Computational Project 2: Optimisation of Signal Analysis in Discovering the Higgs Boson

Tongnian (Minnie) Wang — CID: 02020640

***Abstract*—This numerical investigation maximised the sensitivity of signal analysis in the context of Higgs boson detection. With an emphasis on precision, this investigation examined numerical integration methods and maximised the accuracy of a Gaussian integrator. Using two-dimensional maximisation, the maximum average significance was $S_{max} = 5.142343 \pm 0.000268$, with optimised selection cuts at $123.237502 \pm 0.000701 \frac{GeV}{c^2}$ and $127.158661 + / - 0.000701 \frac{GeV}{c^2}$. The investigation revealed the optimisation of selection cuts maximises the statistical probability of detecting a Higgs boson decay.**

I. Introduction and Aims

THE existence of the Higgs field was first predicted in the 1964 Physics Review Letters symmetry-breaking papers [**1**][**2**]. The Higgs boson was experimentally discovered at CERN in 2012 by the ATLAS and CMS collaborations, validating previous theoretical propositions. With a lifetime of only 10^{-22} s [**3**], the Higgs boson cannot be detected directly, but only by the byproducts of its decay. At the Large Hadron Collider, many different decay events create photons. To claim a discovery, a signal needs to have a five-sigma significance: when the number of photon pairs detected is more than 5 standard deviations away from the expected value from background interactions. The average significance of the measurement is sensitive to the selection cuts, where the selection cuts determine the range of invariant masses subjected to analysis.

The aim of this investigation was to find the selection cuts that maximised the average significance and consequentially maximised the sensitivity of the signal analysis.

II. Abbreviations

For readability, the following numerical methods will be referred to by its abbreviation: Monte Carlo Integration Method (MCIM), Extended Trapezoidal Method (ETM), Midpoint Riemann Method (MRM), Extended Simpson's Method (ESM), Ordinary Differential Equation (ODE) Methods and Runge-Kutta Fourth Order Method (RK4M).

III. Numerical Methodology

Due to the sensitivity of the analysis, it is necessary to minimise the error of the numerical integration of a Gaussian distribution.

Analysis of the error scaling was used to determine the provisional list of numerical integration methods for further accuracy investigation, as shown in Table. I. The ODE Methods were not considered, as the nature of ODE Methods, such as the Euler Method, is identical to the Riemann sums [**4**], with an identical error scaling as MRM [**5**]. RK4M has an identical error scaling to ESM, hence its consideration adds negligible value to the investigation. To minimise the error of the Gaussian integrator, the following factors and their effect on accuracy was investigated: step size, direct or indirect integration, and the value of the upper range of integration.

Numerical Integration Method	Error Scaling
Monte Carlo Integration Method	$\mathcal{O}(h)$
Extended Trapezoidal Method	$\mathcal{O}(h^2)$
Midpoint Riemann Method	$\mathcal{O}(h^3)$
Extended Simpson's Method	$\mathcal{O}(h^4)$
Runge-Kutta Fourth Order Method	$\mathcal{O}(h^4)$

TABLE I: This table shows the analytic error scaling of a selected list of numerical integration methods, where h is the step size. The error scaling indicated the more refined quadrature methods are more accurate, assuming h $\ll$ 1. [**4**]

The minimisation of the selection cuts required an accuracy of $1 \frac{keV}{c^2}$. To achieve this, a two-dimensional gradient descent method was used with a tolerance of $10^{-7} \frac{GeV}{c^2}$. From an observation of the surface of minimisation, demonstrated in Fig. 5, it can be concluded that there is only one minimum within the range $114 \leq \mathrm{m}_l \leq 125$ and $125 \leq \mathrm{m}_l \leq 136$. Although MCIM is a more robust method that

avoids detecting local minima rather than a global minimum, it introduces randomness and is a less precise method due to its stochastic nature. In this minimisation, the robustness of MCIM was not necessary, the choice of numerical method prioritised accuracy instead.

The choice of the learning rate (α) was arrived at empirically: to meet the required precision, it was necessary that $\alpha\nabla f(x_n) \leq 1\frac{keV}{c^2}$ around the minimum, where $\nabla f(x_n)$ is the gradient around the minimum. This avoided overstepping the minimum and terminating the minimisation prematurely.

IV. Code Validation

A. Numerical Integration Validation

The accuracy of the integration was verified by comparing it to the standard error function, described in Eq. 1 [**6**]. The error function is the integral of the normalised Gaussian distribution. By appropriate shifting and scaling, it can output the integral of any arbitrary Gaussian. In Python, there are built-in standard error functions with high accuracy. In this investigation, the accuracy of the integration methods was defined by their deviation from the corresponding result given by math.erf().

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt \tag{1}$$

B. Minimisation

Without the true minimum as a reference, a reliable method of validation was by comparison to the minimisation result from different methods. Trivially, the success of the minimisation can be verified by checking the average significance S value of the optimised parameters against unoptimised parameters, the magnitude of S directly indicates the success of the minimiser. The results here were validated by looking for agreement between two minimisation methods: gradient descent and parabolic method. Furthermore, the results were also validated by initiating the minimiser from different starting points and observing whether it reached the same minimum.

V. Results

A. Optimisation of the Gaussian Integrator

As demonstrated in Fig. 1, the accuracy of all integration methods improved as the step size decreased. The numerical analysis demonstrated surprising results: MRM was consistently the most accurate out of the quadrature methods. MCIM was consistently inaccurate and demonstrated the stochastic nature of its analysis - thus MCIM will be neglected for further analysis.

The dependence of accuracy on the direct/indirect approach to integration is method-dependent. For MRM, the indirect method was consistently more accurate for all step sizes. In contrast, the direct method was consistently more accurate for ESM. For all quadrature methods, the difference in accuracy for direct/indirect methods was negligible for a sample density greater than 2000.

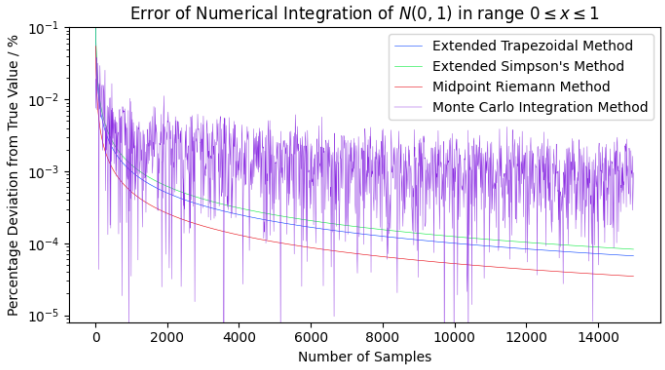

Fig. 1: A logarithmic graph showing the accuracy of different numerical integration methods integrating the standard normal distribution. By considering the percentage deviation from the true value, determined by the standard error function. A decrease in deviation from true value can be observed for a decreasing step size.

Fig. 2 captures the dependence of accuracy on the upper limit of integration. For both ETM and ESM, the accuracy improved as the upper limit of integration a increased. For MRM, the method was the most accurate for a low value of a. This behaviour was consistently observed for sample densities ρ_{sample} in the range $100 \leq \rho_{sample} \leq 4000$. The limit at which MRM is more accurate is $a = 1.625 \pm 0.125$, with no observed dependence on the sample density.

The following considerations were made to minimise the error of the Gaussian integrator:

- The value of a determined the numerical integration method used: for $a \leq 1.625$, MRM was used. Otherwise, ESM was used.

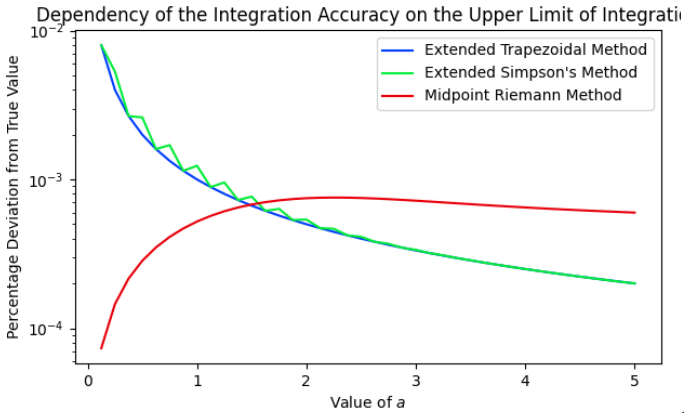

Fig. 2: A logarithmic graph showing the dependency of the accuracy of quadrature methods on the upper limit of integration a, for a set sample density of 1000. The figure demonstrates the variation in accuracy as a increases.

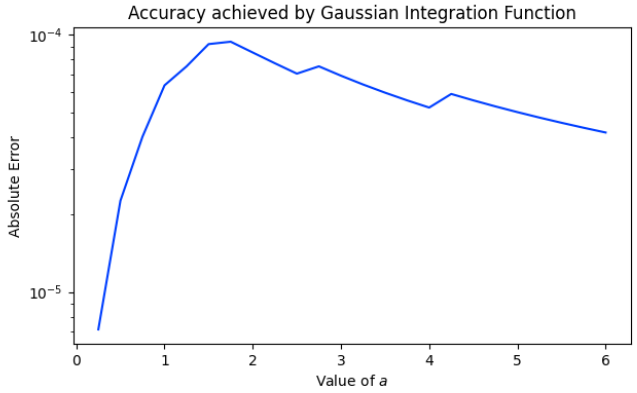

Fig. 3: A logarithmic graph demonstrating the achieved accuracy of the optimised Gaussian integrator as a function of the upper limit of integration a. The function has a maximum error of $9.387{\cdot}10^{-05}$ at $a = 1.625$ and a minimum error of $7.161 \cdot 10^{-06}$ at $a = 0.25$.

- For both MRM and ERM, the direct method was more accurate, therefore the direct method was adopted.
- The integrator optimised the sample density by finding the minimum sample density such that: by increasing the sample density, the percentage error does not decrease by more than $5 \cdot 10^{-5}$. The sample density is allowed to take any value within the range of $400 - 8400$.

The achieved accuracy of the optimised Gaussian integrator is presented in Fig. 3. The accuracy of the integrator was dependent on the value of a, it was most inaccurate for $a = 1.625$.

B. Optimisation of Selection Cuts m_l and m_u

The average significance S is highly sensitive to the selection cuts, where $\mathrm{m_l}$ and $\mathrm{m_u}$ are the lower and upper limits of invariant masses considered for analysis respectively. Preliminary investigations on the dependence of S on the range and position covered by $\mathrm{m_l}$ and $\mathrm{m_u}$ demonstrated the relationship presented in Fig. 4 and Fig. 5. Preliminary results suggested $S_{max} \approx 5.1$. This now presented a two-dimensional maximisation problem, with initial guesses informed by preliminary results in Fig. 4 and Fig. 5.

The optimised selection cuts, accurate to $1\frac{keV}{c^2}$, are as follows:

- $\mathrm{m}_l^* = 123.237502 \pm 0.000701 \frac{GeV}{c^2}$
- $\mathrm{m}_u^* = 127.158661 \pm 0.000701 \frac{GeV}{c^2}$

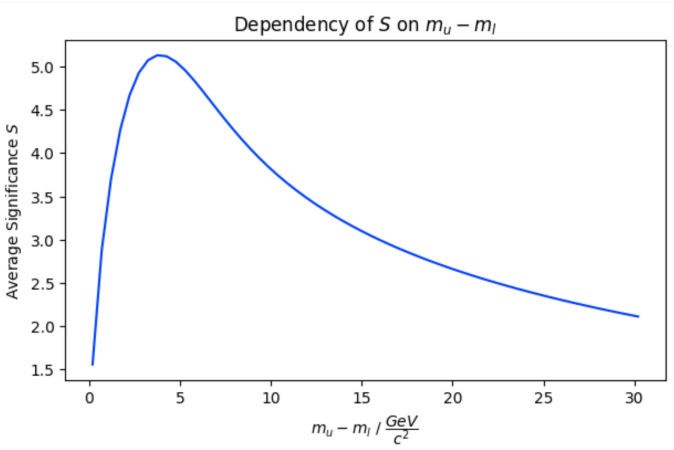

Fig. 4: The figure shows the dependence of S on the range covered by the selection cuts, for a constant $\mathrm{m_l} = 123$. An optimum range can be identified that maximises S where $\mathrm{m_u^*} - \mathrm{m_l^*} \approx 4$.

this achieved $S_{max} = 5.142343 \pm 0.000268$. The range of the optimised selection cuts was 3.9212 ± 0.0014, in agreement with the preliminary results in Fig. 4. Gradient descent and parabolic minimisation agreed on these selection cuts to $\pm 1 \frac{keV}{c^2}$.

C. Probability of a Five-Sigma Signal Measurement

Assuming the validity of the Central Limit Theorem, the sampling distributions of N_B and N_H can be modelled as Gaussian distributions[**7**]. The condition for a five-sigma signal measurement is mathematically represented by Eq. 2, and visualised

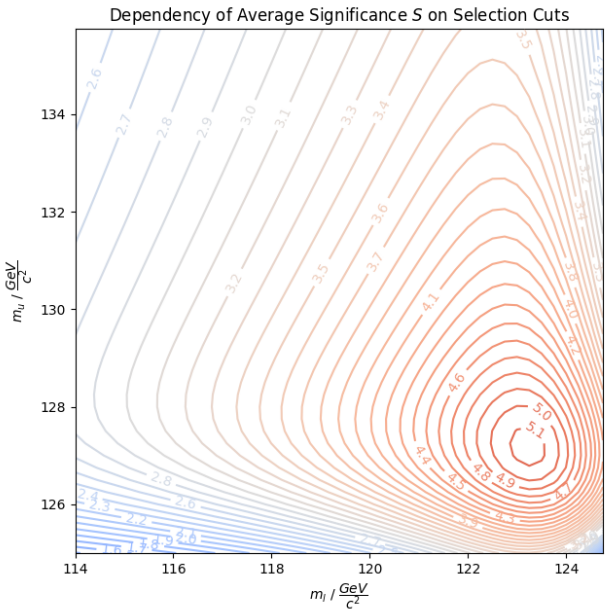

Fig. 5: The figure demonstrates the positional dependence of S on the two-dimensional variable space parameterised by m_l and m_u. The surface of maximisation is smooth with no discontinuities. The maximum is located in the region where $m_l^* \approx 123$ and $m_u^* \approx 127$ (the * notation indicates optimised parameters).

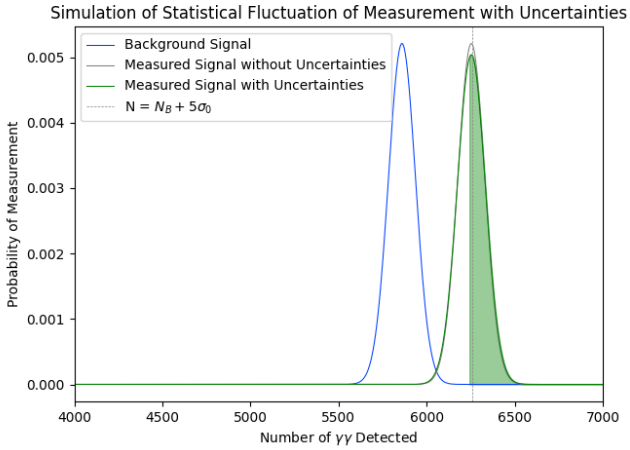

Fig. 6: Sampling distributions of the background and measured number of photons modelled with normalised Gaussian distributions, N_B and N respectively, where $\mathrm{N} = \mathrm{N}_B + \mathrm{N}_H$. The grey line only accounts for the statistical fluctuation of N_H. The green line models the measured signal with consideration of the following sources of uncertainties: the Higgs mass, photon interactions and number of Higgs bosons created. The green area corresponds to the cumulative probability of a five-sigma signal measurement. The gray vertical line indicates the five-sigma limit, where $\sigma_0 = \sqrt{\mathrm{N}_B}$.

by the green shaded area in Fig. 6. The probability of a 5σ measurement is 0.556594 ± 0.000029, with consideration of only the statistical fluctuations of measurement and using optimised selection cuts.

$$\mathrm{P}(5\sigma) = \mathrm{P}(\mathrm{N} \geq \mathrm{N}_B + 5\sigma), \quad \mathrm{N} = \mathrm{N}_B + \mathrm{N}_H \quad (2)$$

The three following sources of uncertainties contribute varying uncertainties to N: the Higgs boson mass, photon interaction fraction and the number of Higgs bosons created. The absolute errors introduced by these uncertainties are:

- $\sigma_{\text{Higgs mass}} = \pm 3.9457553$
- $\sigma_{\text{Photon interaction fraction}} = \pm 5.723282$ assuming 100% single photon interaction.
- $\sigma_{\text{Photon interaction fraction}} = \pm 15.7462014$ assuming 100% photon pair annihilation.
- $\sigma_{\text{Number of Higgs bosons created}} = \pm 11.809651$.

Since $\sigma_{N_H} = \pm 19.8407418$, the only uncertainties with an error of the same magnitude are $\sigma_{\text{Number of Higgs bosons created}}$ and $\sigma_{\text{Photon interaction fraction}}$ assuming 100% photon pair annihilation. Other uncertainties are negligible in comparison with the statistical fluctuation of the distribution. Accounting for the two sources of uncertainties, $\mathrm{P}(5\sigma)$ is 0.554823 ± 0.000029. If all three uncertainties were accounted for, $\mathrm{P}(5\sigma)$ is 0.554756 ± 0.000029, introducing an additional uncertainty of only 0.007%, justifying the negligence of $\sigma_{\text{Higgs mass}}$.

VI. Discussion

A. Gaussian Integrator Accuracy

As demonstrated in Fig. 3, the accuracy of the Gaussian integrator was dependent on the upper limit of integration, and hence the method of integration and the step size. This introduced a variable uncertainty for the rest of the investigation when using the integrator, dependent on a. The accuracy of the integrator was determined by comparison to the corresponding output from the error function contained in the maths module. The built-in error function is also subject to truncation errors, since it is a finite series utilising the Abramowitz and Stegun

numerical approximation [**8**]. The exact approximation used is unknown, but the approximations can introduce a maximum error in the range: $5 \cdot 10^{-4} \leq \epsilon \leq 1.5 \cdot 10^{-7}$[**9**], where ϵ is the truncation error. This introduced a systematic uncertainty to the accuracy of the integrator from an error on the reference true value, but it did not affect the determination of the most accurate integration method by comparison of relative accuracies. MRM did not introduce an inherent overestimate/underestimate, unlike the left/right Riemann sum methods. The surface of maximisation is smooth with no discontinuities, hence it is difficult to determine whether the two methods are an overestimation or underestimation. The role of round-off error and propagation error must also be considered: any floating point number will have an associated round-off error of the order 10^{-16}. The series of arithmetic operations will introduce a propagation error, with a round-off error introduced at each point and accumulated over a sequence of calculations. In this investigation, the attainment of the results do not require sequential calculations, therefore the global truncation error will not be amplified significantly.

B. Minimisation Results and the Probability of a Five-Sigma Signal Measurement

The calculation of S involved two numerical integrals to calculate N_B and N_H. The percentage error on S_{max} from the two integrals, combined in quadrature [**10**], is $5.263147 \cdot 10^{-5}$ %, introducing an absolute uncertainty of $\pm 2.6763 \cdot 10^{-4}$. As a result, there was an uncertainty on the value of the optimised selection cuts. Whilst the local maximisation achieved an accuracy of $1 \frac{keV}{c^2}$, due to the plateau around the maximum, there was a range of selection cuts that produced a value for S within the confidence interval. This introduced an absolute error of $\pm 0.0007 \frac{GeV}{c^2}$.

Maximisation using gradient descent and parabolic method converged on the same optimised selection cuts. The convergence was also witnessed for optimisation from different starting points. The α value for gradient descent was selected such that $\alpha \nabla f(\mathrm{m}_l^*, \mathrm{m}_u^*) \approx 0.5 \frac{keV}{c^2}$ for steps near the maximum to achieve an accuracy of $1 \frac{keV}{c^2}$. Given $\nabla f(\mathrm{m}_l^*, \mathrm{m}_u^*) \approx [-0.00020, 0.00038]$, an alpha value of 0.001 satisfied the accuracy condition.

From observing Fig. 6, the conclusion $\mathrm{P}(5\sigma)_{\text{without uncertainties}} \geq \mathrm{P}(5\sigma)_{\text{with uncertainties}}$ can be arrived at, graphically represented by the change in area over the 5σ limit. The decrease in P(5σ) when considering uncertainties is reasonable due to a greater range of possible measurements and additional sources of statistical fluctuation.

The calculation of P(5σ) involved simplifications that introduce uncertainties. Firstly, the error of the Gaussian integrator introduced an uncertainty of $5.263147 \cdot 10^{-5}\%$, this equated to an absolute uncertainty of $\approx \pm 0.00003$ for both P(5σ) values.

To calculate P(5σ), N_B was approximated as a Gaussian, a common choice for modelling parameter uncertainties, this is also justified by Central Limit Theorem [**11**]. Since a Gaussian integrator had already been optimised for a previous part of the investigation, this simplification improved the accuracy of the integration of the Gaussian. However, the approximation itself introduced uncertainties to the result. For $n \gg 15$, the error for Central Limit Approximation is of $\mathcal{O}(n^{-\frac{1}{2}})$[**12**]. In this case, $\mathrm{N}_B = 5860.1608$, giving an error of order $\approx 1 \cdot 10^{-2}$ on the probability.

The assumption of 100% photon pair annihilation is an overestimation, in reality, the probability of photon pair annihilation will be smaller than one. This assumption was made to account for the maximum possible uncertainty, leading to an overestimation in the combined uncertainty. The assumption that all uncertainties are uncorrelated was likely to have introduced an underestimation of the statistical fluctuation of N. In reality, the number of Higgs bosons created and the Higgs mass will likely have a non-negligible influence on the number of photons that interacts with the detector. This consequently led to an overestimation in P(5σ).

C. Dependency of P(5σ) on the Accuracy of Optimisation

In an attempt to contextualise the efforts of optimising the analysis selection cuts, an investigation was conducted to investigate the relationship between P(5σ) and the distance between a pair of arbitrary selection cuts. The distance $|m - m*|$ can be interpreted as a quantitative measure of how optimised a parameter is. As observed in Fig. 7: the numerically simulated data demonstrated a linear relationship. This result indicated that optimum selection cuts maximise the probability of detecting a five-sigma signal, justifying all the efforts max-

imising accuracy in this investigation. By extrapolation of the fitted line, $\mathrm{P}(5\sigma)_{\mathrm{max}]}$ was predicted to be 0.5565954 ± 0.0000002. This is in agreement with the numerical result up to 5 decimal places. This investigation contextualised the optimisation of selection cuts and its relation to the likelihood of success in detecting a Higgs boson decay.

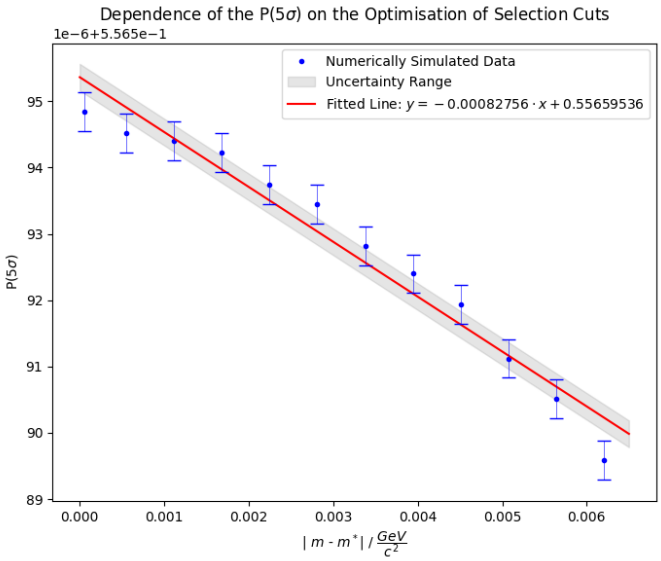

Fig. 7: Demonstration of the linear relationship between the distance of arbitrary selection cuts from the optimised selection cuts and the probability of a 5σ measurement. The dependency can be modelled by $\mathrm{P}(5\sigma) = (-0.0008276 \pm 0.0000550)|\mathbf{m} - \mathbf{m}^*| + (0.5565954 \pm 0.0000002)$.

VII. Conclusion

In conclusion, this investigation provided an approach to maximise the sensitivity of signal analysis for the case of Higgs boson decay at the Large Hadron collider. This was achieved by maximising the average significance S and finding the optimum selection cuts that dictated the range of invariant masses considered for analysis. The optimised result for $S_{max} = 5.142343 \pm 0.000268$ was found at $(\mathrm{m}_l^* = 123.237502 \pm 0.000701, \mathrm{m}_u^* = 127.158661 \pm 0.000701)\frac{GeV}{c^2}$. The exploration of the relationship between selection cut accuracy and the probability of a five-sigma signal measurement provided a subtle insight into the causal relationship between the precision of methodology and achieved results. This investigation was limited by computation power and time, constricting the sample densities and hence the accuracy of the Gaussian integrator.

VIII. References

[**1**] P. W. Higgs, "Broken symmetries and the masses of gauge bosons," *Physical Review Letters*, vol. 13, no. 16, pp. 508–509, 1964. doi:10.1103/physrevlett.13.508

[**2**] G. S. Guralnik, C. R. Hagen, and T. W. Kibble, "Global conservation laws and Massless Particles," *Physical Review Letters*, vol. 13, no. 20, pp. 585–587, 1964. doi:10.1103/physrevlett.13.585

[**3**] "Atlas finds evidence of a rare higgs boson decay," *CERN*, https://home.cern/news/news/physics/atlas-finds-evidence-rare-higgs-boson-decay (accessed Dec. 11, 2023).

[**4**] M. Scott and J. Owen, "Chapter 5. Numerical Calculus," in *Computational Physics Lecture Notes 2023*, Imperial College London.

[**5**] J. C. Butcher, *Numerical Methods for Ordinary Differential Equations*. New York: John Wiley Sons, 2003. ISBN 978-0-471-96758-3.

[**6**] L. C. Andrews, *Special Functions of Mathematics for Engineers*. Oxford: Oxford University Press, 1998.

[**7**] O. Johnson, *Information Theory and the Central Limit Theorem*. London: Imperial College Press, 2004.

[**8**] T. Peters, "cpython/Modules /mathmodule.c," GitHub, https://github.com/python/cpython/blob/main/Modules/mathmodule.c (accessed Dec. 11, 2023).

[**9**] M. Abramowitz and I. A. Stegun, "Chapter 7", *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Washington, D.C.: U.S. Dept. of Commerce, National Bureau of Standards, 1972.

[**10**] D. S. Lemons and P. Langevin, *An Introduction to Stochastic Processes in Physics*. Baltimore: Johns Hopkins University Press, 2002.

[**11**] The ATLAS Collaboration, The CMS Collaboration, and The LHC Higgs Combination Group, *Procedure for the LHC Higgs boson search combination in Summer 2011*, pp. 13–18, Aug. 2011. ATL-PHYS-PUB-2011-11, CMS NOTE-2011/005

[**12**] D. Draper and E. Guo, *The Practical Scope of the Central Limit Theorem*, Nov. 2021. University of California, Santa Cruz