HMC : When is it worth over RWMC ?

Course : Stochastic Simulation
Students : Aude Maier & Tara Fjellman
Fall 2024

1 Introduction

2 Core Theory

2.1 RWMC

2.2 HMC

2.2.1 The Algorithm

2.2.2 Acceptance Rate

To explore how the acceptance rate behaves in HMC, we must consider the quantity $\exp\left[U\left(q^n\right)+K\left(p^n\right)-U\left(q^*\right)-K\left(p^*\right)\right]$ which appears in the expression for α in the Metropolis-Hastings acceptance probability. Using the definition $H(q,p)=U(q)+K(p)$ we can write this quantity as:

$$\exp\left(U\left(q^n\right)+K\left(p^n\right)-U\left(q^*\right)-K\left(p^*\right)\right)=\exp\left[H\left(q^n,p^n\right)-H\left(q^*p^*\right)\right]. \tag{1}$$

Since the Hamiltonian is conserved under the Hamiltonian dynamics :

$$\begin{aligned}\frac{dH}{dt}&=\sum_i\frac{\partial H}{\partial p_i}\frac{dp_i}{dt}+\sum\frac{\partial H}{\partial q_i}\frac{\partial q_i}{dt} &(2)\\&=-\sum_i\frac{\partial H}{\partial p_i}\frac{\partial H}{\partial q_i}+\sum_i\frac{\partial H}{\partial q_i}\frac{\partial H_i}{\partial p_i}=0, &(3)\end{aligned}$$

we find by using this in Eq.1 that if integration is exact, the acceptance rate is always 1.

Under the assumption that the Hamiltonian dynamics is discretised, conservation is there in the best case on average. This implies the acceptance rate will be less than 1.

2.2.3 Convergence to Target Distribution

Gibbs distribution invariance Under the assumption that there is no numerical error, we want to prove that the Gibbs measure is invariant for the chain generated by the hamiltonian dynamics.

This is equivalent to saying that the Gibbs measure π is the same before and after an evolution of t seconds from the hamiltonian dynamics. To prove this we first introduce hamiltonian dynamics operators φ,Φ acting respectively on the phase space and the Gibbs measure : $\varphi_t(q_s,p_s)=(q_{s+t},p_{s+t});\Phi_t[\pi_s]=\pi_{t+s}\quad\forall t\in\mathbb{R}$. The statement we want to prove can then be expressed as

$$\Phi_t[\pi_s](D)=\pi_{s+t}(D)\quad\forall D\in\mathcal{B}(\Omega),\forall s,t\in\mathbb{R}, \tag{4}$$

with Ω the phase space.

We can now write the left hand side of the equation as

$$\begin{aligned}\Phi_t[\pi_s](D)&=\int_D\pi_{s+t}(q,p)\ dqdp &(5)\\&=\int_{\varphi_{-t}(D)}\pi_s(q,p)\ dqdp &(6)\\&=\pi_s(\varphi_{-t}(D)). &(7)\end{aligned}$$

The final result is obtained using the fact that volumes in phase space are preserved by the hamiltonian dynamics (in conservative systems). This result is known as Liouville's theorem, but is mentioned as theorems 2.3 in [cite].

This implies specifically that $q_k \sim \pi$ for all $k \in \mathbb{N}$ if $q_0 \sim \pi$.

If the dynamics is discretised with the Velocity Verlet algorithm, the volume in phase space is preserved up to a small error, which is why the algorithm is used in practice [cite wikipedia].

3 Exploring a 2D example

3.1 Context

In this section we explore the performance of the presented algorithms on a 2D example. The target distribution is taken as $f_1(q_1, q_2) = e^{-\alpha(q_1^2+q_2^2-0.25)^2}$, with $\alpha > 0$ a parameter. This unormalised density is represented in figure Fig.1 for two different values of α. As it can be seen, the density has the shape of a doughnot and α controls its thickness. We expect the $\alpha = 1000$ case to be more difficult to sample from than the $\alpha = 10$ case, as the density is more localised.

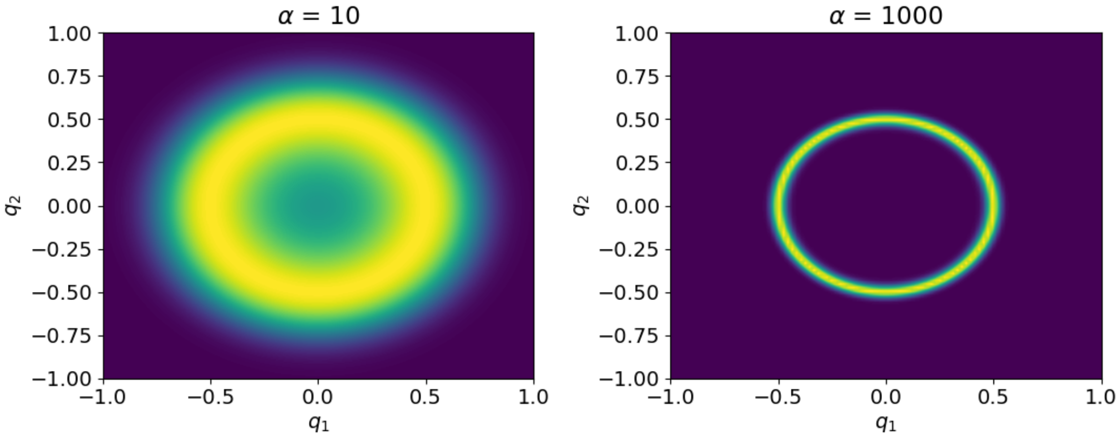

Fig 1: *Density considered in this section for two different values of* α.

3.2 RWMC Solution

As seen previously, the RWMC algorithms only depends on the step size. To find the best RWMC sampler we therefore explore the impact of the step size on performance.

Here and in the following, We decide to quantify performance throught the computation of a similarity based on the Jensen-Shannon divergence. This choice allows us to feed in a discretised version of f_1 (which we can normalise) and the empirical distribution of the samples generated by the algorithm, and get a similarity measure between the two.

The similarities associated to 3000 samples for the different step sizes are presented in figure Fig.2.

- both plots display a peak for a step size in the centre of the range considered (around 8.5×10^{-2} and 6.5×10^{-2} for $\alpha = 10$ and $\alpha = 1000$ respectively). This is expected as the step size is a tuning parameter that should be chosen to match the scale of the target distribution. - the peak is sharper (espescially on the right) for the $\alpha = 1000$ case, which is consistent with the fact that the density is more localised. - the value of the similarity is in all cases smaller than .5, which means that 3000 samples are too few to accurately estimate the target distribution. The value is higher in the $\alpha = 1000$ case, which can at first look surprising. Indeed, this case is meant to be harder than the $\alpha = 10$ one, but the fact that the density is more localised for $\alpha = 1000$ actually means that there are fewer places where the estimate and the target

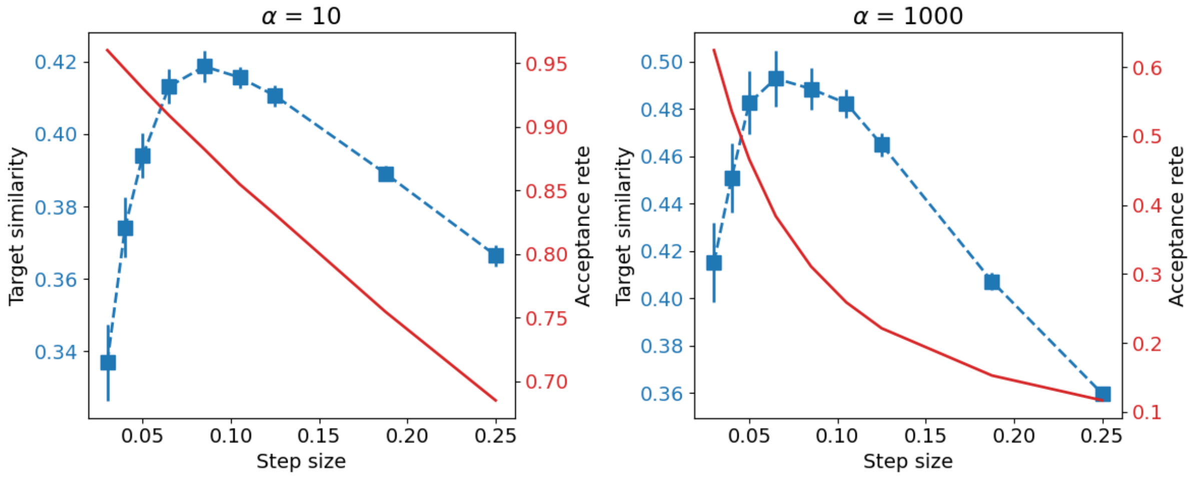

Fig 2: *Similarity as a function of RWMC step size for considered values of α.*

can differ, which can lead to a better similarity. It is therefore most important to consider the relative values of the similarities for the different step sizes, rather than the absolute values. - looking at acceptance rate, we see that they are of course monotonically deacreasing with step size. This means that the obtimal value corresponds with the best exploration-acceptance rate trade-off. The acceptance rate is higher and decreases slower for the $\alpha = 10$ case, which is consistent with the fact that the density is more spread out. The acceptance rate of the $\alpha = 1000$ case associated to the best step size is still around .55, which suggests that this case is still quite easy to sample from.

3.3 HMC Solution

Before exploring the impact of the different parameters of the HMC algorithm, we first present the potential energy landscape associated to the algorithm for this specific problem. The landscape is presented in figure Fig.3. The landscape has polar symmetry, meaning the trajectories will be circles of fixed radius. It has global minima at a radius of 0.5 [CHECK] away from the centre and a local maxima at the centre. The only difference in the landscape for $\alpha = 1000$ w.r.t. the $\alpha = 10$ one is the scale of the potential. This means that the $\alpha = 1000$ will give rise to stronger potential forces, which translates the fact that the density is more localised.

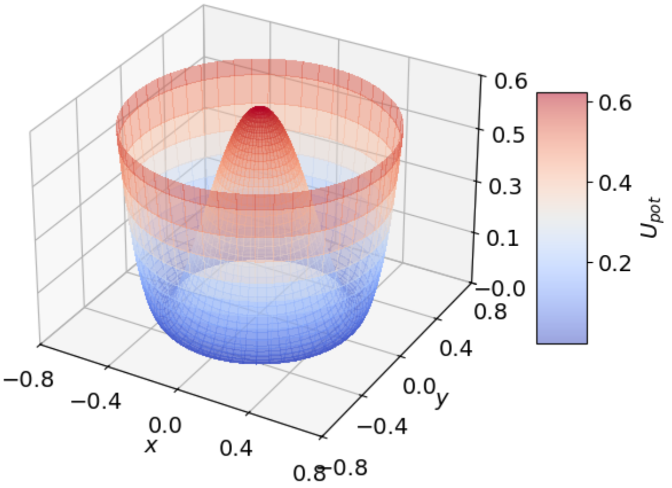

Fig 3: *Potential energy landscape associated to HMC algorithm for $\alpha = 10$. Version for $\alpha = 1000$ is identical, except scales are scaled by a factor of 100.*

3.3.1 Impact of Integration Time

The first parameter we explore is the integration time. For this parameter, we expect the optimal value to be the one that allows the sampler to explore the whole space, without being excessively long (as it would slow down sampling). The similarities associated to 3000 samples for the different integration times are presented in figure Fig.4.

- qualitatively same behaviour : increase of similarity with integration time, followed by a plateau (though more noisy for $\alpha = 1000$ case). - the plateau is reached quicker for the $\alpha = 10$ case, probably because the forces are stronger and the sampler can explore the space more quickly. - good values for the integration time Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

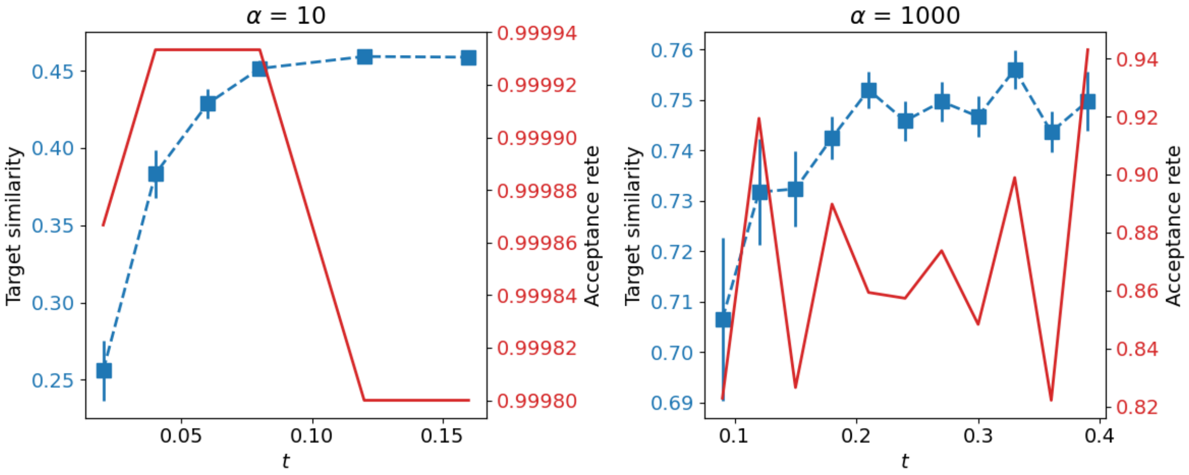

Fig 4: *Similarity as a function of HMC integration time for considered values of α.*

3.4 Impact of Δt

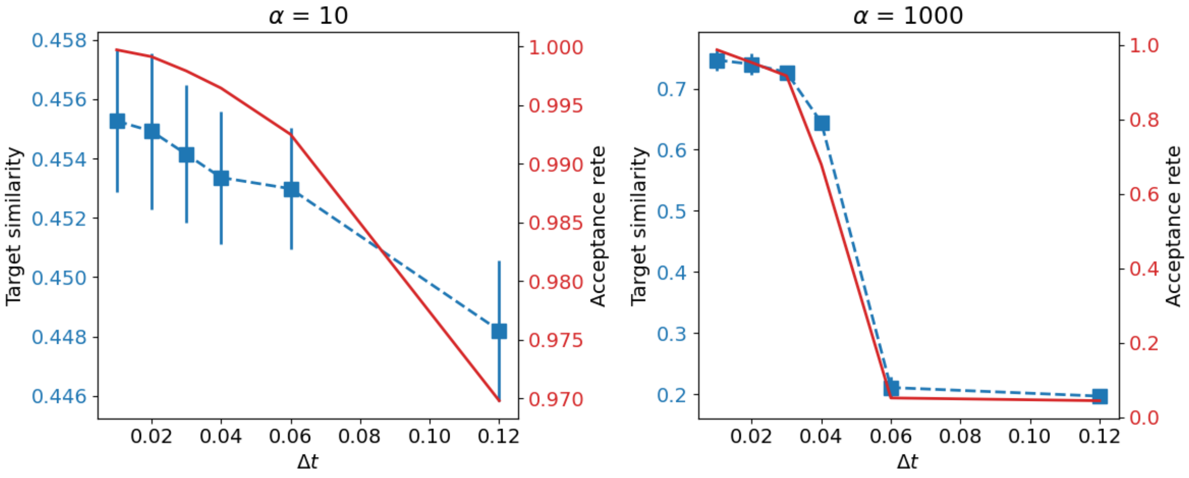

Fig 5: *Similarity as a function of HMC time step for considered values of α.*

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

3.5 Impact of Mass Scale

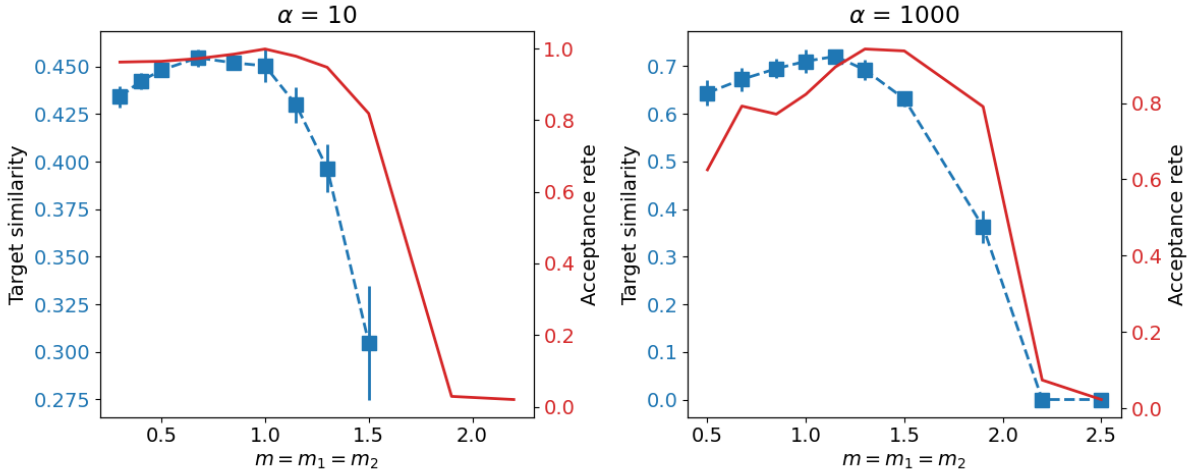

Fig 6: *Similarity as a function of HMC mass scale for considered values of α.*

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

3.6 Impact of Mass Symmetry

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

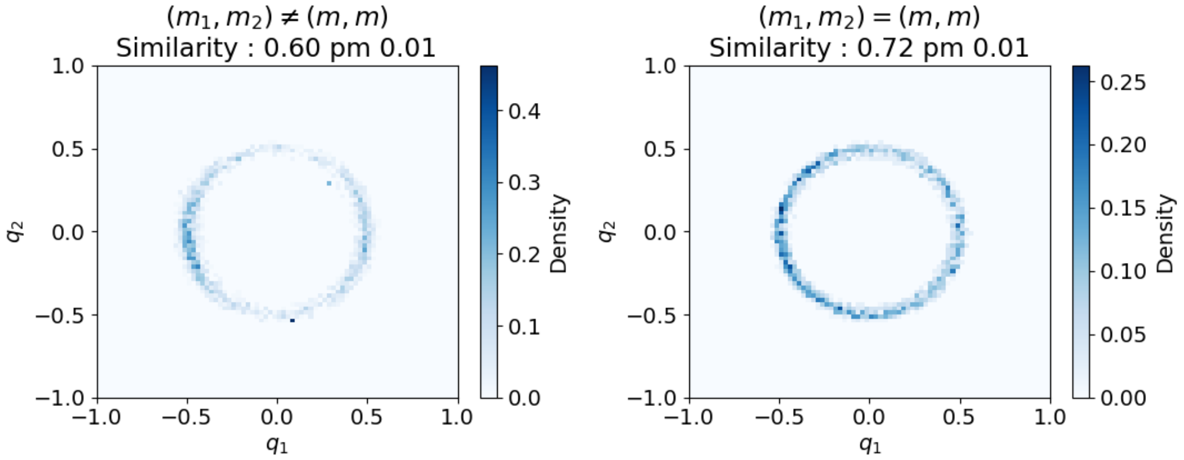

Fig 7: *Similarity for HMC samplers with asymmetric and symmetric masses for* $\alpha = 1000$.

3.7 Comparison and Sample-Size Evolution

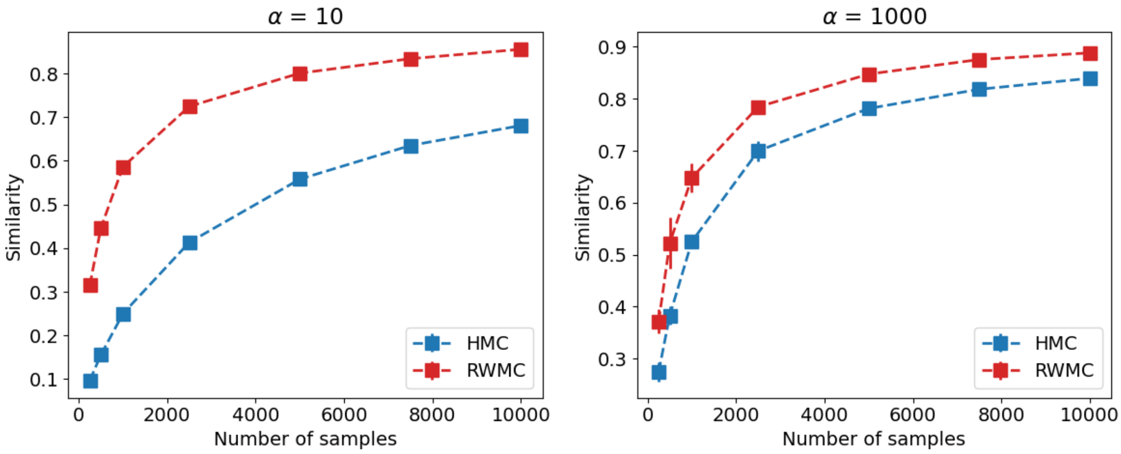

Fig 8: *Similarity as a function of sample-size for RWMC and HMC samplers and considered values of* α. *The sample-size represented on the* x *axis is that associated to the HMC sampler. The associated RWMC sample-size is obtained by matching the number of function evalutations of the HMC sampler.*

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

4 US Birthweight Data

4.1 HMC Solution

4.2 RWMC Solution

5 (f)

6 Section ...

7 Conclusion

Aknowledgements

References

A Commented Code Snippet

[cite stack exchange for format]

```
if transactions: Transaction.create_transactions() # if transactions =
                                                  "true"
node.generate_emptyState() # empty state for all nodes
S.initial_events() # initiate initial events to start with

while not queue.isEmpty() and clock <= targetTime:
  next_e = queue.get_next_event()
  clock = next_e.time # move clock to the time of the event
  Event.execute_event(next_e)
  Queue.remove_event(next_e)

print results
```

B Rejection Sampling Attempt

As an alternative to HMC we consider rejection sampling. We therefore want to find a function $g(q)$ and a constant C such that the following inequality holds for all q:

$$\tilde{f}(q) = e^{q^T X^T (y-1_n)} e^{-1_n^T \log[1+\exp(-x_i^T q)]_{n\times 1}} e^{-\frac{1}{2} q^T \Sigma^{-1} q} \leq Cg(q), \tag{8}$$

where we have denoted $\Sigma = \mathrm{Diag}(\sigma_1^2, ..., \sigma_p^2)$. Given that

$$\log[1+\exp(-x_i^T q)] \leq \log(2) - x_i^T q, \tag{9}$$

$$e^{-\sum_i \log[1+\exp(-x_i^T q)]} = \prod_i \frac{1}{1+\exp(-x_i^T q)} < 1, \tag{10}$$

we can simplify the problem to finding a function $g(q)$ such that

$$\tilde{f}(q) \leq 2^{-n} e^{-q^T X^T 1_n} e^{q^T X^T (y-1_n)} e^{-\frac{1}{2} q^T \Sigma^{-1} q} = 2^{-n} e^{q^T b} e^{-\frac{1}{2} q^T \Sigma^{-1} q} =: Cg(q), \tag{11}$$

with $b = X^T(y - 2_n)$.

By completing the square in the exponent of $Cg(q)$, we can write it in terms of a Multivariate Gaussian distribution with mean $\mu = \Sigma b$ and covariance Σ. Indeed :

$$e^{-\frac{1}{2}(q-\mu)^T \Sigma^{-1} (q-\mu)} = e^{-\frac{1}{2}\mu^T \Sigma^{-1} \mu} e^{q^T \Sigma^{-1} \mu} e^{-\frac{1}{2} q^T \Sigma^{-1} q} \tag{12}$$

$$\implies \tilde{f}(q) \leq 2^{-n} e^{\frac{1}{2}\mu^T \Sigma^{-1} \mu} e^{-\frac{1}{2}(q-\mu)^T \Sigma^{-1} (q-\mu)}. \tag{13}$$

Using now the normalisation constant of the Multivariate Gaussian distribution

$$\sqrt{(2\pi)^p|\Sigma|} = \int_{\mathbb{R}^p} e^{-\frac{1}{2}(q-\mu)^T\Sigma^{-1}(q-\mu)}\ dq, \tag{14}$$

we can define g and C as

$$g(q) = \frac{1}{\sqrt{(2\pi)^p|\Sigma|}} e^{-\frac{1}{2}(q-\mu)^T\Sigma^{-1}(q-\mu)}, \tag{15}$$

$$C = 2^{-n} e^{\frac{1}{2}\mu^T\Sigma^{-1}\mu}\sqrt{(2\pi)^p|\Sigma|} = 2^{-n}\sqrt{(2\pi)^p|\Sigma|e^{\mu^T\Sigma^{-1}\mu}}. \tag{16}$$