Exploring the 3s Radial Probability Distribution of the Hydrogen Atom Using Custom and Ensemble MCMC Methods

Kevin Mesta, Sara Talebi
Department of Chemistry, Syracuse University, Syracuse, New York 13244 USA

November 2024

Introduction

Understanding the radial probability distribution of atomic orbitals is essential for describing the spatial behavior of electrons in atoms. In this study, we focus on the 3s orbital of the hydrogen atom, which exhibits a distinct radial probability distribution with multiple peaks due to quantum mechanical effects. The statistical challenge involves approximating this distribution by sampling the probability density function, which describes the likelihood of finding an electron at a given distance from the nucleus.

To achieve this, we employ two Markov Chain Monte Carlo (MCMC) sampling techniques:

1. A custom Metropolis-Hastings MCMC algorithm, designed with multiple walkers and a proposed step size tuned to explore the 3s radial distribution space efficiently.

2. The `emcee` ensemble sampler, an advanced MCMC method utilizing multiple walkers to explore the distribution space in parallel, enhancing convergence and reducing autocorrelation.

This study aims to evaluate the effectiveness of these two methods in approximating the theoretical 3s radial distribution and compare their performance in terms of accuracy, convergence, and efficiency. The theoretical radial probability function of the 3s orbital serves as the target distribution, allowing for direct comparisons between the sampled and expected distributions.

This analysis provides insights into the applicability of these MCMC methods for quantum mechanical problems. It highlights the practical considerations involved in designing and optimizing custom MCMC algorithms for complex distributions.

Radial Probability Function

The theoretical radial probability density function for the 3s orbital is given by:

$$P(r) = r^2 \left[N\, R_{3s}(r)\right]^2 / a_0^2 \tag{1}$$

Where a_0 is the Bohr radius, N is the normalization factor of radial function, and $R_{3s}(r)$ is the radial function of the 3s level of the hydrogen atom:

$$N = \left(\frac{1}{81\sqrt{3\pi a_0^3}}\right) \tag{2}$$

$$R_{3s}(r) = N\left(27 - 18\frac{r}{a_0} + 2\left(\frac{r}{a_0}\right)^2\right) e^{\left(-\frac{r}{3a_0}\right)} \tag{3}$$

The function 1 is used as the target distribution for both MCMC methods.

MCMC Sampling Techniques

Two MCMC sampling methods were employed to approximate the radial probability distribution:

0.1 Custom Metropolis-Hastings MCMC

The custom MCMC follows the Metropolis-Hastings algorithm and uses multiple "walkers" starting at a random position given by random uniform distribution from some lower and upper bound. Each walker steps through the space in parallel to one another, meaning that during each iteration of the custom MCMC, each walker moves to a new position before starting another iteration. The new positions are proposed according to a Gaussian distribution with a mean of 0 and σ equal to a_0 multiped by some step size parameter. This specific distribution was chosen due to its similar shape to the ideal 3s radial probability distribution and a greater probability of smaller steps than larger abrupt jumps. Additionally, this distribution's symmetry is similar to the symmetry of different regions of the 3s radial distribution. The acceptance probability is calculated based on the ratio of the probability densities at the new and current positions and accepts the new position according to the Metropolis-Hastings algorithm. The custom MCMC returns a list of the positions of each walker throughout the simulation that can be processed by other functions, as well as a method to discard a percentage of initial samples as the MCMC burns in.

0.2 *emcee* Ensemble Sampler

The `emcee` library's ensemble sampler utilizes multiple "walkers" to explore the distribution space more efficiently. Each walker operates in parallel, generally resulting in better convergence and mixing. The `emcee` sampler was configured with 50 walkers, each starting near $5a_0$, and ran for 10,000 steps with a burn-in of 20%.

Results

1 Comparison of Sampling Methods

The trace of the first walker in the custom MCMC method is shown in Figure 1. This plot illustrates the radial distance sampled by the walker over 10,000 iterations. Initially, the walker starts at a high value (far from the target distribution) but quickly moves toward the correct sampling region. After the initial burn-in period, the walker stabilizes and begins exploring the target distribution effectively. This behavior highlights the importance of discarding the burn-in period to ensure the chain is sampling from the correct region of the probability space.

The posterior samples of radial distance obtained from the custom MCMC method are shown in Figure 2. The histogram of the sampled radial distances closely matches the theoretical 3s radial distribution, demonstrating the effectiveness of the custom MCMC approach. The peaks and troughs of the theoretical curve align well with the sampled distribution, indicating that the MCMC method captures the shape and features of the 3s radial probability distribution.

In the case of the custom MCMC, when the walkers are initialized at positions exceedingly far from the region in which the radial proba-

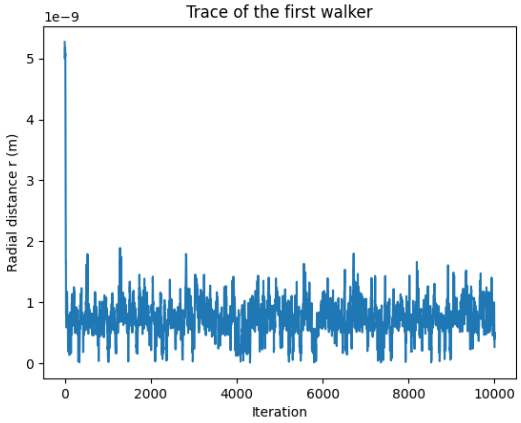

Figure 1: Trace of the first walker in the custom MCMC method, showing the radial distance sampled over 10,000 iterations. The walker stabilizes in the target sampling region after an initial burn-in period.

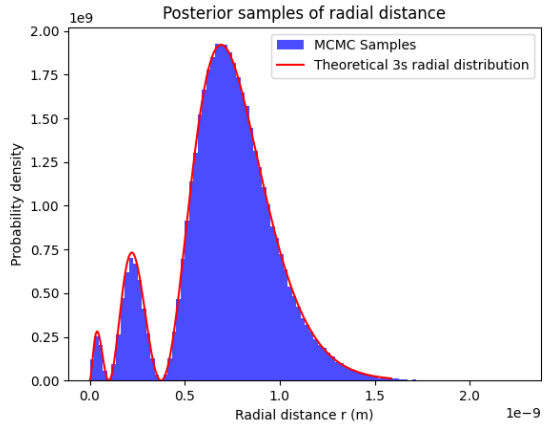

Figure 2: Posterior samples of radial distance obtained from the custom MCMC method. The histogram closely matches the theoretical 3s radial distribution, validating the accuracy of the sampling process.

bility density function is defined, an observable burn-in period is seen. This is because it requires some number of iterations before the walkers are properly exploring the space they are meant to explore. This can be seen in the figure below, where the first 2000 iterations are used to reach the desired space.

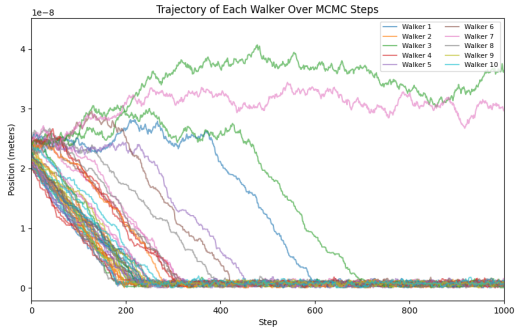

Figure 3: Trace of 50 walkers in the custom MCMC with a step size of 5, lower bound of 200e-10, upper bound of 250e-10, ran for 10,000 iterations and no initial positions discard, burn-in of 0.

Plotting the posterior samples from this set of positions would skew the distribution and produce a plot that is incoherent and not representative of the desired probability distribution. It is then required to discard some fraction of the initial samples, referred to as the burn-in period, to produce the correct probability density graph. Additionally, allowing the MCMC to run for a longer duration also improves the sampling quality since it allows for even the walkers that didn't reach the target space, the few lines above all the other ones, to eventually travel to the target space. The following trace plot is obtained by employing both of these methods, discarding the burn-in period of samples and running the

MCMC for a longer time.

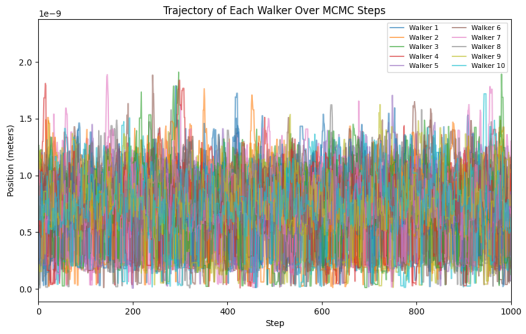

Figure 4: Trace of 50 walkers in the custom MCMC with a step size of 5, lower bound of 200e-10, upper bound of 250e-10, ran for 100,000 iterations and burn-in of 0.2

In this figure, all of the walkers are now sampling the correct space, and only the samples in this burned-in region will be used to construct the probability density plot.

Figure 5 shows a direct comparison between the custom MCMC and `emcee` ensemble sampler. Both methods produce histograms that align closely with the theoretical 3s radial distribution. However, there are some subtle differences:

- The `emcee` samples (green) exhibit slightly better alignment with the theoretical curve around critical peaks and troughs, suggesting better convergence and reduced bias.
- The custom MCMC method shows a slight deviation in certain regions, likely due to a combination of suboptimal proposal step size and longer autocorrelation times.

While both methods are effective, the ensemble approach of `emcee` offers better mixing and faster convergence, as evident from the smoother and more consistent alignment with the theoretical distribution.

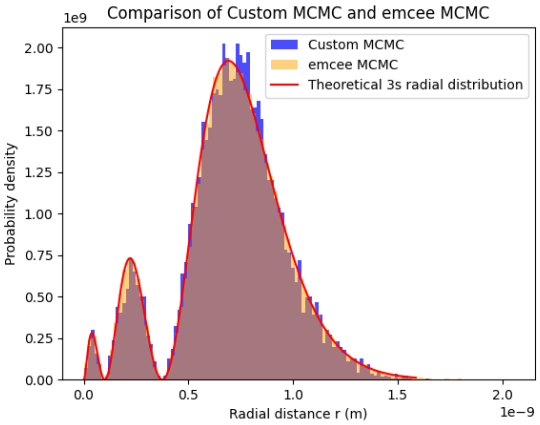

Figure 5: Comparison of histograms from the custom MCMC and `emcee` methods. Both methods align well with the theoretical 3s radial distribution, but the `emcee` samples show slightly better convergence.

2 Average Autocorrelation Length vs. Burn-in Period

The average autocorrelation length (for custom MCMC) decreases significantly as the assumed burn-in period increases, as shown in Figure 6. This trend demonstrates that discarding initial samples (burn-in) allows the Markov chain to move closer to the target distribution, resulting in more effective sampling. Lower autocorrelation lengths indicate better chain mixing, with less correlation between consecutive samples. Based on the results, a burn-in period of approximately 15–20% minimizes autocorrelation length, with further increases providing

diminishing returns.

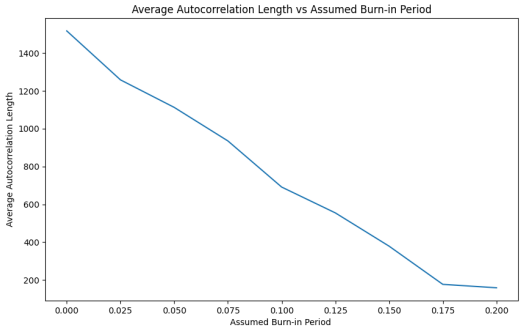

Figure 6: Average autocorrelation length as a function of the assumed burn-in period. A burn-in period of 15–20% significantly reduces autocorrelation length.

The performance of the `emcee` sampler was further quantified using its autocorrelation time and effective sample size. With 50 walkers and 100,000 iterations:

- **Autocorrelation Time:** The autocorrelation time was approximately 33.23 steps. This indicates that after every 33 iterations, the samples become effectively independent. The relatively low autocorrelation time highlights the efficiency of `emcee` in exploring the target distribution.
- **Effective Sample Size:** Based on the autocorrelation time, the effective sample size was estimated to be approximately 120,374. This large effective sample size demonstrates the ability of `emcee` to produce a substantial number of independent samples despite the presence of correlations in the chain.

These metrics confirm the efficiency of the `emcee` ensemble sampler in achieving faster convergence and better mixing compared to the custom MCMC method.

3 Convergence Statistic vs. Burn-in Period

There are many approaches to determine whether an MCMC simulation has converged, some qualitative and some quantitative. Qualitatively, a chain converges once the walkers' positions consistently oscillate within the simulation's target region. This behavior can be seen in Figures 1 and 4, where the walkers travel within the same range of values after a burn-in period for the rest of the simulation. However, in order to better understand the convergence of the chain, the Gelman-Rubin statistic, R, can be calculated for a quantitative comparison of different simulations. In practice, R values of less than 1.1 indicate that the chain has converged and that samples can safely be drawn from the chain to represent the desired probability distribution. As the chain continues to run for longer and longer iterations or as the between-chain variance decreases, the R statistic approaches 1. When comparing the R-value for the custom MCMC and emcee sampling methods for 50 walkers, 100,000 iterations, a step size of 5 and a range of positions from 0 to 10e-10, after discarding the first 20 percent of samples, the R statistic for the custom method is 1.00009964 vs 1.00016111 for the emcee method. Both of these methods are extremely close to 1, which further supports numerically that the chain has converged using both methods.

The convergence statistic approaches the ideal value of 1.0 as the assumed burn-in period increases, as shown in Figure 7. This indicates that the Markov chain samples from the target

distribution more accurately when an appropriate burn-in period is used. Without a burn-in period, the chain starts in suboptimal regions, leading to slower convergence. Discarding 15–20% of the initial samples appears sufficient to ensure proper convergence.

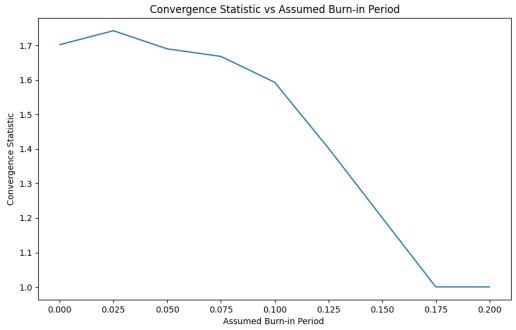

Figure 7: Convergence statistic as a function of the assumed burn-in period. The statistic approaches 1.0 as the burn-in period increases, indicating improved convergence.

4 Optimizing the Custom MCMC Method

The performance of the custom MCMC method can be significantly improved by addressing key aspects of the sampling process, including the burn-in period, total run length, and proposal step size. These parameters directly affect the quality of the samples, convergence rate, and method efficiency. The following optimization strategies are proposed based on the analysis of autocorrelation length and convergence statistics:

4.1 Burn-in Period

The burn-in period represents the fraction of initial samples discarded to ensure the Markov chain stabilizes in the region of the target distribution. As shown in Figures 6 and 7, increasing the burn-in period reduces autocorrelation length and improves convergence. Based on the results:

- A burn-in period of approximately 15–20% of the total iterations is recommended.
- This range ensures that samples used for analysis represent the target distribution while minimizing the influence of the chain's initial starting points.

4.2 Total Run Length

The total number of iterations in an MCMC run must be sufficient to achieve high-quality sampling, especially after discarding the burn-in period. With too few iterations, the results may suffer from insufficient exploration of the target distribution. Based on the analysis:

- A total run length of 50,000–100,000 iterations is suggested.
- This ensures enough valid samples remain after discarding the burn-in, providing robust statistical estimates and a smooth representation of the 3s radial probability distribution.

4.3 Proposal Step Size

The proposal step size determines how far the chain moves in the parameter space during each iteration. Poorly chosen step sizes can result in inefficient exploration, either by causing the

chain to "wander" too slowly (small step size) or by rejecting too many proposed moves (large step size). To improve sampling efficiency:

- The step size should be tuned to balance exploration and acceptance rates. Typically, an acceptance rate of 20–50% is ideal for most MCMC problems.

- Further adjustments to the step size could reduce autocorrelation length, leading to a more efficient sampling process.

By implementing these strategies:

- The custom MCMC method will achieve lower autocorrelation between samples, resulting in better mixing of the Markov chain.

- Convergence to the target 3s radial probability distribution will be faster and more accurate.

- The overall efficiency of the sampling process will improve, making the custom MCMC approach a more competitive alternative to ensemble methods such as `emcee`.

Specifically, the autocorrelation length and convergence statistic R were plotted as a function step size in order to determine the optimal step size for running the custom MCMC. As shown in both figures below, as the step size increases, there is a clear decrease in autocorrelation length, indicating that more of the samples are independent of one another, and a decrease in R, showing that the chains converge more accurately.

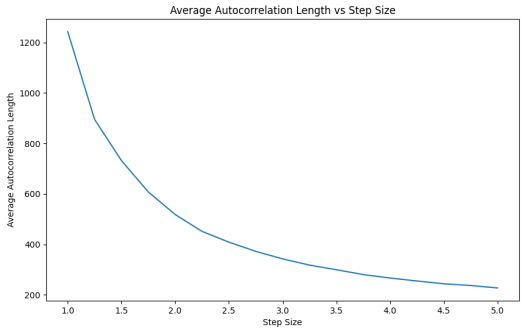

Figure 8: Convergence statistic as a function of the step size period. The statistic approaches 1.0 as the step size increases, indicating improved convergence.

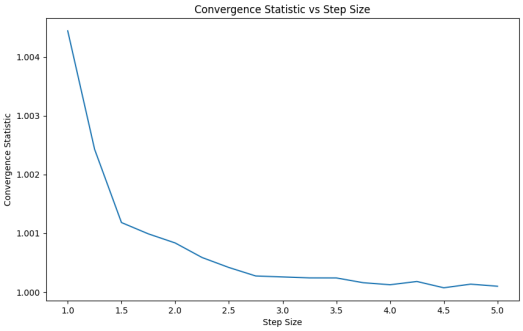

Figure 9: Convergence statistic as a function of the step size. The statistic approaches 1.0 as the step size increases, indicating improved convergence.

Discussion

The results indicate that both MCMC approaches successfully approximate the theoretical radial distribution for the hydrogen atom's 3s orbital. However, the `emcee` ensemble sampler exhibits advantages:

- Accuracy: The `emcee` histogram aligns slightly more closely with the theoretical curve, especially around critical peaks and troughs.

- Convergence: The ensemble approach of `emcee` reduces autocorrelation and improves convergence speed, as shown by the smoother alignment with the theoretical distribution. While the convergence statistic may be slightly larger for `emcee`, it is still comparable to the custom MCMC and reducing the value of R provides diminishing returns after becoming less than 1.1 .

Potential discrepancies in the custom MCMC approach could stem from suboptimal proposal step sizes or insufficient exploration, leading to longer autocorrelation times. In contrast, `emcee` benefits from multiple walkers that reduce the risk of getting trapped in local probability peaks.

Conclusion

This project explored the 3s radial probability distribution of the hydrogen atom using two MCMC sampling methods: a custom Metropolis-Hastings algorithm and the `emcee` ensemble sampler. Both methods successfully approximated the theoretical radial probability distribution, with notable differences in performance and efficiency.

The custom MCMC method demonstrated the capability to capture the features of the 3s radial probability distribution, as shown by the alignment of its posterior samples with the theoretical curve. However, the trace of the walker highlighted the importance of implementing an appropriate burn-in period to discard non-representative initial samples. Additionally, tuning the proposal step size could further improve the sampling efficiency by reducing autocorrelation and enhancing convergence.

The `emcee` ensemble sampler, on the other hand, showed superior performance in terms of convergence and mixing. Its samples aligned more closely with the theoretical distribution, especially around critical peaks and troughs. This improvement is attributed to the parallel exploration of the parameter space by multiple walkers, reducing autocorrelation and sampling bias risk.

In summary:

- Both methods provide accurate approximations of the 3s radial probability distribution, with the `emcee` method demonstrating slightly better performance in terms of accuracy and convergence.

- The custom MCMC method, while effective, requires careful tuning of parameters such as the proposal step size and burn-in period to achieve optimal results.

- For future work, further improvements to the custom MCMC approach could be explored, including adaptive step-size tuning and alternative proposal distributions. Additionally, comparisons with other advanced sampling methods could provide deeper insights into their relative strengths and weaknesses.

The findings highlight the utility of both custom and ensemble MCMC methods for sampling high-dimensional distributions, with `emcee` offering particular advantages for complex problems requiring efficient and accurate sampling.