Random walk Metropolis Hastings with random step size

Son Luu

1 Introduction

The random walk Metropolis-Hastings (RWMH) is a well known Markov chain Monte Carlo (MCMC) sampler used to sample from complex target distributions when direct sampling is difficult. At each step of RWMH, a new state is proposed from a normal distribution centered at the previous state. The standard deviation of the proposal, also known as the step size, is what determines the effectiveness of the sampler for a given target distribution. Therefore, many works have devised methods to find the optimal step size for RWMH [HST01; Gra11; AT08]. However, these methods only consider deterministic step sizes. This leads me to wonder what behaviour random step sizes would induce for RWMH. Hence, my project will be to investigate the behaviour of RWMH with random step sizes.
In particular, the objectives I consider for this project are as follows:

- Investigate how the proposal behave with random step sizes compared to deterministic step sizes.
- Based on the behaviour of random step size RWMH, Devise a way to tune the distribution for the step size to improve convergence.
- Investigate if RWMH with random step size can outperform RWMH with optimally tuned deterministic step size.

2 Background

In this section, several key concepts for the project are introduced. Throughout this project, let π denote the d-dimensional target distribution of interest defined on $\Theta \subset \mathbb{R}^d$, with density $\pi(\theta) \propto e^{-U(\theta)}$, for $\theta \in \Theta$, with respect to Lebesgue measure on $\mathbb{R}^d$. The goal of Markov chain Monte Carlo (MCMC) algorithms is to generate a Markov chain whose invariant distribution is π. After confirming that the chain has sufficiently converged to π, samples from the chain will then be used to estimate quantities from π such as its mean, variance and other kinds of expectations.

2.1 Random walk Metropolis Hastings sampler

Given a current Markov chain $\{\theta_s\}_{s=0}^{t-1}$, the random walk Metropolis Hastings (RWMH) is a family of MCMC algorithms where, at iteration t, the proposal θ^* is a normal random variable centered at the previous value θ_{t-1} of the chain. The proposed value is then accepted with probability

$$\alpha(\theta_{t-1}, \theta^*) := \min\left\{1, \frac{\pi(\theta^*)}{\pi(\theta_{t-1})}\right\}. \tag{1}$$

The next value in the chain θ_t will be θ^* if it is accepted and remain θ_{t-1} otherwise. The pseudo code for vanilla RWMH is presented in Algorithm 2. For the rest of the project, the RWMH kernel, i.e. the procedure to draw θ_t given θ_{t-1} described in Algorithm 1, is denoted as $K_\Sigma(\cdot|\theta_{t-1})$ where Σ is the $d \times d$ covariance matrix (step size) of the proposal.

2.2 Adaptive Metropolis

In practice, the step size of RWMH is often adapted to improve the convergence of the chain. The adaptive Metropolis (AM) algorithm found in [HST01] is an algorithm that iteratively approximate the covariance structure of the target distribution. The pseudo code for AM is presented in Algorithm 3.

Algorithm 1 RW kernel $K_\Sigma(\cdot|\theta_{t-1})$

function RWKernel(θ, Σ)
  $U \leftarrow \text{Unif}(0, 1)$
  Sample $\theta^* \sim N(\theta, \Sigma)$
  $a \leftarrow \alpha(\theta, \theta^*)$
  if $U \leq a$ **then return** θ^*
  elsereturn θ
  end if
end function

Algorithm 2 Vanilla RWMH

Require: Target density $\pi(\theta)$, initial state θ_0, number of MCMC iterations T, step size Σ
for t **in** 1, 2, ..., T **do**
  $\theta_t \leftarrow \text{RWKernel}(\theta_{t-1}, \Sigma)$
end for
return $\{\theta_t\}_{t=0}^T$

2.3 Effective sample size

For an independent sample of size N, the central limit theorem bounds the uncertainty of estimates based on N. This is no longer the case for a dependent sample such as a Markov chain. Effective sample size (ESS) is an estimate of the size of an independent sample that would have the same uncertainty bound as the dependent sample. For example, the estimate error of the mean is proportional to $1/\sqrt{N}$ for an independent sample of size N and $1/\sqrt{ESS}$ for a dependent sample of the same size. The ESS of a Markov chain is defined as

$$\text{ESS} := \frac{N}{1 + 2\sum_{t=1}^{\infty} \rho_t}$$

where N is the chain length and ρ_t is the autocorrelation of the chain at lag $t \geq 0$. Note that $ESS \leq N$ by definition.

3 Algorithms

3.1 Random step size proposal

To get random step size for the RWMH proposal, the step size Σ is sampled from a known distribution q. Denote this new kernel as

$$E(K_\Sigma)(\cdot|\theta) := \int q(\Sigma) K_\Sigma(\cdot|\theta) d\Sigma$$

Algorithm 3 AM algorithm

Require: Target density $\pi(\theta)$, initial state θ_0, number of MCMC iterations T, initial adaptation parameters μ_0, Σ_0
for t **in** 1, 2, ..., T **do**
  $\theta_t \leftarrow \text{RWKernel}(\theta_{t-1}, \Sigma_{t-1})$
  $\mu_t = \mu_{t-1} + t^{-1}(\theta_t - \mu_{t-1})$
  $\Sigma_t = \Sigma_{t-1} + t^{-1}[(\theta_t - \mu_{t-1})(\theta_t - \mu_{t-1})^\top - \Sigma_{t-1}]$
end for
return $\{\theta_t\}_{t=0}^T$

for all $\theta \in \mathbb{R}^d$. Note that this new kernel still have the target π as its invariant distribution since, for all $\theta' \in \mathbb{R}^d$,

$$\begin{aligned}
\int \pi(\theta)E(K_\Sigma)(\theta'|\theta)d\theta &= \int\int \pi(\theta)q(\Sigma)K_\Sigma(\theta|\theta')d\Sigma d\theta && \text{(definition of } E(K_\Sigma)) \\
&= \int q(\Sigma)\left(\int \pi(\theta)K_\Sigma(\theta|\theta')d\theta\right)d\Sigma && \text{(by Tonelli's theorem)} \\
&= \int q(\Sigma)\pi(\theta')d\Sigma && \text{(invariance of } K_\Sigma) \\
&= \pi(\theta')\int q(\Sigma)d\Sigma = \pi(\theta').
\end{aligned}$$

However, since Σ is a $d \times d$ positive definite matrix, specifying a distribution on all of its entries may prove difficult. Therefore, for the rest of the project, random step size RWMH (RRWMH) will refer to the scheme where the correlation structure of Σ is specified and only the scales of its components are random, i.e.

$$\begin{aligned}
&\forall\ i = 1, \dots, d : \sigma_i \sim q_i, \\
&D = \mathrm{diag}(\sigma_1, \dots, \sigma_d) \\
&\Sigma = D^\top S D
\end{aligned}$$

where $q_i, i = 1, \dots, d$ are specified distributions on $\mathbb{R}_+$, S is the specified correlation structure of Σ and $\mathrm{diag}(\sigma_1, \dots, \sigma_d)$ denote the diagonal matrix with elements $\sigma_1, \dots, \sigma_d$. The pseudo code for RRWMH and its kernel are presented in Algorithms 5 and 4, respectively.

Algorithm 4 RRW kernel $E(K_\Sigma)(\cdot|\theta_{t-1})$

function RRWKernel$(\theta, S, q_1, \dots, q_d)$
    for i **in** 1, 2, ..., d **do**
        Sample $\sigma_i \sim q_i$
    end for
    $D \leftarrow \mathrm{diag}(\sigma_1, \dots, \sigma_d)$
    $\Sigma \leftarrow D^\top S D$
    return RWKernel(θ, Σ)
end function

Algorithm 5 RRWMH

Require: Target density $\pi(\theta)$, initial state θ_0, number of MCMC iterations T, correlation structure S of Σ, scale distributions $q_1, \dots, q_d$
for t **in** 1, 2, ..., T **do** $\theta_t \leftarrow$ RRWKernel$(\theta_{t-1}, S, q_1, \dots, q_d)$
end for
return $\{\theta_t\}_{t=0}^T$

3.2 Adaptation of RRWMH

Just randomly choosing the scale distributions $q_1, \dots, q_d$ is often not a good strategy. Therefore, I propose in this section an adaptation strategy for the scale distributions. For the means of $q_1, \dots, q_d$, we can choose them to be the diagonal elements of the adapted step size Σ_t in AM. For the variances of $q_1, \dots, q_d$, we can choose them to be the variance of the chain running variances. Here the running variance of the i^{th} component of a Markov chain at time t is defined as

$$V_t^{(i)} := \frac{1}{k}\sum_{s=t-k}^{t} \theta_s^2[i] - \left(\frac{1}{k}\sum_{s=t-k}^{t} \theta_s[i]\right)^2 \tag{2}$$

where k is the window size indicating how many past values to include. To get moderately stable moving variance without shrinking it to near zero, I set the window size to 100. In addition, for simplicity and flexibility, the scale distributions are all set to be Gamma distributions. The pseudo code for this adaptive RRWMH scheme is presented in Algorithm 6

Algorithm 6 Adaptive RRWMH

Require: Target density $\pi(\theta)$, initial state θ_0, number of MCMC iterations T, initial adaptation parameters μ_0, Σ_0, initial scale distributions $q_1 = \ldots = q_d = \text{Gamma}(1, 1)$, window size k

$S_0 \leftarrow \text{diag}(\Sigma_0)^{-1}\Sigma_0\text{diag}(\Sigma_0)^{-1}$ ▷ Get correlation structure from covariance

for t **in** $1, 2, \ldots, T$ **do**

  $\theta_t \leftarrow \text{RRWKernel}(\theta_{t-1}, S_{t-1}, q_1, \ldots, q_d)$

  $\mu_t = \mu_{t-1} + t^{-1}(\theta_t - \mu_{t-1})$

  $\Sigma_t = \Sigma_{t-1} + t^{-1}[(\theta_t - \mu_{t-1})(\theta_t - \mu_{t-1})^\top - \Sigma_{t-1}]$

  $S_t \leftarrow \text{diag}(\Sigma_t)^{-1}\Sigma_t\text{diag}(\Sigma_t)^{-1}$

  for i **in** $1, 2, \ldots, d$ **do**

    $m_i \leftarrow \Sigma_t[i, i]$

    if $t > k$ **then**

      $V_i \leftarrow \frac{1}{t-k+1}\sum_{s=k}^{t}(V_s^{(i)})^2 - \left(\frac{1}{t-k+1}\sum_{s=k}^{t} V_s^{(i)}\right)^2$

    else

      $V_i \leftarrow 1$

    end if

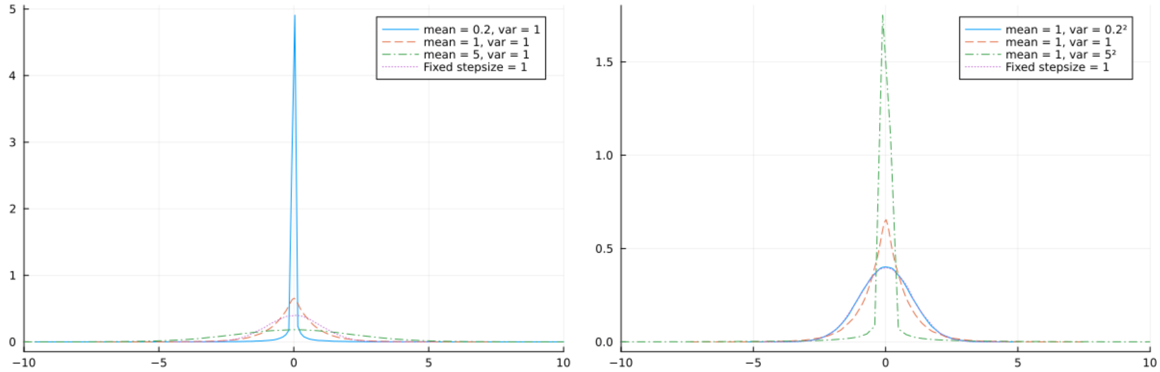
    Choose a_i, b_i so that $\text{Gamma}(a_i, b_i)$ has mean m_i, variance V_i

    $q_i \leftarrow \text{Gamma}(a_i, b_i)$

  end for

end for

return $\{\theta_t\}_{t=0}^{T}$



(a) Varying step size mean.

(b) Varying step size variance.

Figure 1: Proposal behaviour under different step size distributions.

4 Investigation

4.1 Proposal behaviour

I first look at the distribution of the proposal from RRWMH with Gamma scale distributions under various scenarios and summarize key behaviors. Specifically, I generate 10000 samples from proposals with Gamma scale distributions having means and variances in $\{(1, 1^2), (0.2, 1^2), (5, 1^2), (1, 0.2^2), (1, 5^2)\}$. The density plots of these proposals are plotted in Figure 1 along with fixed step size random walk proposals for reference.

Looking at Figure 1a, we can see that the spreads of proposals increase with step size mean similar to the behaviour of fixed step size proposals when we increase the step size. Figure 1b, on the other hand, shows that higher step size variance induces heavier-tailed and more center-concentrated proposals. Based on these results, RRWMH may be beneficial for heavy-tailed or low variance target distributions.

4.2 RRWMH performance

Next, I compare the performance of adaptive RRWMH versus the AM algorithm for different target distributions. The considered targets are as follows

- One-dimensional Normal distributions with means and variances in $\{(2,1^2),(0.2,1^2),(10,1^2),(2,0.2^2),$ $(0.2,0.2^2),(10,0.2^2),(2,5^2),(0.2,5^2),(10,5^2)\}$.
- A standard Cauchy distribution to represent a heavy-tailed target.
- Two-dimensional Normal distributions with mean $(2,3)$, variances $(0.5^2,2^2)$ and correlations in $\{0.1,-0.3,0.5,-0.8\}$.
- Three component Normal mixture

$$\pi = \frac{1}{2}N\left(\begin{bmatrix}1\\2\end{bmatrix},\begin{bmatrix}1 & 0.5\\0.5 & 4\end{bmatrix}\right)+\frac{1}{3}N\left(\begin{bmatrix}-3\\-1\end{bmatrix},\begin{bmatrix}9 & 2\\2 & 1\end{bmatrix}\right)+\frac{1}{6}N\left(\begin{bmatrix}2\\-4\end{bmatrix},\begin{bmatrix}0.25 & -0.15\\-0.15 & 0.49\end{bmatrix}\right)$$

  This target is included to examine how RRWMH handles multi-modal targets.
- Banana distribution [HST01]: a 5-dimensional distribution with the following construction

$$\theta_1 \sim N(0,10^2), \quad \theta_2,\ldots,\theta_5|\theta_1 \sim N(\theta_1^2, 1/10^2)$$

  RWMH samplers are known to have poor performance on this target.
- Logistic regression model with noninformative coefficient priors $N(0,10^2)$. The data comes from a social Network Advertisement dataset from `https://www.kaggle.com/datasets/dragonheir/logistic-regression` with 3 predictors (gender, age and salary). Note that the continuous variables (age and salary) are standardized to get a more well-behaved coefficient posterior.

To evaluate the performance of adaptive RRWMH and the AM algorithm, I compare their median ESS (taken across all dimensions) after 50000 iterations and the algorithm with the higher ESS is deemed to be better. For targets where samples can be drawn exactly, such as the first 4 targets, we can also confirm the validity of the samplers via density plot as well as different moments and quantiles. Note that density plots and ESS calculations are obtained from chains with the first 10000 iterations discarded as burn-in to ensure that the chain has sufficiently converged to the target. In addition, for each target, each algorithm is repeated 10 times to get an estimate of uncertainty.
The ESS of AM and RRWMH for one dimensional targets are plotted in Figure 2. Based on this plot, RRWMH seems to perform better than AM for more concentrated targets but perform worse for more spread out targets and the heavy-tailed Cauchy target. This suggests that the center-concentrated nature of the proposal dominates its heavy-tailedness. Both methods do similarly well estimating the means, variance and each quartile of each target, showing that RRWMH is working properly. The estimation errors for each statistic are shown in Table 1.
For two dimensional Normal targets, the median ESS of RRWMH seems to decrease with the magnitude of the correlation as shown in Figure 3. This trend of deteriorating performance for RRWMH also appear in various statistics estimation as seen in Table 2. I suspect that introducing independent variation to highly correlated targets would go against how these targets want to behave, resulting in less effective chains.
For the remaining three more complex targets, we can see from Figure 4 that RRWMH generally have worse mixing behaviour. For the Normal mixture target, the density plots in Figure 5 shows that RRWMH still produce valid results despite its weaker performance compared to AM as seen in Table 3. For the banana target, however, we can see from Figure 6 that both samplers completely failed to converge to their intended target, which is to be expected. Finally, for the logistic regression problem, both samplers produced quite similar estimates of coefficient statistics as shown in Table 4.

5 Conclusion

In summary, making the step size in RWMH random seems to only make the proposal more center-concentrated while the futher jumps in the proposal tail often get rejected. As a result, RRWMH chains tend to behave similarly to an ordinary RWMH with small step size, which usually performs sub-optimally compared to a well tuned RWMH. I suspect that in order to make a proposal centered at the previous state effective, one needs to somehow incorporate the local geometry information of the target into the proposal, which is hard to achieve with the only random walk proposals.

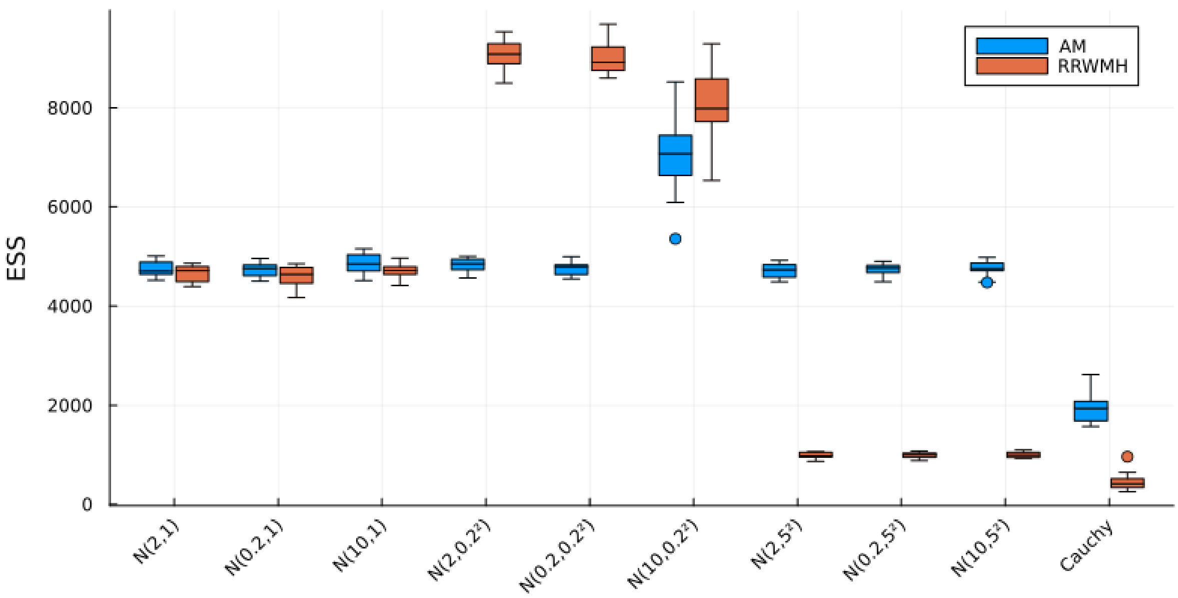


Figure 2: Performance of RRWMH and AM on various 1-dimensional targets.

References

[AT08] C. Andrieu and J. Thoms. "A tutorial on adaptive MCMC". In: *Statistics and computing* 18 (2008), pp. 343–373.

[Gra11] T. L. Graves. "Automatic step size selection in random walk Metropolis algorithms". In: *arXiv preprint arXiv:1103.5986* (2011).

[HST01] H. Haario, E. Saksman, and J. Tamminen. "An adaptive Metropolis algorithm". In: *Bernoulli* 7.6 (2001), pp. 223–242.

Target	**Sampler**	**Mean Error**	**Var Error**	**Median Error**	**Q1 Error**	**Q3 Error**
N(2,1)	RRWMH	4.5×10^{-3}	1.0×10^{-3}	0.7×10^{-3}	0.0×10^{-3}	4.1×10^{-3}
N(2,1)	AM	1.3×10^{-3}	0.6×10^{-3}	1.8×10^{-3}	2.1×10^{-3}	1.1×10^{-3}
N(0.2,1)	RRWMH	1.3×10^{-3}	6.3×10^{-3}	2.4×10^{-3}	4.8×10^{-3}	0.6×10^{-3}
N(0.2,1)	AM	0.1×10^{-3}	3.7×10^{-3}	2.4×10^{-3}	4.5×10^{-3}	3.2×10^{-3}
N(10,1)	RRWMH	4.4×10^{-3}	1.5×10^{-3}	1.4×10^{-3}	2.9×10^{-3}	5.8×10^{-3}
N(10,1)	AM	0.5×10^{-3}	0.7×10^{-3}	1.5×10^{-3}	5.2×10^{-3}	2.6×10^{-3}
N(2,0.2^2)	RRWMH	0.7×10^{-3}	0.3×10^{-3}	0.3×10^{-3}	1.3×10^{-3}	0.3×10^{-3}
N(2,0.2^2)	AM	0.1×10^{-3}	0.1×10^{-3}	0.5×10^{-3}	0.5×10^{-3}	0.1×10^{-3}
N(0.2,0.2^2)	RRWMH	0.8×10^{-3}	0.3×10^{-3}	0.3×10^{-3}	0.7×10^{-3}	0.0×10^{-3}
N(0.2,0.2^2)	AM	0.3×10^{-3}	0.1×10^{-3}	1.0×10^{-3}	1.6×10^{-3}	0.3×10^{-3}
N(10,0.2^2)	RRWMH	0.2×10^{-3}	0.2×10^{-3}	0.5×10^{-3}	0.5×10^{-3}	1.1×10^{-3}
N(10,0.2^2)	AM	0.3×10^{-3}	0.0×10^{-3}	0.6×10^{-3}	0.7×10^{-3}	0.7×10^{-3}
N(2,5^2)	RRWMH	0.09	0.19	0.10	0.06	0.08
N(2,5^2)	AM	0.2×10^{-3}	0.09	3.4×10^{-3}	0.03	3.8×10^{-3}
N(0.2,5^2)	RRWMH	0.08	0.03	0.11	0.07	0.09
N(0.2,5^2)	AM	0.01	0.10	0.01	0.03	6.3×10^{-3}
N(10,5^2)	RRWMH	0.12	0.15	0.13	0.12	0.11
N(10,5^2)	AM	6.5×10^{-3}	0.01	0.8×10^{-3}	0.04	8.7×10^{-3}
Cauchy	RRWMH	-	-	0.04	0.02	0.02
Cauchy	AM	-	-	1.0×10^{-3}	5.2×10^{-3}	6.6×10^{-3}

Table 1: Summary of error metrics by sampler for 1-dimensional targets.

Sampler	**RRWMH**	**AM**	**RRWMH**	**AM**	**RRWMH**	**AM**	**RRWMH**	**AM**
Correlation	0.1	0.1	-0.3	-0.3	0.5	0.5	-0.8	-0.8
Mean error 1	2.3×10^{-3}	2.4×10^{-3}	4.2×10^{-3}	0.6×10^{-3}	0.8×10^{-3}	1.8×10^{-3}	2.4×10^{-3}	0.2×10^{-3}
Var error 1	2.4×10^{-3}	1.0×10^{-3}	2.7×10^{-3}	3.7×10^{-3}	1.8×10^{-3}	3.4×10^{-3}	1.3×10^{-3}	1.0×10^{-3}
Median error 1	2.3×10^{-3}	1.4×10^{-3}	3.1×10^{-3}	4.0×10^{-3}	1.6×10^{-3}	2.5×10^{-3}	1.5×10^{-3}	2.7×10^{-3}
Q1 error 1	4.1×10^{-3}	2.3×10^{-3}	6.8×10^{-3}	0.5×10^{-3}	0.9×10^{-3}	1.6×10^{-3}	2.1×10^{-3}	0.9×10^{-3}
Q3 error 1	1.9×10^{-3}	0.3×10^{-3}	2.5×10^{-3}	1.2×10^{-3}	1.4×10^{-3}	3.8×10^{-3}	2.3×10^{-3}	4.0×10^{-3}
Mean error 2	0.02	3.3×10^{-3}	0.02	2.4×10^{-3}	9.9×10^{-3}	3.0×10^{-3}	0.02	1.0×10^{-3}
Var error 2	0.05	0.02	7.3×10^{-3}	8.9×10^{-3}	0.03	0.02	0.07	3.1×10^{-3}
Median error 2	0.02	5.5×10^{-3}	9.1×10^{-3}	1.2×10^{-3}	0.01	7.5×10^{-3}	5.0×10^{-3}	5.4×10^{-3}
Q1 error 2	0.04	0.01	0.02	9.1×10^{-3}	0.02	0.01	0.03	0.01
Q3 error 2	2.7×10^{-3}	7.0×10^{-3}	3.3×10^{-3}	1.2×10^{-3}	0.7×10^{-3}	1.7×10^{-3}	0.01	1.1×10^{-3}

Table 2: Summary of errors by sampler for 2-dimensional Normal targets.

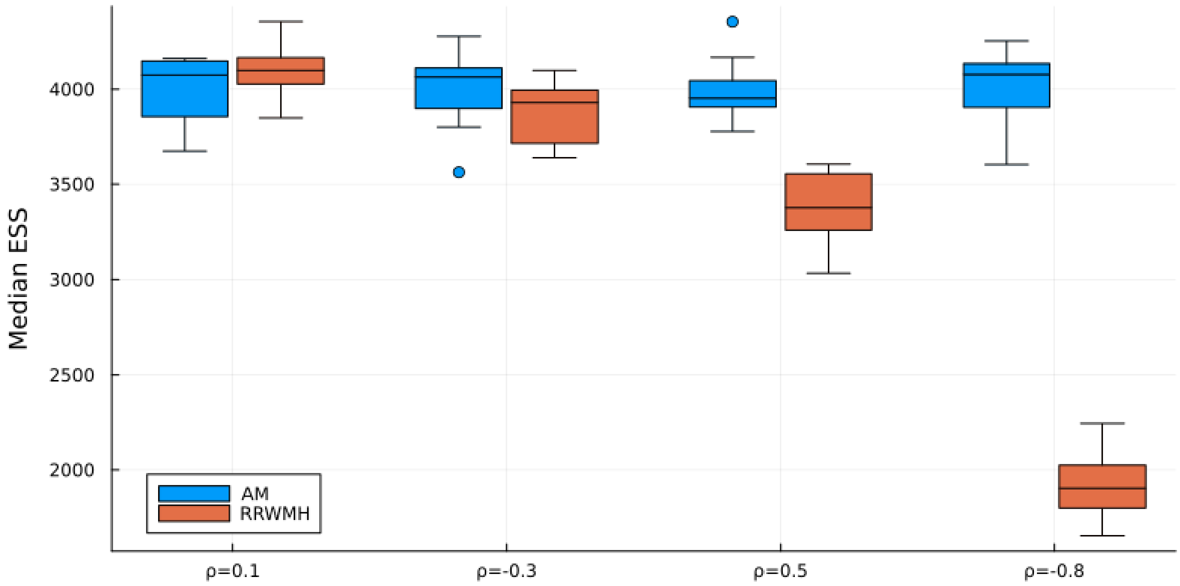


Figure 3: Performance of RRWMH and AM on various 2-dimensional Normal targets.

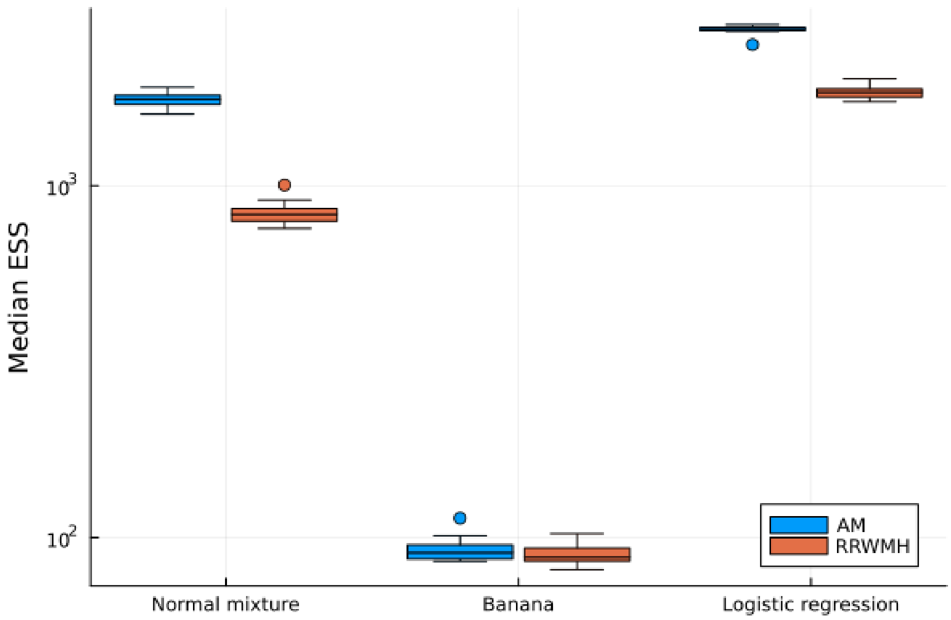


Figure 4: Performance of RRWMH and AM on complex targets.

Sampler	**RRWMH**	**AM**
Mean error 1	0.07	0.01
Var error 1	0.46	0.12
Median error	0.04	5.0×10^{-3}
Q1 error 1	0.20	0.10
Q3 error 1	7.6×10^{-3}	3.4×10^{-3}
Mean error 2	0.04	0.02
Var error 2	0.01	0.03
Median error 2	0.04	0.02
Q1 error 2	0.05	0.02
Q3 error 2	7.8×10^{-3}	0.03

Table 3: Summary of errors by sampler for Normal mixture target.

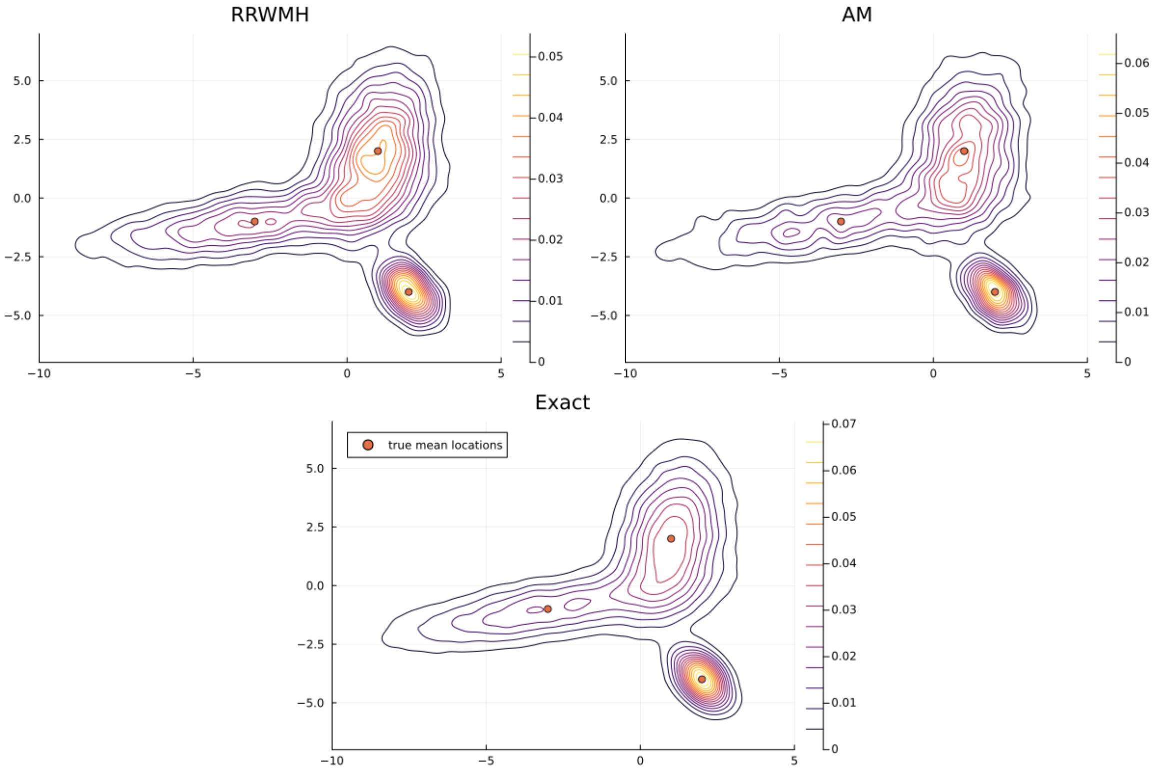


Figure 5: Normal mixture target density plot for RRWMH, AM and exact samples.

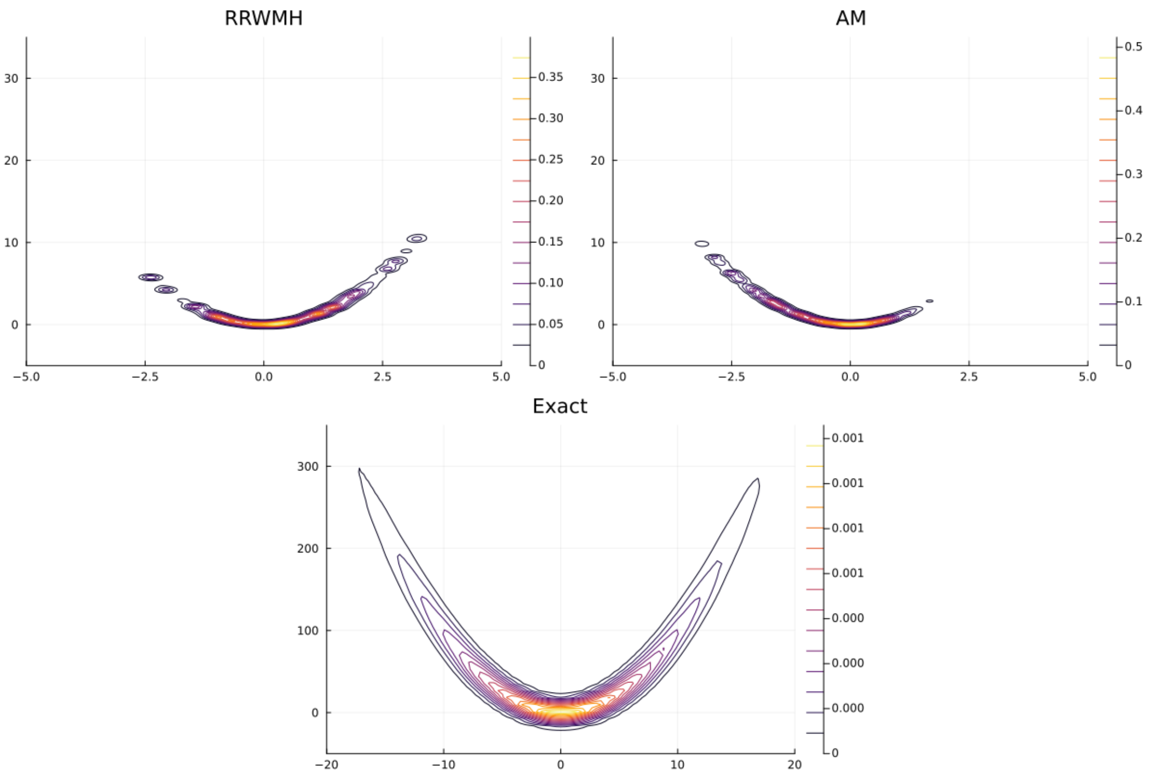


Figure 6: Banana target density plot of the first two dimensions for RRWMH, AM and exact samples.

Sampler	**RRWMH**	**AM**
Intercept mean	-1.00	-1.00
Intercept Var	0.05	0.05
Intercept median	-1.00	-1.00
Gender mean	-0.34	-0.34
Gender var	0.10	0.09
Gender median	-0.34	-0.34
Age mean	2.54	2.54
Age var	0.08	0.08
Age median	2.53	2.53
Salary mean	1.27	1.27
Salary var	0.04	0.04
Salary median	1.26	1.26

Table 4: Posterior coefficient estimations by sampler for logistic regression target.