Final Results

Samantha Mayers, srm2204@columbia.edu

October 24, 2020

Objective

The objective of this project was to determine how to accurately approximate Gaussian curves, with the goal of speeding up calculating an atomic PDF.

Terminology

OG : Original Gaussian (the Gaussian which will be approximated)
Mini Gaussians: The smaller Gaussians that when summed, form an approximation for the original Gaussian
Approximation: The approximation formed from summing the mini Gaussians
Smaller Standard Deviation: The standard deviation used for all the mini Gaussians
Ratio: The ratio of the smaller standard deviation to the OG's standard deviation

Method

The method used was multiple linear regression. Each curve, using a given "small" standard deviation, had generated the intensities of a number of smaller Gaussians (all with the same "smaller" standard deviation), such that when summed together, would yield an approximation of the original Gaussian.

Two different approaches were taken. First, the number of mini Gaussians generated per curve was held constant, at 5. However, as the ratio of the smaller standard deviation to the OG's standard deviation decreased, the accuracy of the approximation plummeted. One way to correct for this drawback is to continually reset the smaller standard deviation as the OG's standard deviation increases. If the data's standard deviations are within a reasonably small ratio, no resets or few resets would be necessary, making this first approach useful for that situation.

The second approach taken was to vary the number of Gaussians generated per curve. For standard deviation ratios (smallest standard deviation/OG's standard deviation) greater than .6 (number taken from Coelho paper), 5 mini Gaussians were used. If the standard deviation ratio was less than .6, 10 Gaussians were used. For this approach, no resets of the smaller standard deviation are necessary.
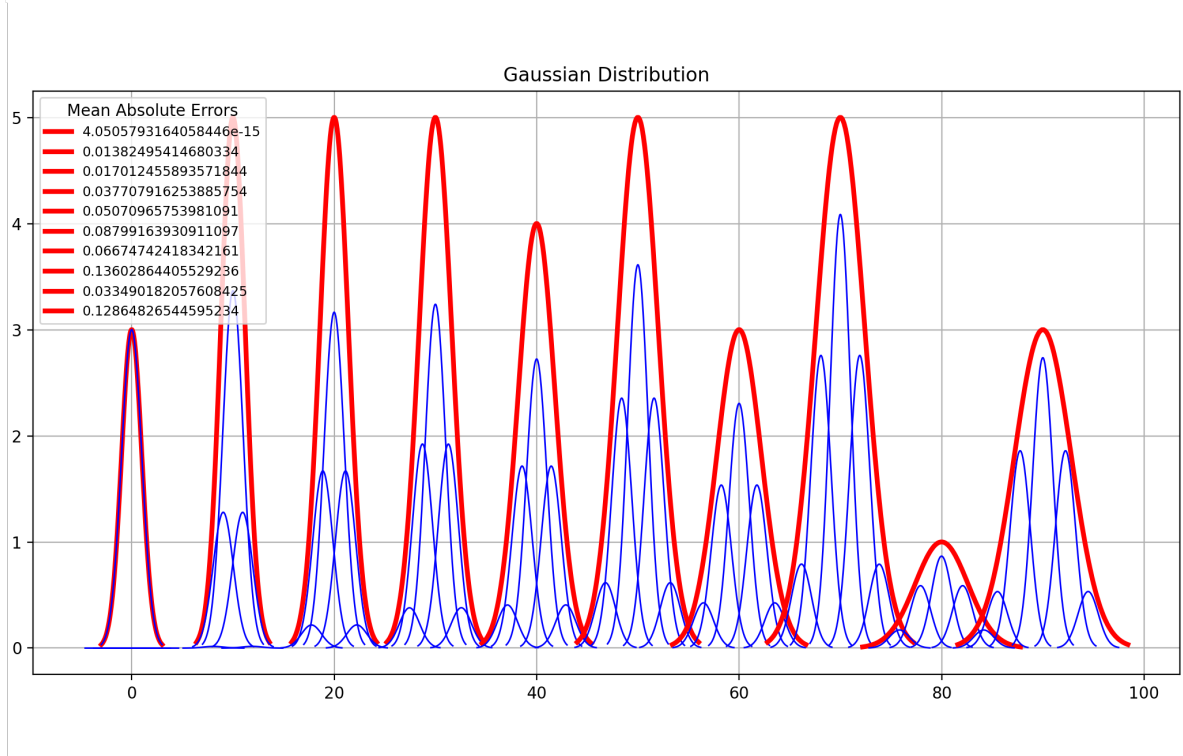
Results

Here are some figures demonstrating the results of the approximation technique.
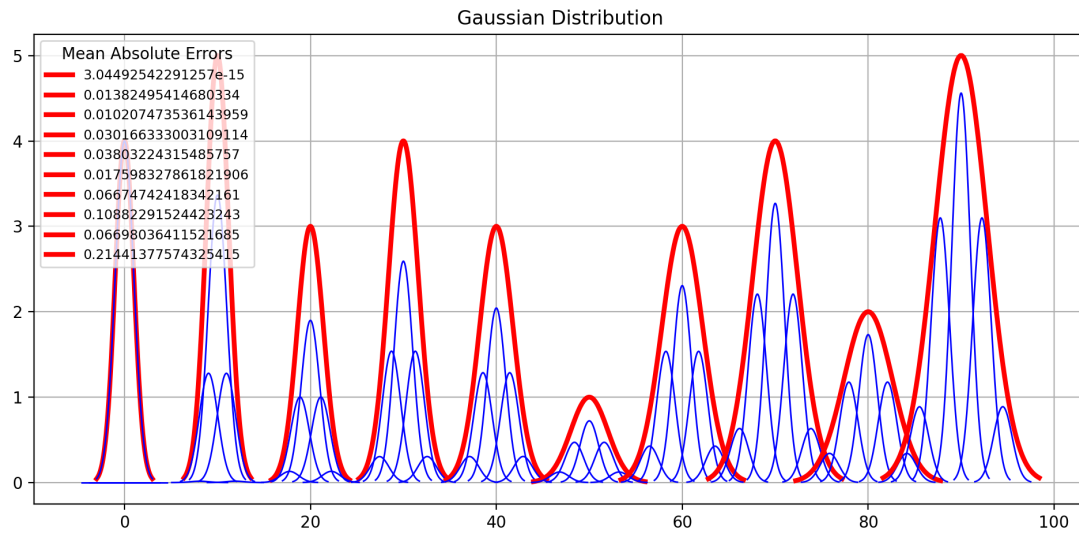
Method 1: Constant Number of Mini Gaussians

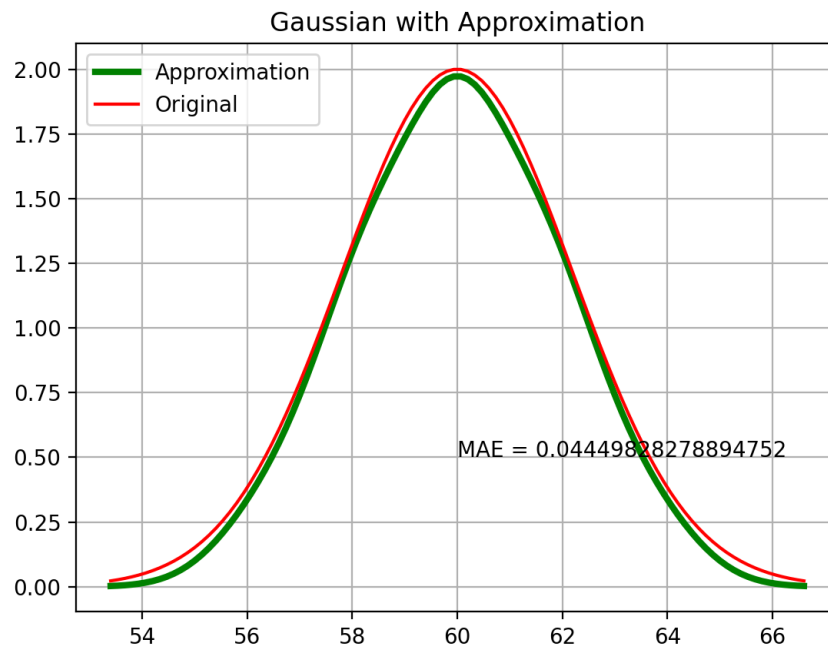First, the constant number of mini Gaussians approach:

No Resets

Within Method 1, first a look at the constant number of mini Gaussians approach with no smaller standard deviation resets:

Gaussian Distribution

Mean Absolute Errors
- 4.050579316405846e-15
- 0.01382495414680334
- 0.017012455893571844
- 0.037707916253885754
- 0.05070965753981091
- 0.08799163930911097
- 0.06674742418342161
- 0.13602864405529236
- 0.033490182057608425
- 0.12864826544595234

5 mini Gaussians were generated each for 10 OGs, with the smaller standard deviation set equal to the first OG's standard deviation, and with no smaller standard deviation resets. The intensities of the OGs were generated randomly from 1-5. The standard deviations increased for the OGs from left to right from 1-3, with a step size of .2. The largest mean absolute error is about .136. However, these errors are dependent on the random generation of OGs, and vary. Here is another figure demonstrating the exact same process, with different MAEs:

Gaussian Distribution

Mean Absolute Errors
3.04492542291257e-15
0.01382495414680334
0.010207473536143959
0.03016633003109114
0.03803224315485757
0.017598327861821906
0.06674742418342161
0.10882291524423243
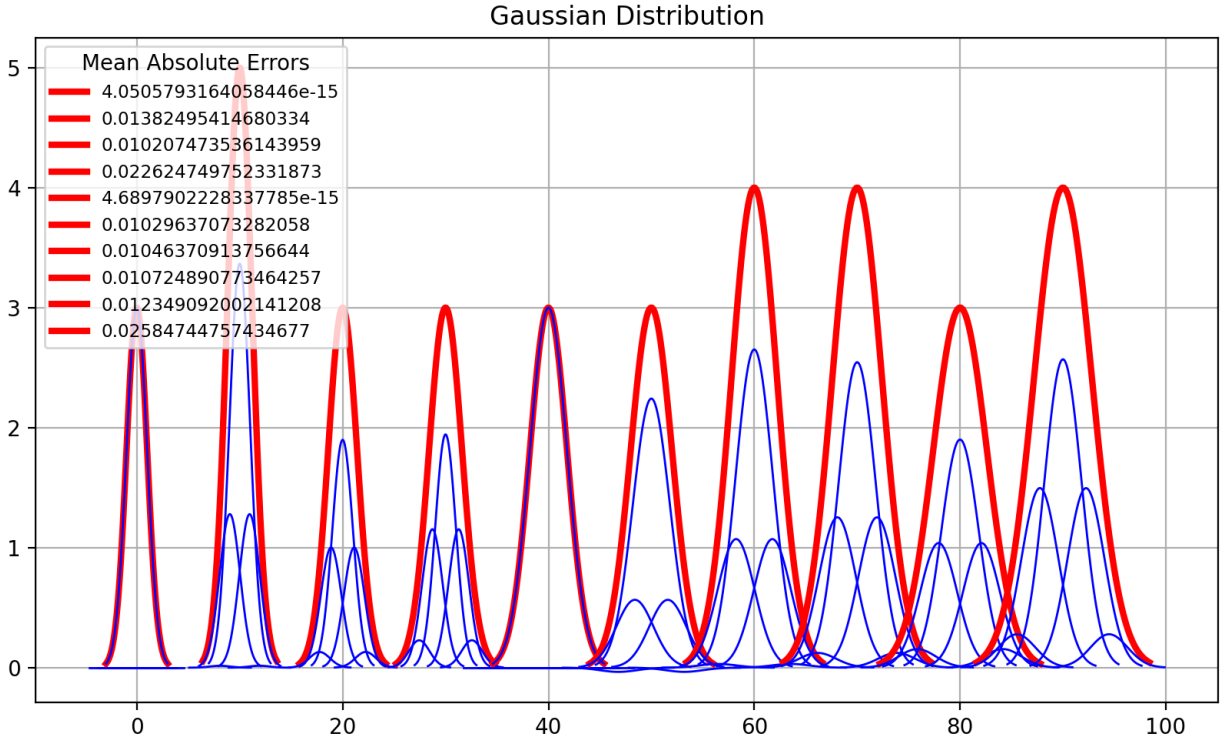0.06698036411521685
0.214413775743254155

The last OG's approximation has a mae of .214. Below is a graph showing a seventh curve (the seventh step from the smallest standard deviation) and the final approximation (the sum of the five mini Gaussians):

Gaussian with Approximation

Approximation
Original

MAE = 0.04449838278894752

4

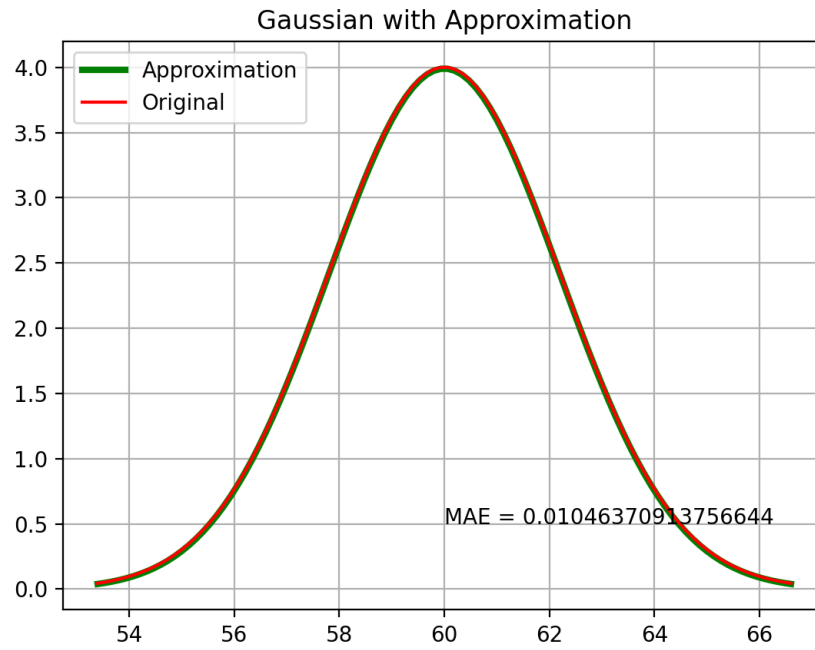For the constant number approach with no resets, the mean absolute error may become too high for small ratios.

Resets

Next, the constant number of mini Gaussians approach with smaller standard deviation resets:

Gaussian Distribution

Mean Absolute Errors
- 4.0505793164058446e-15
- 0.01382495414680334
- 0.010207473536143959
- 0.022624749752331873
- 4.6897902228337785e-15
- 0.01029637073282058
- 0.01046370913756644
- 0.010724890773464257
- 0.012349092002141208
- 0.02584744757434677

5 mini Gaussians were generated each for 10 OGs, with the smaller standard deviation set equal to the first OG's standard deviation, and with the smaller standard deviation being reset at OG number 5 (once the ratio became less than .6). The intensities of the OGs were again generated randomly from 1-5. The standard deviations increased for the OGs from left to right from 1-3, with a step size of .2. The largest mean absolute error is about .026. These errors are again dependent on the random generation of OGs, and vary. The smaller standard deviation reset occurred once the ratio was less than .6 (number taken from Coelho paper). Keeping the ratio above .6 leads to much less error in the approximations. Below is a graph showing a seventh curve (the seventh step from the smallest standard deviation) and the final approximation (the sum of
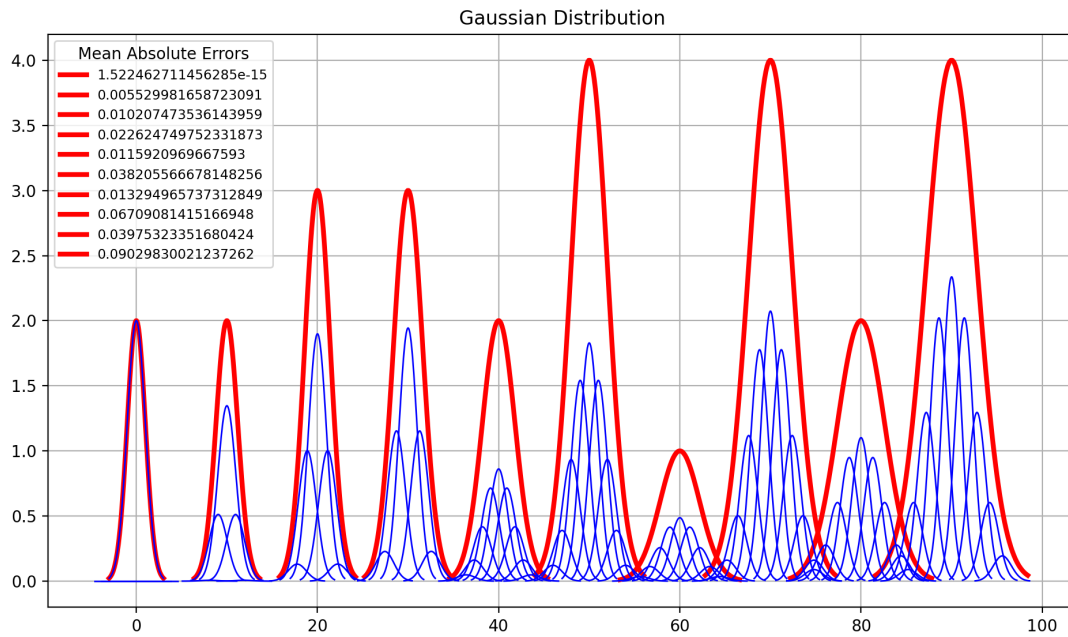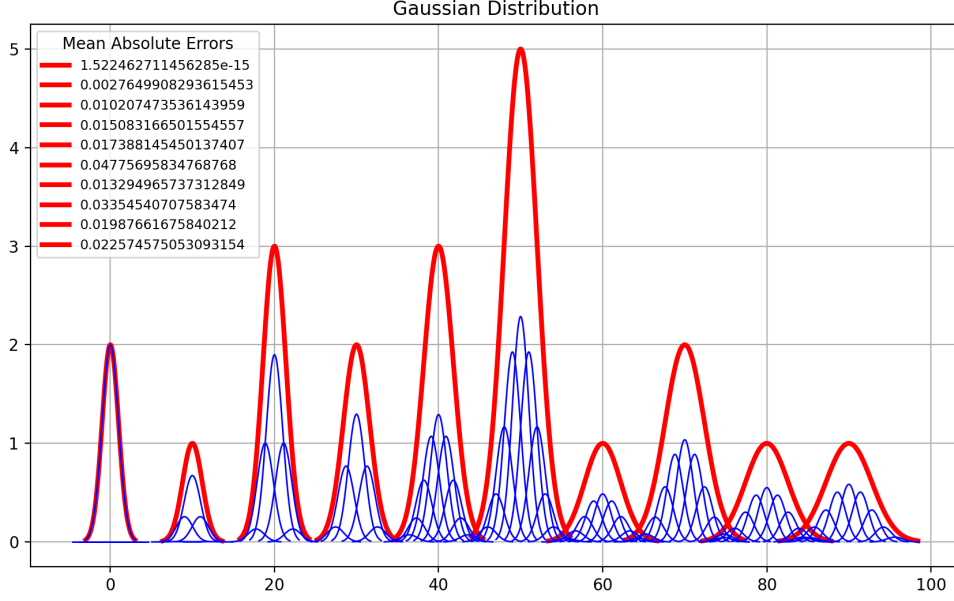
5

the five mini Gaussians):

For the constant number approach with resets, the mean absolute error stays small, with the approximation almost matching the OG, as evident in the figure above.

Method 2: Varied Number of Mini Gaussians

Second, the varied number of mini Gaussians approach:
Here are the results of two different instances of this method.

Gaussian Distribution

Mean Absolute Errors
- 1.522462711456285e-15
- 0.005529981658723091
- 0.010207473536143959
- 0.022624749752331873
- 0.01159209696667593
- 0.038205566678148256
- 0.013294965737312849
- 0.06709081415166948
- 0.03975323351680424
- 0.09029830021237262

Gaussian Distribution

Mean Absolute Errors
- 1.522462711456285e-15
- 0.0027649908293615453
- 0.010207473536143959
- 0.015083166501554557
- 0.017388145450137407
- 0.04775695834768768
- 0.013294965737312849
- 0.03354540707583474
- 0.01987661675840212
- 0.022574575053093154

The mae for this method is small. This method also requires no resets to be accurate, which is an advantage over method 1. However, one issue with this method is that right now, the only choices for the number of Gaussians is 5 or 10, based on the ratio being above or below .6. However, if the ratios in a dataset were even more drastic than the smallest ratio here (the 10th OG) of .357 (or $\frac{1}{2.8}$), then more curves may be necessary in order to generate an accurate approximation. This modification is simple, and would require a simple adjustment to the code. Below is a graph of the eighth curve (the eighth step from the smallest standard deviation) and the final approximation (the sum of the five mini Gaussians):
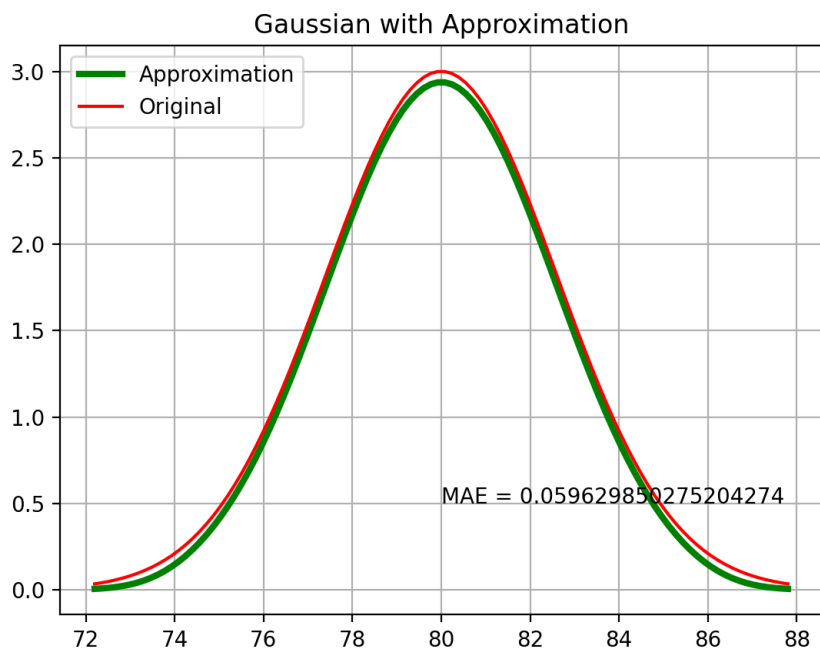
Overall, this method, since it requires no resets to guarantee accuracy, seems more promising, especially since it remains accurate for lower ratios.

Conclusions

Both approaches, Constant Number of Mini Gaussians (without and with resets) and Varied Number of Mini Gaussians were successful in that both were able to accurately approximate Gaussian curves by generating mini Gaussians that when summed, form a close approximation to the OG. The method that should be implemented I think would ultimately depend on the dataset and the smallest ratio of the dataset. Future work would consistent of implementing the Gaussian Approximation techniques in the PDF Calculator in order to speed up the calculations.

Works Cited

Coelho, A. A., Chater, P. A., amp; Kern, A. (2015). Fast synthesis and refinement of the atomic pair distribution function. Journal of Applied Crystallography, 48(3), 869-875. doi:10.1107/s1600576715007487

Code

Code: https://github.com/samrmayers/gaussian$_work$
(Methods in https://github.com/samrmayers/gaussian$_work/blob/master/GaussApproxFINAL.py$)