Metropolis-Hastings (A Markov Chain Monte Carlo Method)

Applications To Linear Regression & Bayesian Estimation

Abstract

Metropolis-Hastings is a computational, statistical method to approximate a probability distribution or approximate an expected value. This method is useful when sampling directly from a distribution is challenging or difficult to get a closed-form solution. In this paper, Metropolis-Hastings is utilized in a Bayesian inference context to sample a posterior distribution and infer parameters from a physical model.

1 The Method [1] [2]

The goal of Metropolis-Hastings (M.H.) is to sample values from a distribution as if the method is exploring the space via random walk. This is done by creating a Markov chain that has a stationary distribution that is exactly the distribution of interest (known but not a closed-form distribution) [1]. Before diving into the theory needed for convergence, the algorithm will be discussed to introduce the reader to the method. In traditional terminology, a probability density that we wish to simulate is known as a target density. Let $\pi(\boldsymbol{x})$ be the target density and $\boldsymbol{x} \in \mathbb{R}^n$. Then, it is necessary to specify a transition density $P(\boldsymbol{y}|\boldsymbol{x}) = q(\boldsymbol{x}, \boldsymbol{y})$ that describes the transition probability from point $\boldsymbol{x}$ to point $\boldsymbol{y}$. This transition density (also called transition kernel) is independent of $\pi(\cdot)$ and will affect the convergence speed of the Markov chain depending on the selection. In practice, $q(\boldsymbol{x}, \boldsymbol{y}) = \mathscr{N}(\boldsymbol{x}, \sigma_q^2 \boldsymbol{I})$ or $q(\boldsymbol{x}, \boldsymbol{y}) = \mathscr{U}(\boldsymbol{x} - \delta_q \boldsymbol{I}, \boldsymbol{x} - \delta_q \boldsymbol{I})$ [2]. For the former, there is a Gaussian distribution around current iterate $\boldsymbol{x}$ and a proposal point $\boldsymbol{y}$ is sampled from it. Whether the point is accepted or rejected is determined by α where $\alpha = \frac{\pi(\boldsymbol{y})q(\boldsymbol{y},\boldsymbol{x}_0)}{\pi(\boldsymbol{x}_0)q(\boldsymbol{x}_0,\boldsymbol{y})}$ [1]. Put differently, the candidate is accepted with probability α. In practice, a uniform random number is drawn from 0 to 1 and if this number is less than α then the candidate point is accepted and $\boldsymbol{x}_1 = \boldsymbol{y}$ otherwise it is rejected and $\boldsymbol{x}_1 = \boldsymbol{x}_0$ [2]. This method is summarized in the pseudo code below.

Algorithm 1 Metropolis-Hastings Algorithm [1]

```
Input: $\boldsymbol{x}_0 \in \mathbb{R}^n, \pi(\boldsymbol{x}), q(\boldsymbol{x}, \boldsymbol{y})$    ▷ initial iterate, target density, transition density
Parameters: $\sigma_q^2, N$    ▷ transition density parameter(s), # of points to collect
  for $k = 0, 1, \cdots, N-1$ do
      $\boldsymbol{y} \sim q(\boldsymbol{x}_k, \boldsymbol{y})$    ▷ get new candidate from some region close to current iterate
      $\alpha \leftarrow \min(1, \frac{\pi(\boldsymbol{y})q(\boldsymbol{y},\boldsymbol{x}_k)}{\pi(\boldsymbol{x}_k)q(\boldsymbol{x}_k,\boldsymbol{y})})$    ▷ probability of acceptance
      $u \sim \mathscr{U}(0, 1)$    ▷ uniform random number to determine whether to accept
      if $u \leq \alpha$ then
          $\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{y}$    ▷ accept current candidate
      else
          $\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k$    ▷ reject current candidate and try again with same iterate
      end if
  end for
  return $\{\boldsymbol{x}_0, \boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$    ▷ return all accepted points
```

Why does this algorithm make sense? Well, assume the transitional probability from $\boldsymbol{x}$ to $\boldsymbol{y}$ is the same as from $\boldsymbol{y}$ to $\boldsymbol{x}$ (i.e. $q(\boldsymbol{x}, \boldsymbol{y}) = q(\boldsymbol{y}, \boldsymbol{x})$). Then, $\alpha = \frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x})}$. This means that if we are moving from a low probability area $\pi(\boldsymbol{x})$ to a high probability area $\pi(\boldsymbol{y})$ then $\alpha > 1$. Thus, regardless of the random number $u \sim \mathscr{U}(0, 1)$, the point $\boldsymbol{y}$ will be accepted. On the other hand, if we are moving from a high probability area to a low one, then $\alpha < 1$ and thus the candidate may be rejected or accepted depending on the ratio. In essence, the algorithm will take many samples from high probability areas and fewer samples from low probability areas. This collection of samples will form the underlying distribution $\pi(\cdot)$.

2 Convergence [3] [4]

For simplicity, assume a symmetric transitional density $q(\boldsymbol{x},\boldsymbol{y}) = q(\boldsymbol{y},\boldsymbol{x})$. We wish to show that the Markov chain $\boldsymbol{X}_i$ (i.e. the accepted samples from the Metropolis algorithm) has a stationary distribution which converges to $\pi(\boldsymbol{x})$. For any Markov chain, the transition matrix can be expressed as $T(\boldsymbol{x}_i,\boldsymbol{x}) = P(\boldsymbol{X}_{i+1} = \boldsymbol{x}|\boldsymbol{X}_i = \boldsymbol{x}_i)$ [3]. This expresses moving from new state $\boldsymbol{x}$ given the current state $\boldsymbol{x}_i$. Suppose $\boldsymbol{x}_i \neq \boldsymbol{x}$ (the new point is accepted), then $T(\boldsymbol{x}_i,\boldsymbol{x}) = P(\boldsymbol{X}_{i+1} = \boldsymbol{x}|\boldsymbol{X}_i = \boldsymbol{x}_i) = P(\boldsymbol{x}|\boldsymbol{x}_i)P(\text{accept } \boldsymbol{x}|\boldsymbol{x},\boldsymbol{x}_i)$ [3]. Of course, the probability of moving state is $q(\boldsymbol{x}_i,\boldsymbol{x})$ by definition and the probability of accepting is $\alpha = \min(1, \frac{\pi(\boldsymbol{x})}{\pi(\boldsymbol{x}_i)})$. Thus, $T(\boldsymbol{x}_i,\boldsymbol{x}) = q(\boldsymbol{x}_i,\boldsymbol{x})\min(1, \frac{\pi(\boldsymbol{x})}{\pi(\boldsymbol{x}_i)})$ [3]. Before moving forward, a new definition is necessary.

Definition 2.1 (Detailed Balance [4]) *$\pi(\cdot)$ is said to satisfy detailed balance equations iff $\pi(a)T(a,b) = \pi(b)T(b,a)\ \forall a,b$. This means that the "mass" moving from a to b is the same as from b to a. Moreover, if $\pi(\cdot)$ satisfies detailed balance condition, then this implies $\pi(.)$ is a stationary distribution since $\pi(b) = \Sigma_a\pi(a)T(a,b) = \Sigma_a\pi(b)T(b,a) = \pi(b)\Sigma_aT(b,a) = \pi(b)$.*

With that said, using the same general notation (a,b), observe that for M.H. the following holds:

$$\pi(a)T(a,b) = \pi(a)q(a,b)\min(1, \frac{\pi(b)}{\pi(a)}) = q(a,b)\min(\pi(a),\pi(b)) =$$

$$q(b,a)\min(\pi(b),\pi(a)) = \pi(b)q(b,a)\min(\frac{\pi(a)}{\pi(b)},1) = \pi(b)T(b,a)$$

Therefore, the Markov chain generated from the Metropolis algorithm satisfies detailed balance property. Thus, by the definition above, the chain has a stationary distribution $\pi(\cdot)$. Not only that, as long as $q(\cdot,\cdot)$ is aperiodic (meaning we will return to starting state after a random number of transitions) then $T(\cdot,\cdot)$ is aperiodic as well [4]. Further, if $q(\cdot,\cdot)$ is irreducible (we can reach any state from arbitrary start) then $T(\cdot,\cdot)$ is irreducible as well [4]. This is stated without proof due to page limitations. With these three conditions met (stationary distribution, aperiodic, and irreducible), the Ergodic theorem for Markov chains can be introduced.

Theorem 2.2 (Ergodic Theorem [4]) *If $(\boldsymbol{X}_0, ..., \boldsymbol{X}_n)$ is an irreducible, aperiodic Markov chain with stationary distribution $\pi(\boldsymbol{x})$, then $\forall\boldsymbol{x},\boldsymbol{x}_0$ the following is known: $P(\boldsymbol{X}_n = \boldsymbol{x}|\boldsymbol{X}_0 = \boldsymbol{x}_0) \to \pi(\boldsymbol{x})$ as $n \to \infty$.*

From the theorem above, the Markov chain constructed from the Metropolis algorithm is said to be an ergodic chain since it has a stationary distribution $\pi(\boldsymbol{x})$ and the chain is both irreducible and aperiodic. Because of this fact, so long as a sufficiently large number of samples are taken, then those accepted samples from the algorithm represent the target density of interest $\pi(\boldsymbol{x})$.

3 Bayesian Inference [2] [5]

Now that M.H. has been discussed, it is time to see how to apply such a method in a Bayesian inference context. Specifically, how to sample a posterior distribution and infer parameters given a likelihood and prior distribution. One interesting topic not covered in class involves posing linear regression problems as likelihood distributions. While the least squares approach works well, it does not quantify uncertainty in the solution which is a disadvantage of a deterministic construction of the problem. To begin, let $y = \boldsymbol{x}^T\boldsymbol{w} + \epsilon$ where $\boldsymbol{x} \in \mathbb{R}^d$, $\epsilon \sim \mathcal{N}(0,\sigma^2)$, and $y,\epsilon \in \mathbb{R}$. This is the linear regression problem for one point which will be extended to multiple points further in this section. From this, the conditional probability distribution (likelihood) can be constructed as follows for a single point: $P(y|\boldsymbol{x},\boldsymbol{w},\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\text{exp}\{\frac{-1}{2\sigma^2}(y - \boldsymbol{x}^T\boldsymbol{w})^2\}$ where $\boldsymbol{x}^T\boldsymbol{w} = \hat{y}$ [5]. Extending to the multi-dimensional case, where the points are expressed as $\boldsymbol{y}$ and $\boldsymbol{X}$ is the "feature matrix", then the likelihood is expressed as [5]:

$$P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w},\sigma^2) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{w},\sigma^2\boldsymbol{I}) = \frac{1}{\sqrt{(2\pi\sigma^2)^N}}\text{exp}\{\frac{-1}{2\sigma^2}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})\}$$

Taking a maximum likelihood approach of the log-likelihood and after some simplification, $\boldsymbol{w}_{\text{MLE}} = \min_{\boldsymbol{w}}(\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}) = \min_{\boldsymbol{w}} \|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2$ (least squares solution) [2]. Moreover, with a similar approach, $\sigma^2_{\text{MLE}} = \frac{1}{N}(\boldsymbol{X}\boldsymbol{w}_{\text{MLE}}-\boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{w}_{\text{MLE}}-\boldsymbol{y}) = \frac{1}{N}\|\boldsymbol{X}\boldsymbol{w}_{\text{MLE}}-\boldsymbol{y}\|_2^2$ (sample average of squared deviation between predictions and training data) [2]. Therefore, the model makes sense with our intuition about the problem. But what if there is prior information about the parameters $\boldsymbol{w}$? In the Bayesian setting, the posterior distribution is expressed as follows [2]:

$$\overbrace{P(\boldsymbol{w},\sigma^2|\boldsymbol{y},\boldsymbol{X})}^{\text{Posterior}} \propto \overbrace{P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w},\sigma^2)}^{\text{Likelihood}}\,\overbrace{P(\boldsymbol{w}|\sigma^2)P(\sigma^2)}^{\text{Prior }=P(\boldsymbol{w},\sigma^2)}$$

The reason why the prior distribution includes belief of σ^2 is because the variance of the noise of the data is usually not known in practice. The Bayesian context requires us to treat σ^2 as a random variable with some unknown distribution. For the numerical experiment, normal and uniform distributions are utilized so that a closed-form solution exists for the posterior distribution. This will be helpful to show convergence of the Metropolis-Hastings algorithm to the true distribution. However, in practice, there may not be a closed-form solution for the posterior (and may even be difficult to calculate) which is where sampling methods such as Metropolis-Hastings prove exceptionally useful. For the next section, a numerical experiment is done to illustrate how this sampling method can be used to sample the posterior in a regression setting.

4 Numerical Results [1] [6] [7]

In this experiment, you are working in a radioactive lab where you received an unknown radioisotope (technician forgot to label it and they are no longer in the building). You know that all radioisotopes created in the lab are calibrated to be 1 Curie in strength (radioactivity). However, you have only received it two hours after it was created. After some thinking, you decide that measuring radioactivity for a few hours should be sufficient to determine the compound so you collect a few samples with an inexpensive Geiger Meter. Unfortunately, there is sensor noise that distorts the measurements so you must estimate in some way.

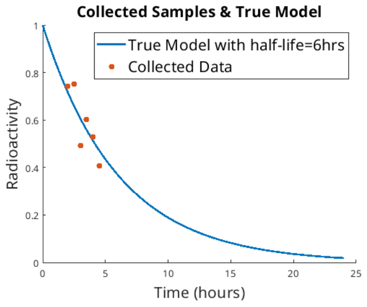

Figure 1: Linear Regression Model.

Being a Bayesianist, you decide to use a linear regression likelihood model. Moreover, you know that the lab only manufactures certain radioisotopes so you use this prior belief to come up with a uniform range for the possible half life values. The goal is to figure out the half life of the compound to be able to identify the radioisotope. Moreover, you read the data sheet of the Geiger Meter and observe that the variance of the noise is reported on the specifications. The radioactive decay model is $N(t) = N_0 \exp -\lambda t$ where λ is related to the half-life of the radioisotope [6]. The posterior distribution is described as $P(\lambda|\sigma^2,\boldsymbol{y},\boldsymbol{X}) \propto \mathcal{N}(\boldsymbol{X}\lambda,\sigma^2\boldsymbol{I})\mathcal{U}(\lambda_{\text{lower}},\lambda_{\text{upper}})$ where $\boldsymbol{y}$ is the recorded data $N(t)$, $\boldsymbol{X}$ is the "feature matrix", and σ^2 is the known variance of the sensor noise. As can be deduced, a Gaussian distribution multiplied by a uniform distribution creates a truncated Gaussian. Since the posterior, likelihood, and prior have closed forms, we can generate graphs for the true distributions as shown below in Figure 4. In Figure 2, M.H. results are shown by running the chain for specific number of iterations. Notice how the histogram of accepted samples converges to the true distribution as collected points range from 10k to 1,000k. This is generated using a fixed step size σ_q^2 for the transition density (normal distribution). In a heuristic manner, the step size σ_q^2 should be large enough so that the entire space can be explored (there may be islands of high probability separated from each other). Yet, small enough to not skip over high probability regions. A good rule of thumb (heuristic) is that σ_q^2 should be the right size such that the acceptance rate of candidates is around 20-80% of the total number of points. Figure 3 illustrates the effect of different step sizes on the histogram of the collected samples under fixed # of points. Notice the acceptance rate changes depending on the step size selected. Moreover, the specific step size value depends on the application and noise variance of the sensor or system. One may be tempted to say that the rightmost plot of Figure 3 is the best one but recall that the posterior distribution for this problem is unimodal. Under a different problem, the target density may be multimodal and thus the super small step size may lead to the chain being incapable of leaving the “island”.

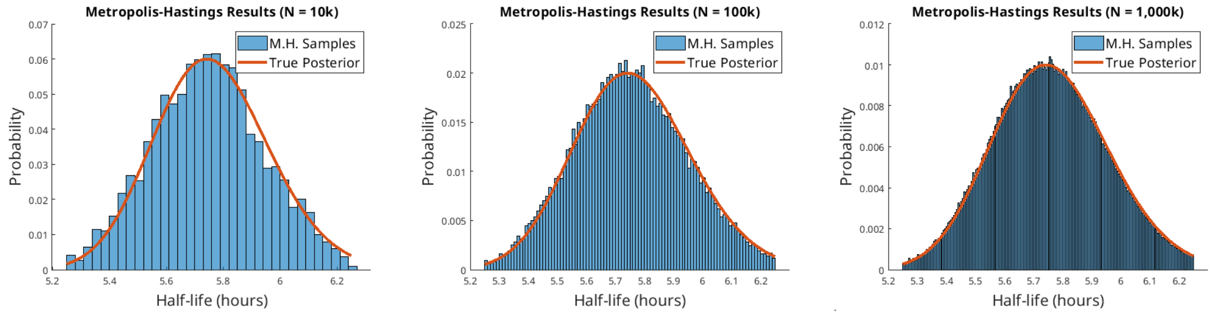

Figure 2: Metropolis-Hastings convergence plots (N = 10k, 100k, 1,000k).

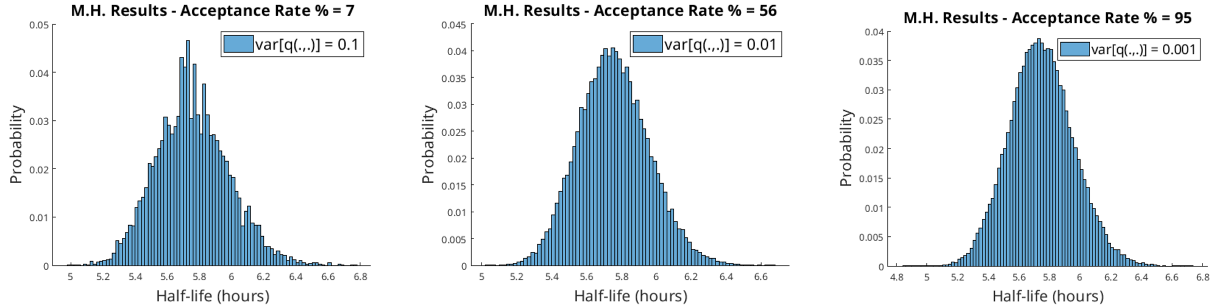

Figure 3: Transitional density parameter σ_q^2 = 0.1, 0.01, and 0.001 under fixed 100k samples.

This method has two main disadvantages. Firstly, the samples are correlated (for any set of nearby samples) even though the entire set of points does follow the target density [7]. This is not ideal since the number of “good” samples taken might be lower than N (the total number of accepted points) due to the correlation effect. This problem does not exist for rejection sampling methods where transition kernels do not exist [7]. Secondly, as shown in the convergence section, the stationary distribution will converge to the target density eventually. However, if the initialization is bad, then the initial samples will not follow the target density. This is known as the “burn-in period” and in practice the first X points are thrown away since the initial points would follow a different distribution [1]. As for the advantages of this method, one key point is that it does not experience the “curse of dimensionality” to the same degree as rejection sampling methods and so M.H. is ideal for high-dimensionality models [1]. This is because traditional rejection sampling techniques will increase the probability of rejection as the dimension increases whereas M.H. algorithm can adjust the transition kernel parameters to have a reasonable rejection rate.

5 Conclusion

In this paper, the Metropolis-Hastings algorithm is discussed as a statistical method to sample distributions. This algorithm is applied to an estimation problem to sample a posterior distribution. In this case, linear regression is proposed within a probabilistic framework where the regression parameter becomes a statistical parameter in the Bayesian context. This regression problem is applied to a physics/chemistry application that utilizes the radioactive decay model for radioactive compounds to determine the half-life value. Additionally, convergence guarantees for the Markov chain are discussed for this computational method to the target density. In summary, this Markov chain Monte Carlo technique is useful for high-dimensional distributions from which direct sampling is difficult or no closed-form distribution exists.

6 Appendix A: Probability Distributions

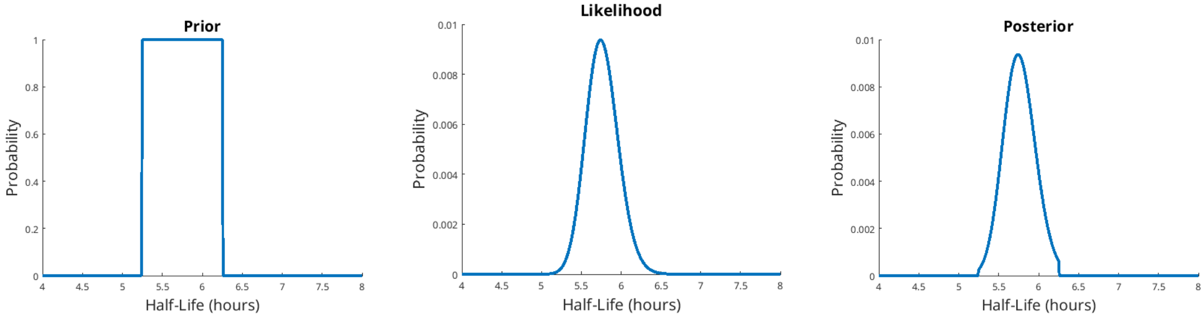

Figure 4: Prior, likelihood, and posterior distributions used in the numerical results section.

7 Appendix B: Predictive Modeling

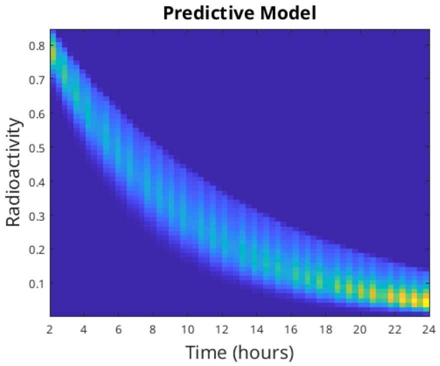

Figure 5: Predictive Model.

Lastly, in Figure 5, under a larger sensor noise situation, one can observe what the radioactive decay would look like for future time points. Since the collected samples from M.H. are the possible half-life values, it is possible to propagate this to the decay model $N(t) = N_0 \exp\{-\lambda t\}$ to see the possible spread of future observations if the experiment were to continue. In the image, low probability areas are purple and blue whereas higher probability areas are yellow and orange. Of course, as the number of observations increase then the spread should decrease as the linear regression solutions become closer to the true solution.

8 Appendix C: Matlab Code

See next page for latex-style pdf printout.

References

[1] S. Chib and E. Greenberg, "Understanding the metropolis-hastings algorithm," *The american statistician*, vol. 49, no. 4, pp. 327–335, 1995.

[2] B. P. Carlin and T. A. Louis, *Bayesian methods for data analysis*. CRC Press, 2008.

[3] K. L. Mengersen and R. L. Tweedie, "Rates of convergence of the hastings and metropolis algorithms," *The annals of Statistics*, vol. 24, no. 1, pp. 101–121, 1996.

[4] S. D. Hill and J. C. Spall, "Stationarity and convergence of the metropolis-hastings algorithm: Insights into theoretical aspects," *IEEE Control Systems Magazine*, vol. 39, no. 1, pp. 56–67, 2019.

[5] J. Gareth, W. Daniela, H. Trevor, and T. Robert, *An introduction to statistical learning: with applications in R*. Spinger, 2013.

[6] S. B. Patel, *Nuclear physics: an introduction*. New Age International, 1991.

[7] W. R. Gilks and P. Wild, "Adaptive rejection sampling for gibbs sampling," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 41, no. 2, pp. 337–348, 1992.