1 Algorithm

Our primary interest in this paper lies in simulating from d-dimensional target densities $\pi(x_d)$, comprising of iid target densities $f_d(.)$ that are complicated and potentially multimodal (for instance, mixture of normal components) in nature. The main problem in using MCMC for target densities of this form is that the MCMC chain may get stuck in one component for a long period of time. This results in poor exploration of the entire state space and gives biased samples. There has been a lot of research in the MCMC literature that focuses on solving this problem. Two very popular approaches are parallel tempering and simulated tempering. Atchade et al (2010) [3] proposed a Metropolis-coupled Markov Cahin algorithm (MC^3) and compared the $ESJD$ for the MC^3 algorithm with that of the Simulated tempering algorithm. In this paper, we are presenting a modification of the MC^3 algorithm by randomizing the inverse temperature spacings for the tempered distributions that are run in parallel.

We define a sequence of heated or tempered density functions (non-normalized) as ${f_d}^{\beta_j}(x) = (f_d(x))^{\beta_j}$ with $0 \leq \beta_n < \beta_{n-1} \leq < \cdots < \beta_1 < \beta_0 = 1$. these β_j are also called inverse temperatures. We propose two algorithms - in one (called RMC^3) we run parallely Metropolis-Hastings chains for each ${f_d}^{\beta_j}$ and for the other (called $RTMC^3$), we run parallely Additive TMCMC chains (refer to Dutta and Bhattacharya (2013) [4], Dey and Bhattacharya (2013) [2] and Dey and Bhattacharya (2013) [5]) for each ${f_d}^{\beta_j}(.)$. Note that in either method, the chain corresponding to β_0 is the cold chain and the main chain of interest as it corresponds to the original non-tempered density. The main hope behind using these heated or tempered chains for varying β_j is that chains with smaller inverse temperatures will be able to mix more easily as they would be flatter. We keep a swapping phase in our algorithm, in which we swap the iterates for chains of different inverse temperatures with the hope that they would add some valuable mixing information to the original cold chain corresponding to b_0.

There are two phases wo the RMC^3 or the $RTMC^3$ algorithms and these two phases are analogous to the ones recommended by Atchade et al (2010) [3].

- Normal iteration step
- Temperature Swap step

In the **first step**, we update each of the chains (corresponding to each β_j or inverse temperature) separately either by the RWMH (Random Walk Metropolis Hastings) or by the Additive TMCMC updates in RMC^3 and $RTMC^3$ respectively. These parallel chains have target densities ${f_d}^{\beta_j}(.)$.

In the **second step**, we swap the iterates corresponding to two inverse temperatures. We choose two inverse temperatures β_j and β_k and try to swap the iterates x_j and x_k corresponding to ${f_d}^{\beta_j}(.)$ and ${f_d}^{\beta_k}(.)$ respectively. We accept the swap with the probability given by

$$Prob(accepting\ swap) = min\left(1, \frac{{f_d}^{\beta_j}(x_k){f_d}^{\beta_k}(x_j)}{{f_d}^{\beta_j}(x_j){f_d}^{\beta_k}(x_k)}\right)$$

Otherwise, the chains remain unchanged.

The main question of interest is that how should we choose the inverse temperatures optimally, or rather, if we want to swap the chain with inverse temperature β_j with another inverse temperature β_k, how should we choose these values optimally. Atchade et al (2010) [3] consider the case where we pick adjacent inverse temperatures for swapping at a specific distance, say β and $\beta + \epsilon$ and our main query of interest then is how to optimally choose ϵ. We propose here an alternative where this ϵ is random and has a distribution on the positive support, for convenience, we consider the case that $\epsilon \sim TN_{>0}(0, \frac{l^2}{d})$ for dimension d and our new focus would be to choose an appropriate scaling l. Note that if l is large, then ϵ would tend to be larger and this would result in greater rejection of the swaps. Again, if l is small, then we shall have higher acceptance of swaps but the two swapping inverse temperatures would be very close to each other and so, ideally, with the swapping iterates also likely to be close, thereby resulting in poor mixing. So, in the next section, we shall deal with the problem of determining the optimal value of the scaling l. We shall show that while in Atchade et al's paper [3], their choice of ϵ leads to an optimal acceptance rate of 0.234, our method corresponding to the optimal scaling l_{opt} gives an optimal acceptance rate of 0.439 which is almost double that of the MC^3 algorithm. Note that while in the MC^3 algorithm, the inverse temperatures are equally spaced, in our proposed methods (RMC^3 or $RTMC^3$) the inverse temperatures are irregularly spaced due to the inherent randomness in ϵ and that lends extra flexibility to our algorithm.

Theorem 1 *Consider the RMC^3 or the $RTMC^3$ chain with a target density π and allow swap of the temperatures between two consecutive inverse temperature heated chains only, namely β and $\beta + \lambda$ where we consider λ to have the distribution $TN_{>0}(0, \frac{l^2}{d})$. Then as the dimension of the chain $\mathrm{d} \to \infty$, the ESJD is maximized when we choose l to be maximize the function $D(\mathrm{l}) = \int_Z \mathrm{l}^2 Z^2 \Phi\left(-\frac{\mathrm{l}|Z|\eta(\beta)}{\sqrt{2}}\right) \phi(Z) dZ$ and this optimal choice of l can be shown to lead to a limiting expected acceptance rate of 0.439 which is almost twice that of the standard MC^3 algorithm.*

2 Proof of the main theorem

We consider the specific problem of determining the optimal temperatures of the sequence of the heated chains in the Randomized Metropolis Coupled MCMC (RMC^3) and Randomized Additive Transformation based MCMC ($RTMC^3$). Suppose we are at the n th iteration and we denote the filtration, or the σ-field formed by the iterates upto the $n-1$ iterations over all the parallely run chains to be $\mathcal{F}_{n-1}$. We consider the swapping between the temperatures β and $\beta + Y$ where we assume that $Y \sim TN_{>0}(0, \frac{l^2}{d})$ where $TN_{>0}()$ represents the truncated normal random variable left truncated at 0. We assume that $\beta > 0$, λ is a random variable that $\lambda > 0$ with probability 1 and $\beta + \lambda \leq 1$ with probability 1 as $d \to \infty$ for $\beta \in (0, 1)$. We wish to find the optimal vale of l under the assumption of stationarity of the chain ($X \sim \prod_{j=1}^{n} {f_d}^{\beta_j}$) Let us define $\beta^\star = \beta + \lambda$ if we accept the swap and $\beta^\star = \beta$ otherwise. If the proposed step is accepted, then we move to the new point $\beta^\star$ and if it is rejected, then we stay where we are at β.The general approach that we follow for finding out an optimal choice of l would be to maximize the stationary expected squared jumping distance (ESJD) as

$$\begin{aligned} ESJD &= E_\pi(\beta^\star - \beta)^2 \\ &= \int \lambda^2 E_\pi \left[min\left(1, \frac{{f_d}^{\beta_j}(x_k){f_d}^{\beta_k}(x_j)}{{f_d}^{\beta_j}(x_j){f_d}^{\beta_k}(x_k)}\right)\right] \end{aligned}$$

where x_j is the present iterate of the chain corresponding to the jth heated chain denoted by the suffix β_j and x_k is the present iterate corresponding to the kth heated chain. Consider the target density to be $\pi_d(x) = \prod_{i=1}^{d} f(x_i)$ and the heated chain with suffix β is given by ${\pi_d}^\beta(x) = \prod_{i=1}^{d} f(x_i)^\beta$. Note that we can write $Y = \frac{lZ}{\sqrt{d}}$ where Z is $TN_{>0}(0, 1)$ random variable. We shall show that for both the RMC^3 and the $RTMC^3$ algorithms, as $d \to \infty$, the ESJD is maximized when l is chosen to maximize $2 \int l^2 z^2 \Phi(-\frac{lz\sqrt{I(\beta)}}{2})$. This is equivalent to finding out the optimal scaling for the Additive TMCMC approach with the same target density. The acceptance probability given the random variable λ is given by $\alpha(\lambda) = min(1, exp(B))$ where B is given by

$$B = ln\left(\frac{{f_d}^{\beta}(y){f_d}^{\beta+\lambda}(x)}{{f_d}^{\beta}(x){f_d}^{\beta+\lambda}(y)}\right)$$

This can be reduced to the expression

$$\begin{aligned} B &= \left(ln({f_d}^{\beta}(y)) - ln({f_d}^{\beta+\lambda}(y))\right) - \left(ln({f_d}^{\beta}(x)) - ln({f_d}^{\beta+\lambda}(x))\right) \\ &= R_d(y) - R_d(x) \end{aligned}$$

Let us denote $h(x) = ln(f(x))$ and we know that ${f_d}^{\beta}(x) = \prod_{i=1}^{d} f(x_i)^{\beta}$. Using this, the above expression reduces to the following

$$R_d(x) = \beta\sum_{i=1}^{d} g(x_i) - (\beta+\lambda)\sum_{i=1}^{d} g(x_i) = -\sum_{i=1}^{d}\frac{lZ}{\sqrt{d}}g(x_i)$$

Now we consider the expectation and the variance of the function g under the density function f^{β}, suitably normalized. Then the expectation and the variance of the function g with respect to the normalized density measure f^{β} is given by

$$E^{\beta}(g) = \frac{\int log(f(x))f^{\beta}(x)dx}{\int f^{\beta}(dx)} = m(\beta)$$

$$V^{\beta}(g) = \frac{\int (logf(x))^2 f^{\beta}(x)dx}{\int f^{\beta}(x)dx} - (E^{\beta}(g))^2 = \eta^2(\beta)$$

Then the conditional expectation and the conditional variance of R_d given the random variable Z is given by

$$E^{\beta}(R_d(x)|Z) = -\sqrt{d}lZE^{\beta}(g) = \mu_{\beta,Z}$$

$$V^{\beta}(R_d(x)|Z) = dl^2Z^2V^{\beta}(g) = \sigma^2_{\beta,Z}$$

where the expressions for $E^{\beta}(g)$ and $V^{\beta}(g)$ are given above. It can be shown by differentiating $\mu_{b,Z}$ with respect to β that

$$m^{'}(\beta) = \eta^2(\beta) \qquad\qquad \mu^{'}_{b,Z} = -\frac{\sqrt{d}}{lZ}\sigma^2_{\beta,Z}$$

This implies that

$$R_d(y) - R_d(x) = -\frac{lZ}{\sqrt{d}}\sum_{i=1}^{d}(g(y_i) - g(x_i))$$

Our aim is to derive the asymptotic distribution of the quantity $R_d(y) - R_d(x)$ and the characteristic function is given by

$$\begin{aligned}
\phi_d(t|Z,\mathcal{F}_{n-1}) &= E|_{Z,\mathcal{F}_{n-1}}\left[exp\quad\left(\quad it(R_d(y)-R_d(x))\right)\right]\\
&= E|_{Z,\mathcal{F}_{n-1}}\left[exp\left(-it\frac{lZ}{\sqrt{d}}\sum_{i=1}^{d}(g(y_i)-g(x_i))\right)\right]\\
&= E_{Y|Z,\mathcal{F}_{n-1}}\left[exp\left(-it\frac{lZ}{\sqrt{d}}\sum_{i=1}^{d}\{g(y_i)-m(\beta+\epsilon)\}\right)\right]E_{X|Z,\mathcal{F}_{n-1}}\left[exp(\left(it\frac{lZ}{\sqrt{d}}\sum_{i=1}^{d}\{g(x_i)-m(\beta)\}\right)\right]\\
&\qquad\times E|_{Z,\mathcal{F}_{n-1}}\left[exp\left(-it\frac{lZ}{\sqrt{d}}(m(\beta+\epsilon)-m(\beta))\right)\right]
\end{aligned}$$

where $\mathcal{F}_{n-1}$ is the filtration generated by the $n-1$ iterations of all the parallel chains in the RMC^3 or the $RTMC^3$ algorithm corresponding to all the inverse temperatures (meaning all the heated chains with various degrees of heating) along with the cold chain- the main chain of interest.

By the application of the Central Limit theorem we know that $\frac{1}{\sqrt{d}}\sum_{i=1}^{d}(g(x_i)-m(\beta))\sim N(0,\eta^2(\beta))$ where $\mathbf{x}\sim {f_d}^{\beta}$. This implies by the Taylor series theorem that

$$E_{Y|Z}\left[exp(-it\frac{lZ}{\sqrt{d}})\sum_{i=1}^{d}g(y_i)-m(\beta+\lambda)\right]=1-\frac{l^2Z^2}{2d}\eta^2(\beta+Y)+o\left(\frac{1}{d}\right)\tag{1}$$

$$E_{X|Z}\left[exp(-it\frac{lZ}{\sqrt{d}})\sum_{i=1}^{d}g(x_i)-m(\beta)\right]=1-\frac{l^2Z^2}{2d}\eta^2(\beta)+o\left(\frac{1}{d}\right)\tag{2}$$

Also we have given Z,

$$itlZ\sqrt{d}(m(\beta+Y)-m(\beta))\to itl^2Z^2\eta^2(\beta)\qquad as\ d\to\infty\tag{3}$$

This follows from the fact that given Z, the variable Y is known and then from the definition

$$\underset{d\to\infty}{Lt}\frac{m(\beta+Y)-m(\beta)}{Y}\to m^{'}(\beta)=\eta^2(\beta)$$

So from the above equations **Eq 1**, **Eq 2** and **Eq 3**, it follows that $\phi_d(t|Z,\mathcal{F}_{n-1})$ converges to

$$\phi^{\star}(t|Z,\mathcal{F}_{n-1})=exp(-itl^2\eta^2(\beta)Z^2-l^2t^2Z^2\eta^2(\beta))$$

.

This is the characteristic function of $N(-l^2\eta^2(\beta)Z^2,l^2t^2Z^2\eta^2(\beta))$ density given the random variable Z.

Note that a standard result (see Roberts et al. 1997 [1], Dey and Bhattacharya [2]) that for a random variable $X\sim N(\mu,\sigma^2)$

$$E\left(min(1,exp(X))\right)=\Phi(\frac{\mu}{\sigma})+exp(\mu+\frac{\sigma}{2})\Phi(-\sigma-\frac{\mu}{\sigma})$$

From the above calculations, we have shown that the limiting distribution of $B\sim N(-l^2\eta^2(\beta)Z^2,l^2t^2Z^2\eta^2(\beta))$ and hence, the limiting expected acceptance rate of the swap given λ or equivalently given Z is given by

$$E|_{Z,\mathcal{F}_{n-1}}\left(min(1,exp(B(Z)))\right) \approx 2\Phi(-\frac{lZ\eta(\beta)}{\sqrt{2}})$$

However, as per our assumption the heatings are not uniformly distributed as in the standard MC^3 and the randomness is expressed in terms of Z which has a $TN_{>0}(0,1)$ distribution. Under this set up, we have

$$ESJD = \int_{Z>0} \frac{l^2Z^2}{d} E_{Z,\mathcal{F}_{n-1}}\left(\quad min\left(1, exp(B(Z))\right)\right) \phi^{+}(Z)dZ$$

Then in the limit we can write this as

$$ESJD \approx 2\int_{Z>0} \frac{l^2Z^2}{d}\Phi\left(-\frac{lZ\eta(\beta)}{\sqrt{2}}\right)\phi^{+}(Z)dZ \tag{4}$$

where $\phi^{+}(.)$ is the density function for the $TN_{>0}(.)$ random variable. Note that instead of generating a Z from $TN_{>0}(0,1)$ random variable and updating β by $\beta + \frac{lZ}{\sqrt{d}}$, we can equivalently generate Z from $N(0,1)$ distribution and then update β by $\beta + \frac{l|Z|}{\sqrt{d}}$. Using this analogy, it trivially follows from **Eqn 4** that

$$ESJD \approx 4\int_{Z} \frac{l^2Z^2}{d}\Phi\left(-\frac{l|Z|\eta(\beta)}{\sqrt{2}}\right)\phi(Z)dZ \tag{5}$$

In order to maximize the $ESJD$ with respect to l we aim to optimize the function

$$\mathcal{D}(l) = \int_{Z} l^2Z^2\Phi\left(-\frac{l|Z|\eta(\beta)}{\sqrt{2}}\right)\phi(Z)dZ$$

with respect to l. This is analogous to the expression for the diffusion speed we obtained in Dey and Bhattacharya (2013) [2]. As has been shown in that paper already, the maxima of this function is attained at the value of l

$$l_{opt} = \frac{2.426}{\eta(\beta)}$$

and the value of the optimal acceptance rate corresponding to this optimal value of l is given by

$$\alpha_{opt} = 4\int_{Z} \Phi\left(-\frac{l_{opt}|Z|\eta(\beta)}{\sqrt{2}}\right)\phi(Z)dZ \approx 0.439$$

This optimal value of the acceptance rate is higher than that of the MC^3 algorithm derived in Atchade et al. (2010) [3]. We emphasize the fact that under the optimal scaling for our proposed algorithm, we have higher limiting expected acceptance rate of the swaps, meaning higher degree of exchange between the parallel chains and this will lead to better exploration of the state space. Also the graphs of $\mathcal{D}(l)$ implies that the $ESJD$ for both the MC^3 and that of RMC^3 / $RTMC^3$ are close corresponding to l_{opt} but that of MC^3 is slightly higher and for higher values of l, the $ESJD$ for the MC^3 chain falls rapidly while that of the RMC^3 or $RTMC^3$ chains remain more or less stable for a wide range of l.

Note that one big assumption in this calculation is that the process is in stationarity. The higher value of $D(l_{opt})$ in MC^3 implies that the ESJD would be marginally higher under stationarity compared to that of RMC^3 or $RTMC^3$ algorithms for the optimal scaling. But to have a valid comparison of the two methods, we must compare the performance even under non-stationarity, and under non-stationarity, our algorithm has a clear edge over MC^3 because it has higher acceptance rate of the swaps and hence has better exploration of the state space compared to the MC^3.

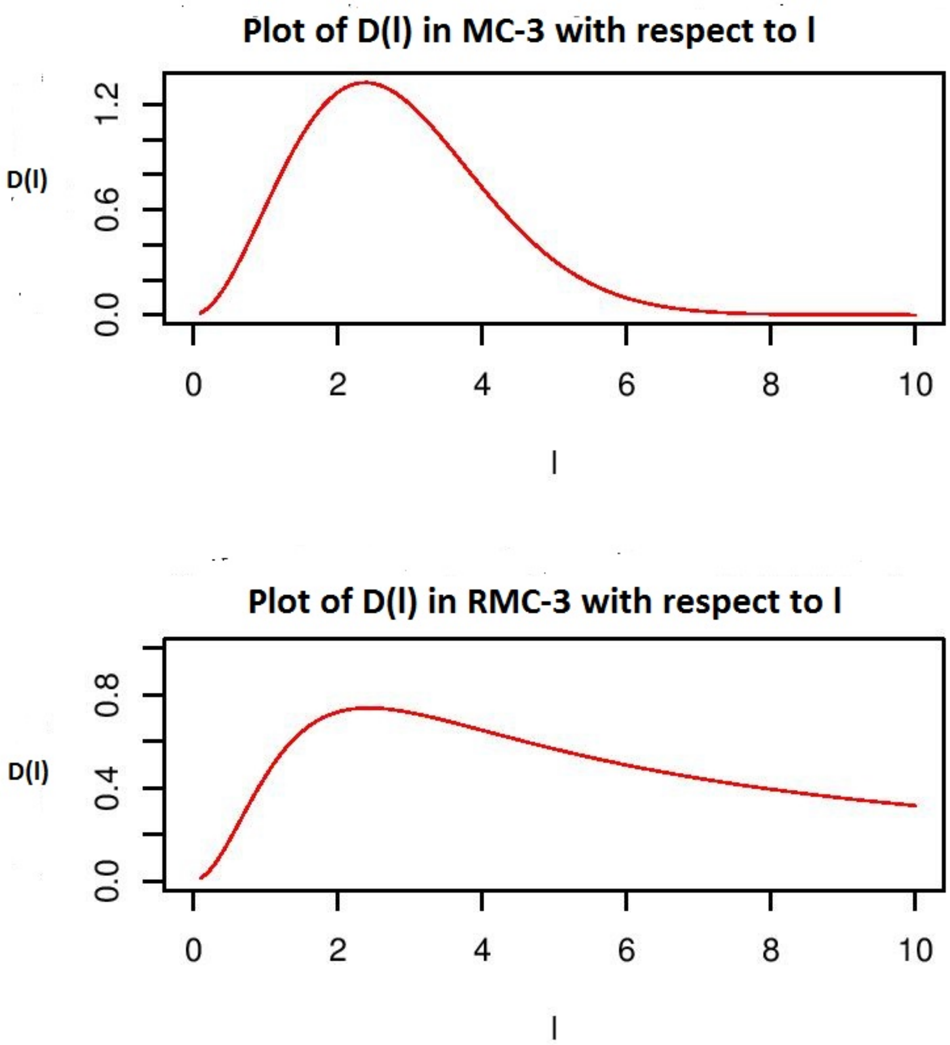

Figure 1: The plots of $D(l)$ which is proportional to the $ESJD$ with respect to l for the MC^3 and the RMC^3 algorithms.

References

[1] Roberts, G.O., Gelman, A. and Gilks, W.R. (1997), ”Weak convergence and Optimal Scaling of Random Walk Metropolis algorithms” *Ann.Appl.Prob*, 7, 110-120

[2] Dey K.K. and Bhattacharya S.(2013), ”On Optimal Scaling of Additive Transformation Based Markov Chain Monte Carlo”, *preprint*, arxiv:1307.1446

[3] Atchade, Y.F., Roberts, G.O. and Rosenthal, J.S. (2010), ”Towards Optimal Scaling of Metropolis-coupled Markoc Chain Monte Carlo” *Stat Comput*, 21, 555-568

[4] Dutta, S. and Bhattacharya, S. (2013), ”Markov Chain Monte Carlo Based on Deterministic Transformations” *Stat Methodology*, 16, 100-116

[5] Dey, K.K. and Bhattacharya, S. (2013), ”On Geometric Ergodicity of Additive and Multiplicative Transformation Based Markov Chain Monte Carlo in High Dimensions”, *preprint*, arXiv:1312.0915