Sampling Algorithm for Continuous Distribution," The t-Walk"

Abhilash Singh [191003]
Ankit Gupta [191016]
Ojasvi Rajpoot [20108270]
Rajat Agarwal [191104]
Vinay Sharma [191171]

13th May 2021
IIT Kanpur

Acknowledgement

We would like here to express our heartfelt gratitude to Dr. Dootika Vats for helping us with every difficulty which we faced during this project and has also provided us with a practical insight of the theoretical knowledge gathered during the course: MTH-707A (Markov Chain Monte Carlo).

We also take this opportunity to thank the authors and publishers of the various books, journals and blogs we have consulted. Without those, this task would not have been completed. We would also like to thank our parents for their extensive support throughout the session. Their constant encouragement has enabled us to complete the project within the stipulated time-period.

Contents

1 Introduction

In this project work, we discuss "**t-walk**", the MCMC sampler for continuous distribution that requires no tuning. The sampler is introduced by J. A. Christen and C. Fox in his paper " A General Purpose Sampling Algorithm for Continuous Distributions (the t-walk) 2010". They apply Metropolis-Hasting algorithm and choose proposal with the help of four moves according to t walk design which we will discuss in details.They choose a proposal distribution in such a way that an algorithm is invariant to scale and affine transformation to the state space.
In general, this algorithm is only a small factor less efficient than optimally tuned algorithm but performed very well in comparison to general random walk M-H samplers that are not tuned for a particular problem (we later compare these two with some examples). Also, this algorithm remains effective for target distribution for which correlation structure is very different in different regions of state space.
Thus, this MCMC sampling algorithm neither contains adaptivity nor tuning parameter, and we can sample from a target with arbitrary scale and correlation structure. As t-walk construct as M-H algorithm on product space, so it converges under mild conditions.

2 Algorithm

t-walk works on the concept of the M-H algorithm. The main purpose of the t-walk sampler is choosing a proposal so the sampler maintains mixing of the chain, aperiodicity, irreducibility of chain over arbitrary target, no matter what correlation structure between two or several variablse of a target. So, fulfill this condition J. A. Christen and C. Fox choose four types of moves(proposals) that we discuss later(walk move, traverse move, hop and blow move) and they set suitable weights to there moves to form a complete proposal. So to implement t-walk, they give weight of probabilities 0.4918, 0.4918, 0.0082, 0.0082 for respectively w_w, w_t, w_h, w_b (these values are chosen on the basis of minimum integrated auto-correlated time).
If q_m is corresponding proposal where $m \in \{w, t, h, b\}$ then we will randomly choose from four different proposals, each characterised by a particular function h(.,.). First, we choose an option from

$$(y, y') = \begin{cases} (x, h(x', x)) & ;\text{ with probability } 0.5 \\ (h(x, x'), x') & ;\text{ with probability } 0.5 \end{cases}$$

and then create the proposal (y, y') from the corresponding h function and then apply the Metropolis-Hasting (MH) algorithm.
In very general case the steps of MH algorithm are as follows. Let $X_n = X$, is the initial observation, to obtain the next observation X_{n+1}. we follow the steps :-

□ $Y \sim q(x, .)$ where q(x,.) is proposal distribution
□ $Draw U \sim uniform(0, 1)$ (Y and U are independent)
□ $p = \frac{f(y)q(y,x)}{f(x)q(x,y)}$, if $U < min\{1, p\}$ set $x_{n+1} = y$, else
□ set $x_{n+1} = x$

3 Theory

The objective function $\pi(x)$, where $x \in \chi$ has the form:

$$f(x, x^{'}) = \pi(x)\pi(x^{'})$$

on the product space $\chi \times \chi$. General proposal has the form $q\{(y, y^{'})|(x, x^{'})\}$. The two restricted proposals are given as;

$$(y, y^{'}) = \begin{cases} (x, h(x^{'}, x)) & ; \text{ with probability } 0.5 \\ (h(x, x^{'}), x^{'}) & ; \text{ with probability } 0.5 \end{cases} \tag{1}$$

where $h(x', x)$ is a random variable and each move only changes one of x or x'. As a result, the entire process is based on $\chi \times \chi$. The selection of a proposal is determined by a specific function $h(.,.)$, and has four options. Since the t walk is based on the MH Algorithm, then the corresponding acceptance ratio is,

$$\text{acceptance ratio} = \frac{\pi(y^{'})}{\pi(x^{'})} \frac{g(x^{'}|y^{'}, x)}{g((y^{'}|x^{'}, x)}$$

for the case 1 and ,

$$\text{acceptance ratio} = \frac{\pi(y^{'})}{\pi(x^{'})} \frac{g(x^{'}|y^{'}, x)}{g((y^{'}|x^{'}, x)}$$

for the case 2 , where $g(.|x, x^{'})$ is density of $h(x, x^{'})$
If h(.,.) is invariant to affine transformation ϕ then sampler will be in invariant under that transformation.
For high dimension problems, J. A. Christen and C. Fox select a random subset of coordinates to be updated in each step as follows; In each of the four moves, they simulated a Bernoullian sequence of independent indicator variables $I_j \sim B(p)$, $j = 1(1)n$, if $I_j = 0$ then coordinate j is not updated. The probability p of updating a given coordinate is chosen so that np=min(n, n_1) and they set $n_I = \sum_{j=1}^{n} I_j$(they set $n_1 = 4$).

3.1 Walk move

In many applications, particularly with weak correlations, J. A. Christen and C. Fox find that mixing of the chain is primarily achieved by a scaled random walk that they called the walk move.
The walk move function h_w is defined as;

$$h_w(x, x^{'})_j = \begin{cases} x_j + (x_j - x^{'}_j)\alpha_j & ; \ I_j = 1 \\ x_j & ; \ I_j = 0 \end{cases}$$

for $j = 1(1)n$ where $\alpha_j \in \mathbb{R}$ are IID random variable with density $\psi_w(.)$
Now for considering case 2 of (eq. 1). $g(y|x, x^{'}) = \prod_{I_j=1} g_j(y_j|(x_j, x^{'}_j))$, where $g_j(y_j|(x_j, x^{'}_j)) = \psi_w(\frac{y_j - x_j}{x_j - x^{'}_j})/|x_j - x^{'}_j|$ where g(.|.,) is proposal.

This proposal is symmetric (see the paper by J. A. Christen and C. Fox) when $\psi_w(\frac{-\alpha}{1+\alpha}) = (1 + \alpha)\psi_w(\alpha)$
J. A. Christen and C. Fox achieve this by setting,

$$\psi_w(\alpha) = \begin{cases} \frac{1}{k\sqrt{1+\alpha}} & ;\ \alpha \in [\frac{-a_w}{1+a_w}, a_w] \\ 0 & ;\text{otherwise} \end{cases}$$

for any $a_w > 0$, this density is simple to from using inverse cummulative distribution as

$$\alpha = \frac{a_w}{1+a_w}(-1 + 2u + a_w u^2)$$

where $u \sim U(0,1)$. By setting this Hasting ratio for second case is one. Hence the acceptance probability is ratio of target densities. Similar process can be done for first case.

3.2 Traverse move

When there is a clear association between a few or more variables in density, samplers that use random walk moves have trouble. Rotating and scaling the coordinates of the state variables, or, equivalently, the proposal distributions, is a common solution. However, with distributions where the correlation structure changes over time, this is not possible. For those applications, the efficiency of the sampler is greatly enhanced by the 'traverse move' (see the paper by J. A. Christen and C. Fox) defined by

$$h_t(x, x')_j = \begin{cases} x_j + (x_j - x'_j)\beta & ;\ I_j = 1 \\ x_j & ;\ I_j = 0 \end{cases}$$

Where $\beta \in \mathbb{R}+$ are random variable with density $\psi_t(.)$

Since here one random variable is used in the proposal so it is not possible to make both the proposal and acceptance ratio independent of the dimension of state space. Except for the case $\beta = 1$, so J. A. Christen and C. Fox suggest $\psi_t(\beta)$;

$$\psi_t(\beta) = \frac{a_t - 1}{2a_t}\{(a_t+1)\beta^{a_t}1_{(0,1]}(\beta)\} + \frac{a_t+1}{2a_t}\{(a_t-1)\beta^{-a_t}1_{(0,\infty]}(\beta)\}$$

this is mixture of two distributions and easily sampled from ,

$$x\beta = \begin{cases} u^{\frac{1}{1+a_t}} & ;\ \text{with probability} = \frac{a_t-1}{2a_t} \\ u^{\frac{1}{1-a_t}} & ;\ \text{with probability} = \frac{a_t+1}{2a_t} \end{cases}$$

where $u \sim U(0,1)$ they find optimal $a_t = 6$

3.3 Hop and Blow moves

The walk and traverse moves are insufficient to guarantee chain irreducibility over arbitrary target distributions. Furthermore, both the walk and traverse moves will result in extremely slow mixing for highly correlated distributions. Therefore J. A. Christen and C. Fox introduces hop and blow moves, hop move is defined by the function;

$$h_h(x,x^{'})_j = \begin{cases} x_j + \frac{\sigma(x,x^{'})}{3} z_j & ; \ I_j = 1 \\ x_j & ; \ I_j = 0 \end{cases}$$

where $z_j \sim N(0,1)$, $\sigma(x,x^{'}) = \max_{I_j=1}|x_j - x_j^{'}|$. For this proposal

$$g_h(y|(x,x^{'})) = \frac{(2\pi)^{-n_I/2}3^{n_I}}{\sigma(x,x^{'})^{n_I}} \exp\left\{\frac{-9}{2\sigma(x,x^{'})^2}\sum_{I_j=1}(y_j - x_j)^2\right\}\Pi_{I_j}\delta_{x_j}(y_j)$$

clearly here, this move is centred at x.

Similarly the blow move defined by;

$$h_b(x,x^{'})_j = \begin{cases} x_j + \sigma(x,x^{'}) z_j & ; \ I_j = 1 \\ x_j & ; \ I_j = 0 \end{cases}$$

where $z_j \sim N(0,1)$, For this proposal;

$$g_b(y|(x,x^{'})) = \frac{(2\pi)^{-n_I/2}}{\sigma(x,x^{'})^{n_I}} \exp\left\{\frac{-1}{2\sigma(x,x^{'})^2}\sum_{I_j=1}(y_j - x_j^{'})^2\right\}\Pi_{I_j}\delta_{x_j}(y_j)$$

clearly here, this move is centred at $x^{'}$.

3.4 Convergence

In this section we review convergence, Let K_m be the M-H transition kernel for the proposal q_m, where $m \in \{w.t, h, b\}$. Strong aperiodicity is ensured by positive probability of rejection in M-H scheme. J. A. Christen and C. Fox forms the transition kernel;

$$K\{(x,x^{'}),(y,y^{'})\} = \sum_{m\in\{w.t,h,b\}} w_m K_m\{(y,y^{'})|(x,x^{'})\}$$

Where $\sum_{m\in\{w.t,h,b\}} w_m = 1$, which also satisfy the detailed balance condition with f. Assuming K is f irreducible(hop and blow moves ensure f-irreducibility), then f is the limit distribution of K (see Robert and Casella 1999, Chapter 6, for details).

4 Example

In this section we will work on a bi-modal, correlated objective function having the form;

$$\pi(x) = C\exp\left\{ -\tau(\sum_{i=1}^{2}(x_i - m_{1,i})^2)(\sum_{i=1}^{2}(x_i - m_{2,i})^2)\right\}$$

where C is normality constant, τ is a scale parameter and $(m_{1,1}, m_{1,2})$, $(m_{2,1}, m_{2,2})$ are approximately bi-modal on $\mathbb{R}^2$.

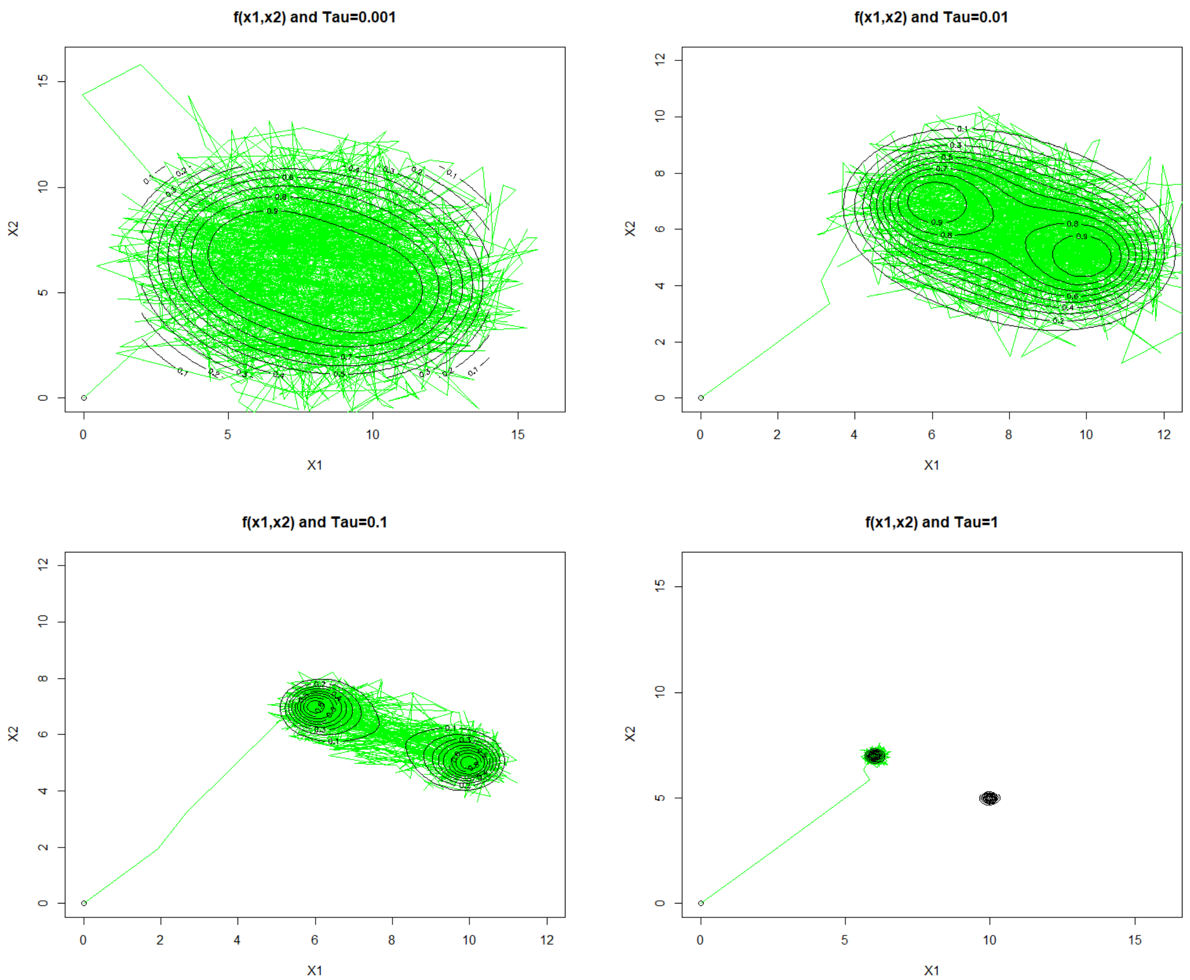


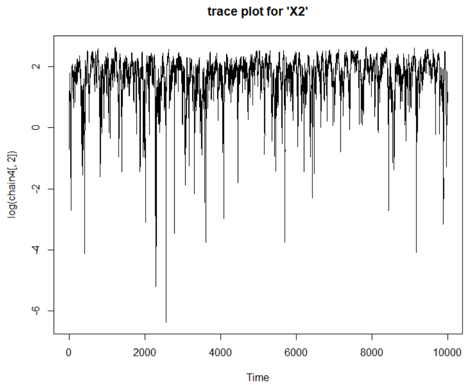
Figure 1: Sample path for t-walk sampler for τ=(0.001,0.01,1,0.1)

Here we will generate a sample from this target objective function using **t-walk** and **Random Walk MH algorithm**. We will represent sample paths for T-walk for $\tau = 0.001, 0.01, 0.1, 1$ and then try to compare their trace plots and acceptance probability.

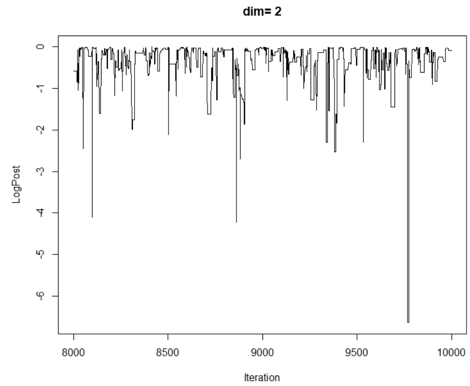
In the above fig we represent the contour density plot of our target distribution for different $\tau = 0.001, 0.01, 0.1, 1$ (respectively from upper left corner in anti-clock wise direction) based on sample generated by the **t-walk sampler**.In this example we take m1=(10,5) and m2=(6,7) as the bi-modal points of density and which can be clearly seen by contour plots for each τ.It is observed that as the τ increases, the sampler spans a lesser portion of support. For example, when $\tau = 1$ all the generated points are concentrated around (6,7) and the sampler keeps moving around this point. And if we further increase τ,(say $\tau = 1000$), the distribution shape can not be distinguished and it is reduced to a point. This shows that increasing the value of τ is not a good idea. (Note :- Here green jig-jag lines show the path of the sample generated).

4.1 Comparison between t-walk and Random Walk MH sampler

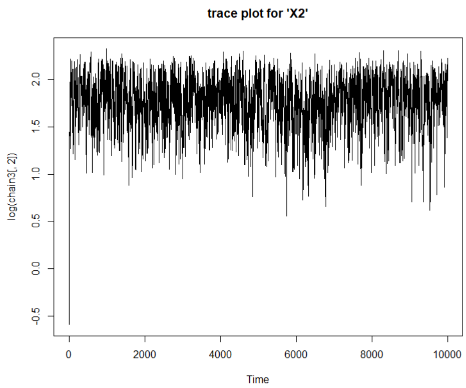
In this subsection we will compare the t-walk sampler with a tuned M-H MCMC sampler named Random walk M-H sampler. In the Random walk M-H sampler, we take bi-variate normal as a proposal distribution. We compare trace plots of both samplers in the figure.



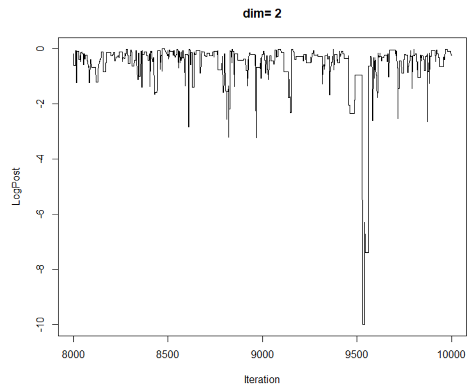
(a) Random walk MH, τ=0.001



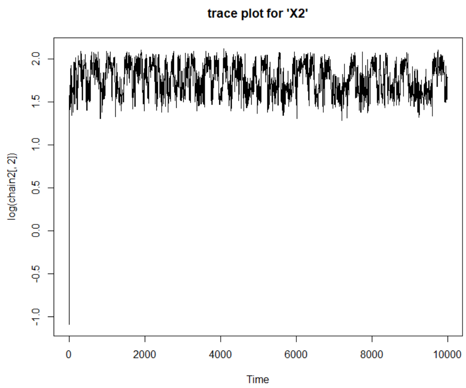
(b) T walk, τ=0.001



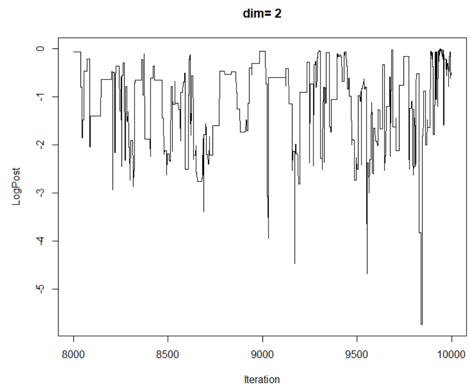
(c) Random walk MH, τ=0.01



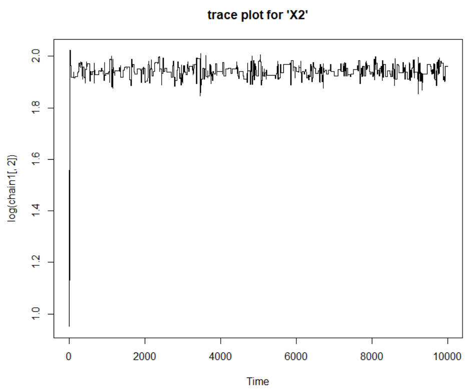
(d) T walk, τ=0.01



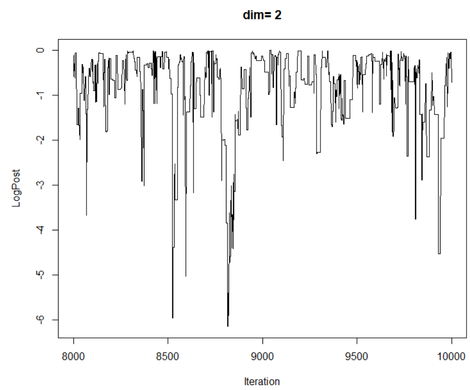
(e) Random walk MH, τ=0.1



(f) T walk, τ=0.1



(g) Random walk MH, τ=1



(h) T walk, τ=1

Figure 2: Comparison between trace plots of "X_2" for Random walk MH(left) and walk(right)

While comparing sampler using trace plots, the sampler having more zig-zag trace plots, exploring the more sample space, is the better one compared to the other one. So clearly from the trace plots given above, for $\tau = 0.1 and 1$ t-walk is better as it explore more sample space while for τ =0.01 and 0.001 Random walk MH sampler is better.

As this method is based on visualisation and not a quantified method. Now we will compare on the basis of acceptance probability and effective sample size per second (ESS per second).

τ	t walk	Random walk MH
τ=1	0.4345	0.0503
τ=0.1	0.3652	0.3439
τ=0.01	0.4157	0.7226
τ=0.001	0.4529	0.8497

Table 1: comparison of "Acceptance Probability" between t walk and Random walk MH for different values of τ

The above table shows the acceptance probability for both samplers for different τ. So we can easily conclude that for $\tau = 1$ "t-Walk is much better while for $\tau = 0.01, 0.001$ random walk is much better. And for $\tau = 0.1$ both perform similar (trace plots are also very much similar for this case)
One important thing we can notice here, the acceptance probability for "t-walk" is not varying too much on changing τ while for Random walk,it is changing very much.So this is a kind of benefit of "t-walk" that we don't have to tune scale parameter τ.

τ	T walk	Random walk
τ=1	600	5454
τ=0.1	43	1565
τ=0.01	260	5637
τ=0.001	278	2836

Table 2: Mean effective sample size per second

The above table shows approximately mean effective sample size per second. From the table, one can easily conclude that Random Walk MH is much better than t-walk in the sense of effective sample size per second. So, if time is constrained, then we should prefer Random walk MH sampler over t-walk sampler.

5 Conclusion

So, based on the preceding discussion, we conclude that if the scale parameter tau is optimally tuned, Random walk M-H outperforms t walk. But if we don't have a good idea about tuning of τ and initial points from where we should start, then t walk is a good choice of sampler as compared to the M-H sampler. So, for the average person who is not a statistician and has no idea how to choose tuning parameters, t walk is a better sampler.

6 Pros And Cons of t-walk

6.1 Pros:

1. In Adaptive algorithms, our effort is to learn the scale and correlation structure of the complex target distribution. But the t-walk is defined in such a way that it is invariant to this type of structure.

2. t-walk doesn't require tuning parameters.

3. t-walk is useful as a black box sampling algorithm. It allows researchers to focus on the data analysis rather than MCMC. Because implementation of the t-walk algorithm is very easy, it only requires the log of target distribution and two initial points in the parametric space.

4. t-walk is useful in multiple data analysis where details of the posterior distribution depend sufficiently on a particular data set that adjustment would be required to proposal in a standard MH algorithm allowing automatic use of MCMC samplings.

6.2 Cons:

1. It may be possible that the t-walk is not as efficient as a well tuned algorithm.

2. t-walk is probably convergent under some usual mild conditions. We choose those proposal distribution or move that produce an algorithm that is invariant to scale and approximately invariant to affine transformation of the state space.

7 References

.

- J.Andŕes christen and Colin Fox. "A General Purpose Sampling Algorithm for Continuous Distribution(the t-walk)" Bayesian Analysis(June 2010) 5,Number 2, pp. 263-282 DOI: 10.1214/10-BA603.

- `http://www.cimat.mx/~jac/twalk/`