Adaptive Rejection Metropolis Sampling

Corrà Sara Shaboian Goar
5206191 5217162

1. Problem statement

In computational statistics, it becomes increasingly important to develop sophisticated algorithms that enable addressing issues arising from complex statistical models. Where numerical algorithms fail to yield the necessary result, researchers often turn to approximation algorithms, such as Monte Carlo methods and Markov chains. Pharmacokinetics is a scientific field where addressing complex models is particularly prevalent. It aims at studying how certain drugs are absorbed, distributed and eliminated from the bodies of patients. Due to the considerable number of factors influencing this problem, and the overwhelming presence of noise and measurement error, researchers oftentimes encounter statistical models which are not straightforward to work with. Such situation arose in Wally R. Gilks, Best, and Tan (1995), where, to determine the appropriate dosage of gentamicin for newborns treated for infections, the concentration y_{ij} of gentamicin in the blood flow was modelled as follows:

$$log\ Z_{ij} = log\ E\left[z_{ij} \mid V_i, C_i\right] + \epsilon_{ij},\ where\ z_{ij} = log\left(y_{ij}\right),$$

$$E\left[z_{ij}\right] = log[\sum_{l:\, t_{ij}>s_{il}} \frac{d_{il}}{V_i}\ exp\{-\frac{C_{ij}}{V_i}\ \left(t_{ij} - s\right)]$$

In this framework, C_i, V_i and ϵ are random quantities that are estimated in their turn, with vague prior distributions assigned to the parameters.

The complicated nature of the model in the Bayesian setting resulted in non-log-concave nature of full conditional densities, sampling from which was necessary to perform posterior inference. Thus, researchers had to adapt appropriate methodology: as will be shown below, standard methods fail in approximating non-log-concave densities. For that reason, Adaptive Rejection Sampling algorithm (ARMS) was introduced: the detailed methodology and examples of implementation are described in this report.

2. Rejection sampling

Rejection sampling involves sampling from an easier distribution and adjusting the probability by randomly discarding candidates. The objective is to obtain a sample conforming to the target distribution $f(x)$, which can be only known up to a multiplicative constant. It will be sufficient to sample from an alternative, widely recognized distribution, and subsequently retain solely those values that adhere to the condition (Robert and Casella (1999)):

$$u < \frac{f(y)}{\alpha g(y)}, \qquad U \sim \text{ Unif }(0,1)$$

It follows that the constant α can be interpreted as a measure of the efficiency of the algorithm since the random total number of iterations will depend on the proportion of rejections (Givens and Hoeting (2012)). This may require optimization or a clever approximation to the target function in order to ensure that the envelope can be constructed to exceed the target everywhere.

As an illustrative example, in *Figure* **1** a potential envelope $\alpha g(y)$ is presented, subsequently denoted as $h(x; \alpha)$) for the density function of a Gamma distribution with a shape parameter of 8 and a rate parameter of 1.

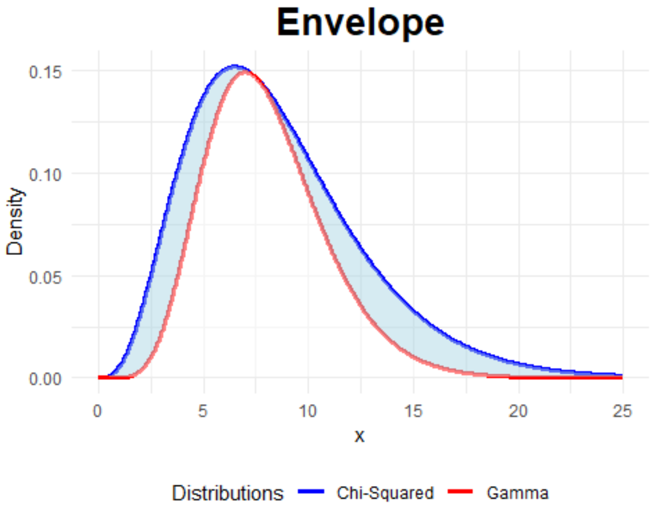

Figure 1: *Chi-Squared envelope for a Gamma distribution*

As mentioned earlier, the decision to accept a particular value is based on the ratio $\frac{f(y)}{\alpha g(y)}$, which, in the given example, can be expressed as follow:

$$\frac{\frac{b^a}{\Gamma(a)} * x^{a-1} * exp - bx}{\alpha * \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}} = \frac{dGamma\ (x;\ \alpha = 8,\ \beta = 1)}{\alpha * dChisq(x;\ k = 8.5)}$$

The parameter α is chosen to be equal to 1.406434 so that the condition $\alpha^* = \frac{f(x)}{g(x)}$ is satisfied and the divergence between the two density lines can reach its minimum. Furthermore, the envelope distribution possesses thicker tails compared to the target distribution in order to ensure comprehensive coverage of the Gamma function across all points.

2.1. Squeezing function

The *Squeeze principle* was proposed by Marsaglia (1977). It states that for s to be a suitable squeezing function, $s(x)$ must not exceed$f(x)$ anywhere on the support of f. The incorporation of this additional function leads to a decrease in the number of evaluations required for f. This arises from the algorithm now incorporating two distinct stages. Firstly, a value is accepted if it falls below the Squeezing function. Only when this condition is not satisfied it becomes necessary to evaluate the function in that point to assess whether the sample value falls under the desired curve (thus is accepted) or not. The proportion of iterations in which evaluation of f is avoided is (Givens and Hoeting (2012)):

$$\frac{\int s(x)\ dx}{\int e(x)\ dx}$$

Furthermore, while the squeezing function always possesses a bounded support, the same cannot be said for the envelope line.

3. Adaptive rejection sampling

Adaptive Rejection sampling approach was introduced by Gilks and Wild (Walter R. Gilks and Wild (1992)) to decrease the proportion of rejections by simultaneously refining the envelope and squeezing function during the generation of sample draws.

$$s(x) \leq f(x) \leq h(x)$$

Two available alternatives for the possible envelope structure encompass the tangent and secant method. For the purpose of this discussion, the focus will be on the secant procedure, which uses secants lines avoids the need for the specification of derivatives (Walter R. Gilks and Wild (1992)). Nevertheless, it is important to acknowledge that this advantage comes at the cost of reduced algorithm efficiency.

By leveraging the property that any concave function can be bounded both from above and below by its secants and chords, the Adaptive Rejection Sampling (ARS) method diminishes the need for time-consuming optimization by reducing the number of function evaluations of the target density (Meyer, Cai, and Perron (2008)). In this framework, log-concavity is defined as follows:

$$\ln \mathrm{f(a)} - 2\ln(\mathrm{b}) + \ln \mathrm{f(c)} < 0, \qquad \forall \mathrm{a,b,c} \in \mathrm{D}, \qquad \mathrm{a < b < c}$$

Fortunately, there is a wide range of functions that are log-concave, making them suitable for application of the ARS method.

3.1.1. Function definitions

Let $S_n = x_i;\ i = 0, \dots, n+1$ define a current set of abscissae in ascending order. Then, for $1 \leq i \leq j \leq n$ let $L_{ij}(x; S_n)$ denote the straight line through points $[x_i, Inf(x_i)]$ and $[x_j, Inf(x_j)]$; for other (i,j) let $L_{ij}(x; S_n)$ be undefined (Wally R. Gilks, Best, and Tan (1995))[1]. Consequently, the piecewise linear function $h_n(x)$ as may be specified as:

$$h_n(x) = min[L_{i-1,i}(x; S_n),\ L_{i+1,i+2}(x; S_n)] \qquad x_i \leq x < x_{i+1}$$

Rejection sampling can be performed with the sampling distribution given by:

$$g_n(x) = \frac{1}{m_n} exp\ h_n(x),\ \ m_n = \int exp\ \ h_n(x)\ dx,$$

where m_n is the normalizing constant of the piecewise exponential distribution.

It follows that the rejection envelope can be identified as an exponential linear function which allows for efficient sampling of values via inversion (Devroye (2006)). Moreover, it can be

[1] Refer to this paper for the algorithm specification.

stated that since each accepted draw is made using a rejection sampling approach, the draws are an independent and identically distributed sample precisely from target distribution f.

An interesting feature of this algorithm is that the set S_n is only updated when $f(x)$ has been previously computed. As the algorithm produces variable $X \sim f(x)$, the envelope and squeezing function become increasingly accurate and, therefore, the number of evaluations of f is reduced. Also, only rejected samples have previously necessitated a target function evaluation, accepted points may have been accepted through the evaluation of the squeezing function. This method will therefore be useful in situations where the evaluation of $f(x)$ is computationally expensive.

3.2. Envelope updating

The plot presented in *Figure* **2** displays the progression of the envelope and squeezing function following the addition of a single value to the sequence in S_n. It is important to note that, in the first plot, for illustrative purposes, all points are assumed to be equidistant.

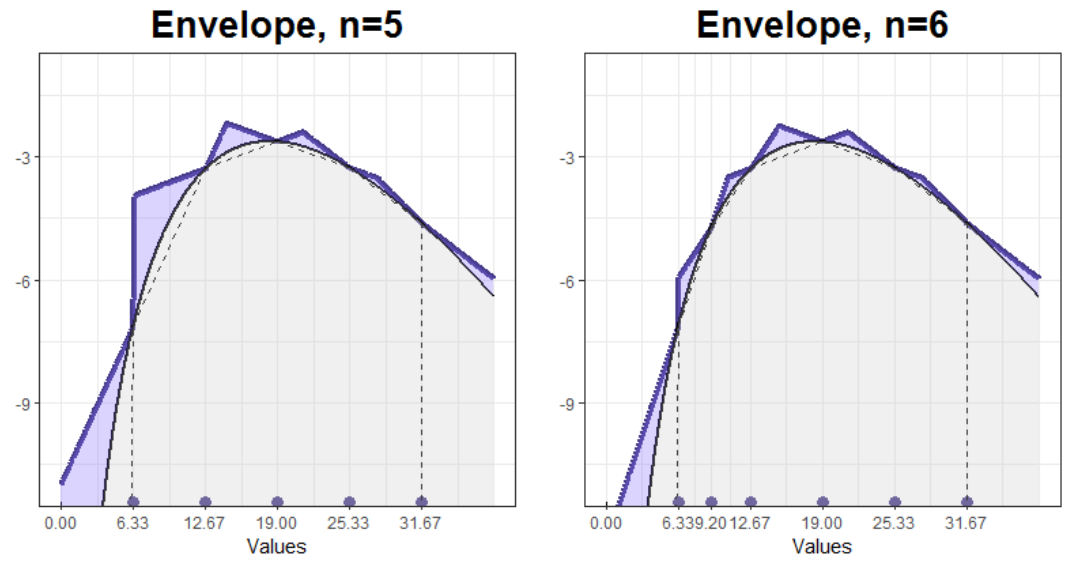

Figure 2: *Envelope and squeezing function with 7 and 8 points in S*

In general, one would prefer the grid (S_n) to be most dense in regions where f(x) is largest,thus near the mode of f. Fortunately, this happens automatically, since such points are most likely to be kept in subsequent iterations and included in updates to S_n.

4. Gamma sampling with ARS

For computations, the *armspp* package in R was used[2]. The function *arms* takes as input the logarithm of the kernel density. In this case, the following holds:

$$log\ f(x) \propto (a-1)log(x) - bx = (8-1)log(x) - x$$

[2] Implements the code used by Gilk for research on Wally R. Gilks, Best, and Tan (1995). C++ code was adapted to R using Rcpp by Michael Bertolacci (Bertolacci (2019))

For this particular example, the recommended practice from Gilks et al. (Wally R. Gilks, Best, and Tan (1995)) was followed by employing four starting abscissae. This approach is typically beneficial, unless the density exhibits an exceptionally high level of concentration.

Moreover, it is crucial to verify that if the support of f extend to $-\infty$, the lower value of the set S_n, namely x_0, must be chosen such that $\ell'(x_0) > 0$ where $\ell = logf(x)$. Similary, if the support of f extends to ∞, x_{n+1} must be chosen such that $\ell'(x_0) < 0$. As a result, the extreme values of the set S_n evaluated at ℓ will inevitably lie before and after the mode of the distribution.

4.1. Comparison of empirical distributions

In *Figure* **3**, the approximation of the two rejection method to the theoretical Gamma distribution is demonstrated. Notably, both lines closely track the actual density line. Hence, the disparity between the two methods becomes most pronounced when assessing their respective efficiencies within the algorithm.

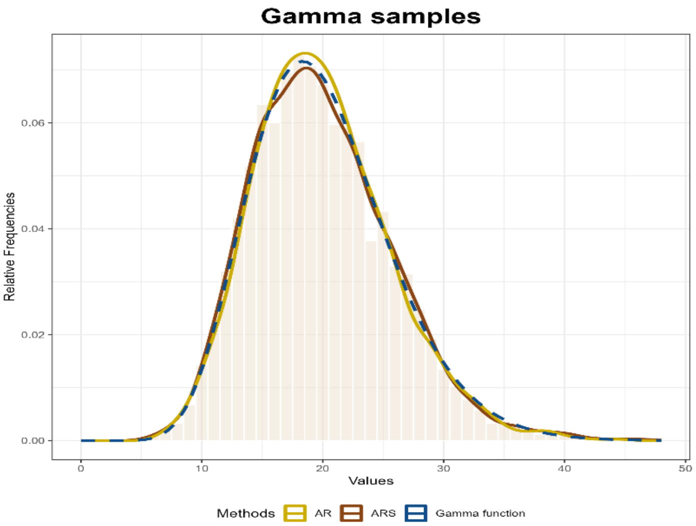

Figure 3: *Comparison of empirical distributions, histogram refers to the ars sample*

5. Adjustment of initial values

In *Figure* **4**, a graphical comparison of two distinct scenarios is presented. The number of evaluations of the target density, which serves as a measure of algorithm efficiency, is represented by both lines. The disparity between the two lines stems from the utilization of different approaches. The blue line corresponds to equidistant starting abscissae, while the red line represents the condition where the initial x-values are adapted to conform to the shape of the target distribution, selected through appropriate quantiles.

The outcomes reveal that opting for suitable starting abscissae consistently leads to a lower number of evaluations. However, it is important to note that there is a considerable level of uncertainty associated with this observation.

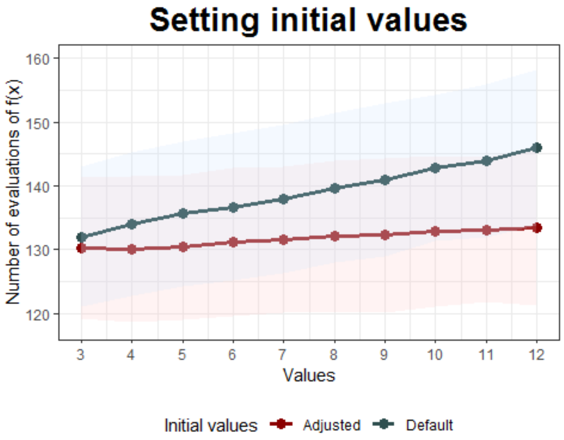

Figure 4: *Comparison of the number of evaluations of the target distribution.*

6. Sampling methods for complex distributions

Markov Chain Monte Carlo algorithms provide tools for drawing inference from vague distributions, for which importance or rejection sampling cannot be applied (Gelman (1993)). The increase in the power of the sampling framework comes at the expense of the independence of the elements in the chain: by construction, Markov Chain algorithms produce a sample in which each value only on its current state.

Metropolis-Hastings sampling is a Markov Chain Monte Carlo algorithm used for tackling high-dimensional parameter vectors by reducing the framework to a sequential sampling from univariate distributions for each component, especially for cases where sampling directly from the target distribution is not feasible (Efron and Hastie (2013), p. 252). To achieve that, a proposal distribution is used to draw samples for target parameters. In the framework of this analysis, this property is of great importance, since Metropolis-Hastings allows to approximate non-log-concave distributions. The decision on whether to accept or reject the new X is determined by the acceptance ratio, which illustrates the difference in the values for the density for the current state and the new proposed state:

$$r = \frac{f(X)\ q(x_i \mid X)}{f(x_i)\ q(X \mid x_i)}$$

The ratio of proposals within the acceptance ratio provides that, if a certain value has a higher probability of being included, the probability would be downweighed to prevent bias in the sample (Hoff (2009), p.183). A special case of the Metropolis-Hastings is the Gibbs sampling algorithm.

However, when sampling multiple parameters from complex density distributions, one instance being non-log-concave densities, the frequent occurrence of rejections at the Metropolis-Hastings step leads to increased computational burden and slow convergence rate. To address this issue, the addition of an adaptive rejection step is proposed to ensure that, for certain values of X sampled from the piecewise exponential distribution, a rejection at the ARS step will adapt the envelope to closer approximate the target distribution, thus guaranteeing faster convergence. The proposed algorithm combines Adaptive Rejection sampling with a Metropolis-Hastings step (ARMS), and is defined as follows:

Table 1. Adaptive Rejection Metropolis algorithm

0. Initialize the starting abscissae and construct the set S_n Note that they must be independent of the current value, X_{cur}			
1. Sample $X \sim g_n(x)$			
2. ARS rejection step	Sample $U \sim Unif(0,1)$	If $U > \frac{f(X)}{exp\, h_n(x)}$, • X is rejected and used to adapt the envelope: • set $S_{n+1} = S_n \cup X$; relabel points in $S_n + 1$ in ascending order; increment n, • go back to sampling $X \sim g_n(x)$;	Otherwise, • Accept the proposed value, $\boldsymbol{X_A = X}$
3. Metropolis-Hastings rejection step	Sample $U \sim Unif(0,1)$	Calculate acceptance ratio: $r = \frac{f(X_A) \min\{f(X_{cur}), exp\, h_n(X_{cur})\}}{f(X_{cur})\, min\,\{f(X_A), exp\, h_n(X_A)\}}$	
		If $U > min(1, r)$ • the proposed value is rejected, set $\boldsymbol{X_M} = \boldsymbol{X_{cur}}$	Otherwise, • set $\boldsymbol{X_M} = \boldsymbol{X_A}$
4. Return $\boldsymbol{X_M}$			

The algorithm described above is applicable to ARMS-within-Gibbs sampling, where the conditioning on other variables being sampled is notationally suppressed. It can be discerned that, in case the density for one of the parameters is log-concave, the Metropolis-Hastings rejection step would always result in "accept" decision, effectively transforming the algorithm into a simple Adaptive Rejection sampling.

For Gibbs and Metropolis-Hasting sampling it can be stated that they are almost always theoretically convergent (Robert and Casella (2010), p. 170), if the proposal distribution meets certain requirements. However, the time required for the convergence to be achieved might be very lengthy: hence, to increase the efficiency, the adaptive-rejection step is introduced. In the ARMS framework, $h_n(x)$ is not strictly an envelope of $log\ f(x)$, due to non log-concavity of the target distribution. The proposal distribution is then set as $q\ (x|X_{cur},\ S_N)\ \alpha\ minf\ (x),\ exp\big(h_n\ (x)\big)$, where S_N are regarded as auxilliary variables (Besag and Green (1993), p. 30) which provides a good way to downweigh sampled values that have a lower probability of rejection by the ARS step.

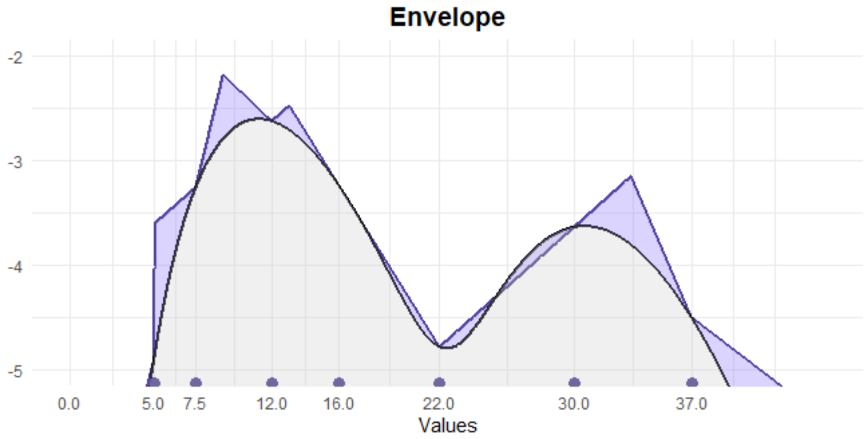

Figure 5: *Envelope for ARMS, 9 starting abscissae*

The piecewise linear function with a non-log-concave distribution is defined as:

$$h_n(x) = max\big[L_{i,i+1}(x; S_n), min\big\{L_{i-1,i}(x; S_n),\ L_{i+1,i+2}(x; S_n)\big\}\big] \qquad x_i \le x < x_{i+1}$$

The squeezing function is not utilized due to non-log-concavity of the target distribution. It should be stressed that the abscissae chosen for constructing the envelope are not dependent on the value for the current state of the chain: hence, no dependency on the envelope is observed in the right-hand side of the definition for the proposal. This indicates that this yields an independence chain, which is irreducible and aperiodic if $g\ (x) > 0,\ where\ f\ (x) > 0$ (Givens and Hoeting (2012), p.204). The dependence on the current state enters in the Metropolis-Hastings step, where the decision is made based on the acceptance ratio, which incorporates the value for the proposal and the density of the current state. This is confirmed by the fact that the Markov transition function $P(X_M \mid X_{cur}, S_N)$, which allows to construct the detailed balance equation:

$$f(X_{cur})\ P(X_M \mid X_{cur},\ S_N) = f(X_M)\ P(X_{cur} \mid X_M,\ S_N),$$

The sampling method allows to obtain X_m, a sample from full conditional of x, thus preserving the stationarity distribution of the chain.

ARMS was applied to a simulated gamma mixture which has non-log-concave density. To demonstrate the gain provided by introducing the adaptive rejection step, the standard Metropolis-Hastings was applied first, along with the ARMS (using *armspp* package in R). The analyzed distribution is a Gamma mixture with two mixture components, with probability density function defined as follows:

$$f(x) = 0.3\ dGamma\ (x;\ \alpha = 10,\ \beta = 0.8) + 0.7\ dGamma\ (x;\ \alpha = 47,\ \beta = 1.5)$$

As proposal distribution for Metropolis algorithm, a symmetric normal distribution centered around the current state of the variable was chosen in order to partially account for the asymmetry of the target distribution. For ARMS, six initial points were chosen, using 5%, 30%, 45%, 55%, 70% and 95% centiles of the previous chain, in accordance with methodology devised by Gilk. For each algorithm, 5000 values were sampled, and the results of the approximation are presented below.

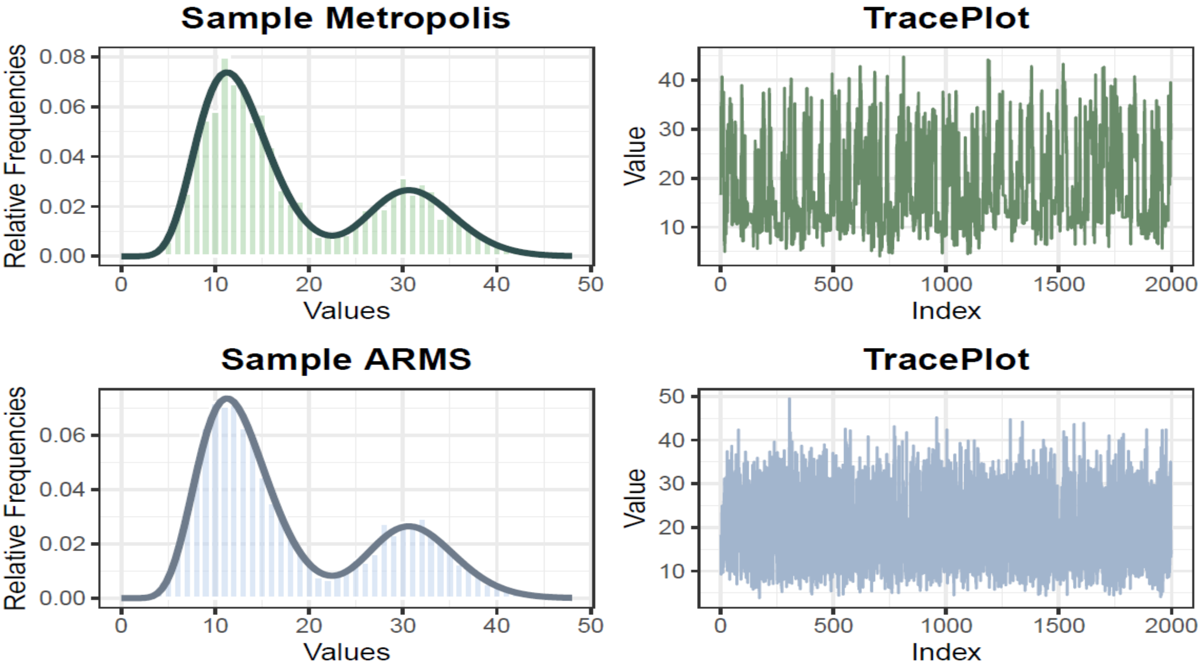

Figure 6: *Histograms and trace plots for Metropolis sampling and ARMS on Gamma Mixture*

Table 2. Autocorrelation values

Method	ESS	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Lag 8	Lag 9	Lag 10
Metropolis	621	0.76	0.59	0.45	0.36	0.27	0.19	0.14	0.11	0.08	0.06
ARMS	5000	0.05	0	-0.01	-0.01	-0.01	0.01	0	0	0.02	-0.01

Only initial 2000 iterations were included in *Figure* **6** trace plots in order to better observe patterns, and to track how the chain behaves in the first iterations. From *Figure* **6** and *Table* 2, it is evident that ARMS algorithm outperforms standard Metropolis algorithm for every diagnostic measure. This outcome is attributed to the fact that, considering the complexity of the distribution, the proposal distribution does not provide good approximation of the target density, contributing to many rejections and, subsequently, higher autocorrelation in the chain. Adaptive Rejection Metropolis sampling, however, effectively adapts the piecewise exponential envelope, tailoring it well to the target distribution, thus creating a chain with low autocorrelation that approximates the target well.

7. Discussion

Adaptive rejection metropolis sampling provides a way to tackle non-log-concave densities, but the efficiency of the algorithm decreases as the severity of non-log-concavity increases, demonstrating a larger number of rejections. Over the recent years, some improvements to the ARMS algorithm have been proposed. One instance is using a mixture of trapezoidal densitial as the proposal, which guarantees fast and efficient sampling (Cai, Meyer, and Perron (2008)). Another possibility to improve the algorithm is to use Lagrange interpolation polynomial of the second degree to build a piecewise quadratic envelope (Meyer, Cai, and Perron (2008)). This method is named ARMS2 and is also extended to non-log-concave densities by approximating non-log-concave parts with linear interpolations. Another procedure implements stochastic variance-reduction using differential equations, and is referred to as gradient Langevin dynamics (Zou, Xu, and Gu (2021)).

In conclusion, it is evident that modern computational statistics methodologies provide a wide range of algorithms to approximate complex distributions. The selection of a suitable algorithm should be guided by careful consideration of several factors, including the specific research task, the complexity of the target distribution to be approximated, and the available computational resources.

8. References

1. Balasubramanian, Krishna, Sinho Chewi, Murat A Erdogdu, Adil Salim, and Shunshi Zhang. 2022. “Towards a Theory of Non-Log-Concave Sampling: First-Order Stationarity Guarantees for Langevin Monte Carlo.” In *Conference on Learning Theory*, 2896–2923. PMLR.

2. Bertolacci, Michael. 2019. *Armspp: Adaptive Rejection Metropolis Sampling (ARMS) via ’Rcpp’*. https://CRAN.R-project.org/package=armspp.

3. Besag, Julian, and Peter J Green. 1993. “Spatial Statistics and Bayesian Computation.” *Journal of the Royal Statistical Society: Series B (Methodological)* 55 (1): 25–37.

4. Bishop, Christopher M, and Nasser M Nasrabadi. 2006. *Pattern Recognition and Machine Learning*. Vol. 4. 4. Springer.

5. Cai, Bo, Renate Meyer, and François Perron. 2008. “Metropolis–Hastings Algorithms with Adaptive Proposals.” *Statistics and Computing* 18: 421–33.

6. Devroye, Luc. 2006. “Nonuniform Random Variate Generation.” *Handbooks in Operations Research and Management Science* 13: 83–121.

7. Efron, Bradley, and Trevor Hastie. 2013. “Computer Age Statistical Inference.” Cambridge University Press 2016, vol. 5.

8. Gelman, Andrew. 1993. “Iterative and Non-Iterative Simulation Algorithms.” *Computing Science and Statistics*, 433–33.

9. Gilks, Wally R, Nicky G Best, and Keith KC Tan. 1995. “Adaptive Rejection Metropolis Sampling Within Gibbs Sampling.” *Journal of the Royal Statistical Society Series C: Applied Statistics* 44 (4): 455–72.

10. Gilks, Walter R, and Pascal Wild. 1992. “Adaptive Rejection Sampling for Gibbs Sampling.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 41 (2): 337–48.

11. Givens, Geof H, and Jennifer A Hoeting. 2012. *Computational Statistics*. Vol. 703. John Wiley & Sons.

12. Hoff, Peter D. 2009. *A First Course in Bayesian Statistical Methods*. Vol. 580. Springer.

13. Maire, Florian, Nial Friel, Antonietta Mira, and Adrian E Raftery. 2019. “Adaptive Incremental Mixture Markov Chain Monte Carlo.” *Journal of Computational and Graphical Statistics* 28 (4): 790–805.

14. Marsaglia, George. 1977. “The Squeeze Method for Generating Gamma Variates.” Computers & Mathematics with Applications 3 (4): 321–25.

15. Meyer, Renate, Bo Cai, and François Perron. 2008. “Adaptive Rejection Metropolis Sampling Using Lagrange Interpolation Polynomials of Degree 2.” *Computational Statistics & Data Analysis* 52 (7): 3408–23.

16. Robert, Christian P, and George Casella. 2010. *Introducing Monte Carlo Methods with r*. Vol. 18. Springer.

17. Robert, Christian P, George Casella, and George Casella. 1999. *Monte Carlo Statistical Methods*. Vol. 2. Springer.

18. Zou, Difan, Pan Xu, and Quanquan Gu. 2021. “Faster Convergence of Stochastic Gradient Langevin Dynamics for Non-Log-Concave Sampling.” In *Uncertainty in Artificial Intelligence*, 1152–62. PMLR.