structures and Algorithms*, 9(1-2):223–252, 1996.

[21] Christian Robert and George Casella. *Monte Carlo Statistical Methods*. Springer-Verlag New York, 2004.

[22] Il'ya Meerovich Sobol'. Pseudo-random numbers for constructing discrete Markov chains by the Monte Carlo method. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 14(1):36–44, 1974.

[23] Hakon Tjelmeland. Using all Metropolis–Hastings proposals to estimate mean values. *Statistics*, 4, 2004.

[24] Seth D Tribble, Art B Owen, et al. Construction of weakly CUD sequences for MCMC sampling. *Electronic Journal of Statistics*, 2:634–660, 2008.

[25] Xiaoqun Wang and Kai-Tai Fang. The effective dimension and quasi-Monte Carlo integration. *Journal of Complexity*, 19(2):101–124, 2003.