Potentials and limitations to speed up MCMC methods using non-reversible chains

Balázs Gerencsér

***Abstract*— Mixing time is the quantity to measure the speed of MCMC sampling. We compare the cases of using reversible chains, which are better understood with non-reversible chains, which offer more degree of freedom. It turns out that non-reversible chains can provide significant speedup in some cases but no improvement in others.**

I. Introduction

In both the theory and practice of Markov Chain Monte Carlo (MCMC) methods it is vital to know how fast the underlying Markov chain reaches its stationary distribution. Distributed averaging methods can also be interpreted as Markov chains reaching the uniform distribution provided that this is the stationary distribution. See [1] or [2] for details. Our scope is limited to discrete time, finite state space chains, with a unique stationary distribution π. The speed of approaching the stationary distribution can be quantified as the *mixing time*, defined as follows:

$$t_{\mathrm{mix}}(\varepsilon) = \max_{\sigma\in\mathcal{P}(\Omega)} \min\left\{k : \|\sigma^{(k)} - \pi\|_{\mathrm{TV}} \le \varepsilon\right\}.$$

Here $\mathcal{P}(\Omega)$ is the set of probability distributions on the state space Ω, $\sigma^{(k)}$ is the distribution after k step with initial distribution σ, π is the stationary distribution of the chain, and $\|.\|_{\mathrm{TV}}$ is the total variation norm defined as:

$$\|\mu\|_{\mathrm{TV}} = \max_{A\subset\Omega} |\mu(A)|.$$

We should note that there are other similar quantities used in the literature to quantify the speed of reaching the stationary distribution, for example one can replace the TV norm with the l_2 or l_∞ norm.

A Markov chain is reversible if starting from the stationary distribution, the probability of the consecutive pair (i, j) is the same as the probability of the consecutive pair (j, i). Formally:

$$\pi_i p_{ij} = \pi_j p_{ji} \quad \forall i, j,$$

where p_{ij} is the transition probability from i to j.

An important feature of a Markov chain is the connectivity graph. This is defined with the states as nodes and edges (i, j) if either $p_{ij} > 0$ or $p_{ji} > 0$. We often omit self loops from this graph. As we will see the connectivity graph may impose some restrictions (lower bounds) on achievable mixing times.

Balázs Gerencsér is with the Department of Probability Theory and Statistics, Eötvös Loránd University, 1117 Budapest, Hungary gebaboy@cs.elte.hu

This work has been carried out while visiting the Massachusetts Institute of Technology, Laboratory for Information and Decision Systems. Supported by the Hungarian-American Fulbright Commission.

Reversible Markov chains play a prominent role in MCMC theory. Allowing non-reversible chains considerably enhances our flexibility, and the question arises, as to what improvement can be achieved. The present paper is describing a research in progress in this direction. First we provide a brief survey of existing results. Then we show for the first time that decreasing the mixing time by allowing non-reversible chains may be impossible assuming certain connectivity graphs.

II. Efficient tools for reversible chains

Depending on what type of information can we easily extract from the Markov chain, there are several techniques to get estimates for the mixing time.

A classic algebraic tool is using the eigenvalue structure of the transition matrix. It is known that the mixing time can be estimated using the *spectral gap*, which is

$$1 - \max_{i\ge 2} |\lambda_i|,$$

where λ_i are the eigenvalues of the transition matrix, $\lambda_1 = 1$. For fixed ε, the mixing time can be estimated by the inverse of the spectral gap, up to a factor of $\log \min_i \pi_i$. One might get tighter bounds by an alternative approach using log-Sobolev constants (see [3]). This has a smaller $\log\log \min_i \pi_i$ factor between the lower and upper bound. However, this method works only for the l_2 mixing time, and the log-Sobolev constant can be substantially harder to compute than the spectral gap (see [4]).

Another class of tools are using properties of the underlying graph. An important example is the concept of *conductance* (see [5], [6]). The idea is to find the worst decomposition of the graph into two parts so that the chain is rarely going from one part to the other. This can then be used to bound the overall mixing on the graph. The lower and upper bounds we can get here differ by a square factor.

Besides estimating mixing time for a single chain, quite some work has been done to find the fastest mixing reversible chain given the connectivity graph, see [7] and [8].

All these results depend on reversibility in some way. Some of the claims simply do not hold for non-reversible chains. In some other cases the result we can get is less sharp in the general setting. Finally, it also happens that a tool is applicable in the general case, but it is hard to compute the upper or lower bounding quantity.

ISBN 978-963-311-370-7

For example, using the spectral gap we get a much weaker upper bound for the mixing time for non-reversible chains. Moreover, when moving away from reversible chains we lose the orthonormal structure of the eigenvectors of the transition matrix. Thus we may find ourselves in a more difficult situation when trying to determine the spectral gap.

The reversibility assumption is often technical necessity but not a constraint explicitly imposed by the motivating application. A detailed and complete survey on these techniques can be found in [9] or [10].

III. Significant speedup by non-reversible chains

Non-reversible Markov chains are much harder to work with. However, it turns out that in some settings they can be substantially faster than reversible chains. One of the key ingredient here, the concept of *lifting* was introduced by Diaconis, Holmes and Neal in [11], extended by Chen, Lovász and Pak (see [12]), and optimized by Gade and Overton (see [13]). The idea is to split each state into two or more new states and determine transition probabilities in an appropriate way so that the marginal of the new chain will behave like the original chain. The additional structure allows more flexibility and faster mixing, it can decrease mixing time up to the square root of the original.

Let us mention Example 6.6. in [10]. Here the mixing time of a non-reversible chain is the square root of the mixing time of a similar reversible one. We need to add that the mixing time of the reversible chain is nearly optimal for the specific connectivity graph. In this example we could achieve speedup without changing the connectivity graph. This is the type of improvement we are aiming for.

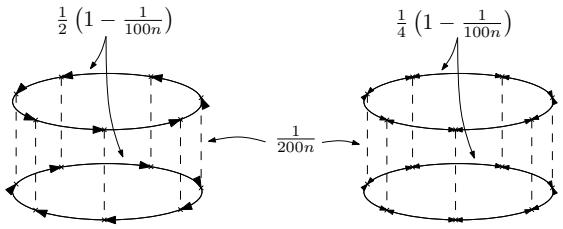

Fig. 1. Example 6.6 of [10], two chains on the double cycle

IV. Speedup is not always possible

In certain cases, one might not want to change the connectivity graph, but still try to use a non-reversible chain to improve mixing performance. Our result shows that this is impossible for certain graphs. We restrict ourselves to the uniform sampling problem, that is, to chains with uniform stationary distribution.

Theorem 1: Consider a Markov chain on a cycle with n nodes. Suppose the stationary distribution is uniform. Then, for some global constant $C > 0$,

$$t_{\mathrm{mix}}(1/8) \geq Cn^2.$$

Without including a complete proof, let us highlight a few key points of it.

First we observe that any chain considered in the theorem can be obtained from a reversible chain by adding some "rotation". This means increasing all, say, clockwise transition probabilities and decreasing counterclockwise ones by the same amount.

We try to emulate the evolution of the Markov chain probabilities by a geometric rotation defined in the following way. Let us start from a function f on the unit circle. Fix some *observation points* $Z_1, Z_2, \ldots, Z_n$ on the circle, and define

$$y_i^0 = f(Z_i).$$

And also

$$y_i^\alpha = f^\alpha(Z_i),$$

where f^α equals to f rotated by an angle α. Now we want the rotation noted for the transition probabilities relate to the geometric rotation just defined. We can formulate our goal as to ensure that

$$y^\alpha P = y^{\alpha+\varphi} + d^\alpha,$$

with φ fixed and for all α, while keeping d^α as small as possible. In order to achieve this we have to choose $f, \varphi, Z_i, d^\alpha$ carefully.

It is relatively hard to track the effect of an iterated matrix multiplication. On the other hand, it is very easy to tell what happens after multiple geometric rotations. We can switch our focus from the Markov chain to a series of rotated vectors y^α if we make sure that the cumulative error added by the d^α terms is small enough.

We also have to scale y^α properly. The first element of the series should be a probability distribution. Then the scaling should stay constant so that our previous observations remain true. In order to get an estimate on the mixing time, we have to find a strong lower bound on

$$\left\| \frac{y^{\alpha+k\varphi}}{\|y^\alpha\|_1} - \frac{\mathbf{1}}{n} \right\|_{TV}$$

for some $k \geq Cn^2$. We should not let k be very large though, as the cumulative estimation error has to stay under control. The proper choice of the initial α is also a key part of the process.

Surprisingly most of the time it is possible to set up these variables so as to satisfy all our requirements. But unfortunately not always. We bump into a technical constraint that the "rotation" needs to be strong.

Because of this we need to treat the case of small "rotation" in a different way. We can argue that such a chain is almost reversible, and we can show that the difference from a reversible chain is small enough to still have a mixing time of the same magnitude.

Using conventional methods it is easy to see that reversible chains on a cycle can also easily mix in $O(n^2)$ time. Consequently, relaxing the reversibility condition does not help in this case.

V. Adding a single edge

At last, let us describe a recent observation. Although there is no rigorous mathematical statement here, it demonstrates what can happen close to the previous case.

Let us change the cycle by adding a single edge, a kind of a diameter. One can see that this does not help the reversible chain at all, we will get the same optimal mixing time of the order of n^2.

Without knowing the best possible non-reversible chain, we have done simulations based on some heuristics. It turns out that there is an improvement, probably by just a constant factor. Numerically for a cycle of length 1000 the non-reversible chain had 4 times smaller mixing time. Roughly what happens is that we are forcing the chain to move clockwise. This rotation allows the chain to visit the extra feature (the additional edge) rather frequently, and take advantage of it.

This idea of creating a "highway" might also provide some improvements in other cases. It might help mixing in an indirect way, by allowing the chain to visit parts of the graph where local mixing is faster.

VI. Conclusions

In this paper we summarized the present state of an ongoing research project. Our goal is to find and exploit the extra capabilities provided by the additional freedom of non-reversible Markov chains.

The results available now are quite far from providing a comprehensive picture. Still, we could accomplish the next small step for understanding the nature of these chains, and give some promising examples as well.

Acknowledgments

I would like to express my thanks to Professor John Tsitsiklis, my host and supervisor during my stay at MIT, for motivating me to investigate the subject matter of this paper, and for his guidance during my research, and also to my supervisor Professor Gábor Tusnády for his persistent support and valuable comments.

References

[1] A. Olshevsky and J. N. Tsitsiklis, "Convergence speed in distributed consensus and averaging," *SIAM J. Control Optim.*, vol. 48, no. 1, pp. 33–55, 2009.

[2] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Trans. Inform. Theory*, vol. 52, no. 6, pp. 2508–2530, 2006.

[3] P. Diaconis and L. Saloff-Coste, "Logarithmic Sobolev inequalities for finite Markov chains," *Ann. Appl. Probab.*, vol. 6, no. 3, pp. 695–750, 1996.

[4] G.-Y. Chen, W.-W. Liu, and L. Saloff-Coste, "The logarithmic Sobolev constant of some finite Markov chains," *Ann. Fac. Sci. Toulouse Math. (6)*, vol. 17, no. 2, pp. 239–290, 2008.

[5] A. Sinclair and M. Jerrum, "Approximate counting, uniform generation and rapidly mixing Markov chains," *Inform. and Comput.*, vol. 82, no. 1, pp. 93–133, 1989.

[6] R. Kannan, L. Lovász, and R. Montenegro, "Blocking conductance and mixing in random walks," *Combin. Probab. Comput.*, vol. 15, no. 4, pp. 541–570, 2006.

[7] S. Boyd, P. Diaconis, and L. Xiao, "Fastest mixing Markov chain on a graph," *SIAM Rev.*, vol. 46, no. 4, pp. 667–689 (electronic), 2004.

[8] S. Boyd, P. Diaconis, P. Parrilo, and L. Xiao, "Fastest mixing Markov chain on graphs with symmetries," *SIAM J. Optim.*, vol. 20, no. 2, pp. 792–819, 2009.

[9] L. Saloff-Coste, "Lectures on finite Markov chains," in *Lectures on probability theory and statistics (Saint-Flour, 1996)*, vol. 1665 of *Lecture Notes in Math.*, pp. 301–413, Berlin: Springer, 1997.

[10] R. Montenegro and P. Tetali, "Mathematical aspects of mixing times in Markov chains," *Found. Trends Theor. Comput. Sci.*, vol. 1, no. 3, pp. x+121, 2006.

[11] P. Diaconis, S. Holmes, and R. M. Neal, "Analysis of a nonreversible Markov chain sampler," *Ann. Appl. Probab.*, vol. 10, no. 3, pp. 726–752, 2000.

[12] F. Chen, L. Lovász, and I. Pak, "Lifting Markov chains to speed up mixing," in *Annual ACM Symposium on Theory of Computing (Atlanta, GA, 1999)*, pp. 275–281 (electronic), New York: ACM, 1999.

[13] K. K. Gade and M. L. Overton, "Optimizing the asymptotic convergence rate of the Diaconis-Holmes-Neal sampler," *Adv. in Appl. Math.*, vol. 38, no. 3, pp. 382–403, 2007.