Incorporating Stepping-Stone Sampling Into BayesWave

Seth Moriarty and Mentors: Sophie Hourihane, Katerina Chatziioannou
(Dated: September 2022)

`BayesWave` is a library of code used to analyze data from LIGO's gravitational wave detections. `BayesWave` uses Bayesian statistics to reconstruct signals and determine possible sources. The likelihoods of various models can be compared so that `BayesWave` can determine the most likely sizes, locations, and types of sources that could produce a certain detected signal. Currently, `BayesWave` uses Thermodynamic Integration (TI) to calculate the likelihoods of various models. An alternative method is called Stepping-Stone (SS) sampling. In other fields, SS has been shown to be as accurate as TI while also being less computationally expensive. This project explores the comparison between TI and SS methods when each is applied inside `BayesWave`, to determine if SS is a viable replacement for TI to be used for analysis of LIGO's fourth detection run in 2023.

I. INTRODUCTION/BACKGROUND

A. LIGO

Gravitational waves are ripples in space-time caused by high-energy events in outer space, such as supernovae and the collisions of black holes. The Laser Interferometer Gravitational-wave Observatory (LIGO) [1] is a large ground-based interferometer used to detect those ripples. Such ripples change the way that light travels and LIGO measures that change as slight fluctuations in its interferometric "arm" length. LIGO first detected gravitational waves in 2015, and has been carrying out observational runs since, finishing its most recent observing run in 2020[2], and is planned to start its fourth observing run (O4) in 2023. A significant amount of data processing is required to convert detected signals into astrophysical data that can be used to describe properties of the signal's source.

B. Bayesian Statistics

Bayesian statistics is a form of statistics in which earlier probability distributions (priors) can be updated to account for new data to produce new distributions (posteriors). This is done using Bayes' Theorem, given by equation 1.

$$P(A|B) = \frac{\mathcal{L}(B|A)\pi(A)}{P(B)} \tag{1}$$

The "posterior", $P(A|B)$, represents the probability of some event A given that B is true. The "prior" $\pi(A)$ is the probability of some event A independent of other events, and is used to encode our understanding of the statistics before beginning an analysis. The "likelihood", $\mathcal{L}(B|A)$, is the measure of how well out model fits the data. In order to normalize a posterior distribution (which is necessary in order to compare probabilities of different models), you integrate the numerator of 1 over the parameter space. The normalization factor (written as $P(B)$) is called the "evidence". Evidences are used to compare the probabilities between different models, and are generally compared using the "Bayes Factor":

$$\frac{P(B_1)}{P(B_2)} \tag{2}$$

Bayes factors are an essential tool in model comparison. In LIGO analyses Bayes factors are used to determine the most probable source of gravitational wave signals. For our purposes we use them to determine whether excess power is astrophysical, Gaussian noise, or non-Gaussian noise. As more GW detections are confirmed, LIGO's pool of confirmed sourches increases an our priors become more accurate to what traits we expect sources to have.

As mentioned, evidence calculation is an integral; when parameter spaces are large, performing such an integral can be computationally challenging. In this study we will discuss the uses of thermodynamic integration and the stepping-stone algorithm in the context of evidence calculation in LIGO analyses using `BayesWave`.

C. `BayesWave`

`BayesWave` [3] is a library of code which analyzes LIGO data using Bayesian statistical methods. is able to account for multiple models in its analyses including Gaussian noise (PSD), non-Gaussian noise ("glitches"), and coherent power between detectors ("signals"), and recently compact binary coalescence templates ("CBC"). The signal and glitch models are both reconstructed using wavelets where the signal model has the additional constraint that the wavelets must be coherent between detectors. Models can then be compared with their Bayes' Factors, to determine which are good fits for the observed data. This is an important tool for determining whether a detection is actually from an astrophysical source or from a glitch within the interferometer's system.

When `BayesWave` is used to analyze interferometric signals, it creates and stores model data and diagrams in a directory, useful for analysis. Figure 1 shows a reconstruction of the 150914 signal at the Hanford detector

using a coherent wavelet (signal) model. Figure 2 shows the reconstructed spectrogram from the same signal.

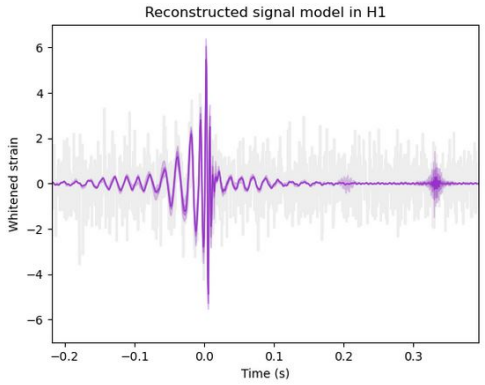

FIG. 1: reconstructed waveform using `BayesWave`

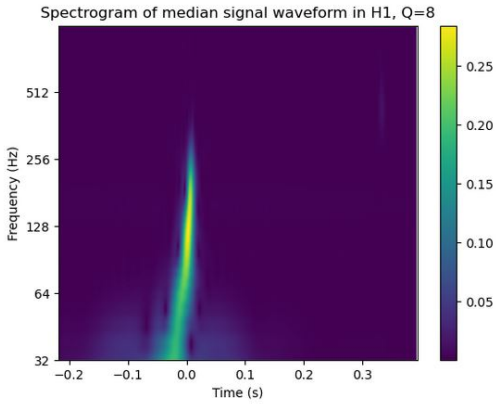

FIG. 2: reconstructed signal spectrogram using `BayesWave`

An important part of `BayesWave` that influences this project is `BayesLine`, which is used to more accurately estimate the power spectral density (PSD) of the instrumental noise during a detection [3] than other PSD estimation methods. There are many sources of noise such as ground motion and optical thermal fluctuations. The more accurately that noise can be quantified and parsed through, the more accurately a signal can be reconstructed. [3]. The details of `BayesLine`'s algorithm are not relevant to this project, but it is important to be aware that applying `BayesLine` to a run of `BayesWave` produces different evidence estimates than when it is off, due to its effect on signal reconstruction.

II. EVIDENCE CALCULATION

Evidence calculation is the integration of likelihoods in a high dimensional space. When the likelihood is sharply peaked in a large space, this integration can be tricky. We need methods such as thermodynamic integration and stepping-stone, which can be used to calculate multiple evidences on likelihoods which can be scaled to smooth out their peak. The factor called "temperature" scales the likelihood, with a larger temperature making the likelihood peak easier to find since it will be non-negligible on a larger set of parameter space compared to the unscaled likelihood. The set of points used to calculate the likelihood at a given temperature is called a "chain".

A high temperature corresponds to a chain which, similar to the behavior of a thermodynamic system, has a higher chance of jumping to less likely states. This means that the likelihood function flattens to approach the prior distribution (flattening out in parameter space). Likewise, a low temperature corresponds to chains that are more likely to stay in areas of high probability [4], resulting in a peaked probability distribution approaching that of the posterior. We define β as the inverse of the temperature such that a β value moving from zero to one corresponds to chains "cooling down", while starting at one and moving to zero corresponds to a chain "heating up".

$$p_i(\theta) = \frac{\mathcal{L}(D|\theta, M_i)\pi(\theta|M_i)}{z_i} \tag{3}$$

We will rewrite Bayes' theorem (Eq 1) as the above in order to show what it looks like on a given chain and to motivate both thermodynamic integration and the stepping stone algorithm. We introduce an index i which corresponds to the chain or equivalently "temperature" of the likelihood we are evaluating. Here $p_i(\theta)$ is the posterior probability density for some model, $\pi(\theta|M_i)$ is the prior distribution $\mathcal{L}(D|\theta, M_i)$ is the likelihood function of some data, D, given that the model is true. We will also write the evidence corresponding to a given temperature, z_i, more explicitly below [4]:

$$z_i = p(D|M_i) = \int \mathcal{L}(D|\theta, M_i)\pi(\theta|M_i)d\theta \tag{4}$$

The marginalized likelihoods can be compared using the log Bayes' factor which we will call μ:

$$\mu = ln(\frac{z_1}{z_0}) \tag{5}$$

What we can do is rewrite equation 5 in terms of the symbol β.

$$\mu = ln(\frac{z_1}{z_0}) = ln(z_1) - ln(z_0) = \int_0^1 \frac{\partial ln(z_\beta)}{\partial \beta} d\beta \tag{6}$$

III. THERMODYNAMIC INTEGRATION

Currently `BayesWave` uses thermodynamic integration (TI) [5] to calculate evidences for potential models. To

do so TI estimates the integral form of equation 6[6]. TI is also known as path sampling, because it involves taking samples along a path of temperatures. β (again, inverse temperature) begins at one extreme, either zero or one, and travels along a path to the other extreme as the temperature changes. $\beta = 0$ corresponds to the chain containing posterior samples, and $\beta = 1$ corresponds to the chain containing prior samples.

TI estimates μ by taking many discrete steps as β moves between 0 and 1 and taking samples at each temperature. The samples can then be used to estimate the integral. As can be expected, the estimate of an integral curve using a finite number of values will introduce discretization bias into the estimate. The more samples we take, the smaller this bias, and the more accurate the estimates are.

$$E_\beta[log(\mathcal{L}(D|\theta, M))] = \int log(\mathcal{L}(D|\theta, M))^\beta \pi(\theta) d\theta \quad (7)$$

To estimate the integral in equation 6, we rewrite the integral in terms of the expectation value of the log likelihood, affected now by the term β, as seen above [7]. β is the inverse of the temperature of a sampling chain. From this integral, the logarithmic evidence can be rewritten as the following [7]:

$$log(\mu) = log(\frac{z_1}{z_0}) = log(z_1) = \int_0^1 E_\beta[log(\mathcal{L}(D|\theta, M))] d\beta \quad (8)$$

For each chain location along β's path, samples will be taken and a sample average acquired. Then, those values can be used to estimate the integral [7], as per equation 8. This takes some time, as each β chain is run one at a time, rather than in parallel. An estimate can be seen in the following plot.

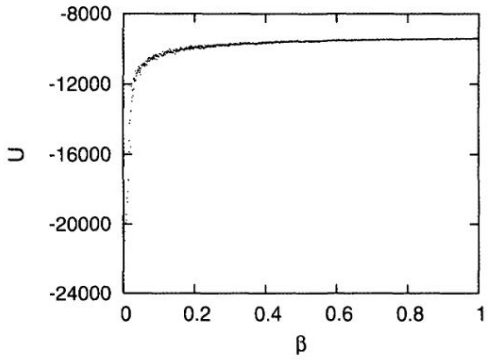

FIG. 3: Estimate of the marginal likelihood using 1000 points, from "Computing Bayes' Factors Using Thermodynamic Integration" [4]

Thermodynamic integration has proven to be a very reliable method, provided enough samples are taken to minimize bias and produce accurate estimations of the desired curve. TI experiences thermic lag bias [4] as it changes β values and adjusts to each new value. This has been shown to cause TI to slightly underestimate marginal likelihood values if β is integrated from zero to one, and a slight overestimate if it is taken from one to zero. It also experiences discretization bias, because of the limits on the accuracy with which a discrete number of values can estimate a continuous integral. Despite these points, and the fact that it is computationally expensive, TI is significantly more accurate than simpler methods, and thus a very helpful tool.

IV. THE STEPPING-STONE METHOD

The stepping-stone algorithm [7] is a method for finding evidences that is similar to TI but is less computationally costly. like TI, SS calculates marginal likelihoods directly, producing similar, and actually slightly more accurate [8] estimates.

the stepping-stone method calculates evidences differently than TI. Rather than calculating the average likelihood at each β value and summing them to estimate the integral as in Eq 8, SS compares marginal likelihoods between each discrete β_i value and that of the one before it, in a process called importance sampling. Then the product of those ratios can be used to estimate the evidence [7]. This is shown in equation 9, where K is the number of chains (β values) used.

$$z = \frac{z_1}{z_0} = \prod_{k=1}^{K-1} \frac{z_{\beta_k}}{z_{\beta_{k-1}}} \quad (9)$$

In other fields, this method has been shown to be more accurate than the TI method of averaging samples at each step[8]. LIGO's fourth observational run is expected to detect significantly more events than previous runs, so it is possible that the SS algorithm will make an important addition to the tools `BayesWave` has at its disposal to analyze that data.

V. OBJECTIVES

Ultimately, the goal of this project has been to compare Thermodynamic Integration to a Stepping-Stone algorithm, to compare their abilities to accurately estimate evidences for modeling GW signals. These comparisons provide information about which method ought to be used for future gravitational wave analyses in the upcoming 04 run in 2023. If it seems worthwhile, Meg Millhouse's branch of `BayesWave`, which incorporates SS sampling, can be incorporated into the main branch.

To do this, many runs of `BayesWave` have been done using both TI and SS, so that the output data could be compared. Run data has been organized into dataframes

inside Python so that it can be easily observed and plotted. Conclusions drawn from this are detailed in the following section.

VI. RESULTS

The first step of this project was to familiarize myself with the functions of `BayesWave`, and with the other background knowledge required for this project. I began by working through the GWOSC open data workshop for signal analysis, a helpful basis for understanding the steps that go into matching models to raw signal data. This made looking at the output data from `BayesWave` runs more intuitive.

Rather than working from the main branch of `BayesWave`, runs for this project were done on Meg Millhouse's branch, which includes the stepping-stone algorithm in addition to thermodynamic integration. Runs provide evidences calculated using each of those methods. Early in the process of this project, I did various runs to familiarize myself with the process of configuring them, using Condor to automate runs, and reading the output files. These runs were done using an injected signal of the 150914 waveform, with a fairly high signal-to-noise ratio.

To test the stepping-stone algorithm, we varied multiple parameters including the number of chains used, the number of iterations, and whether `BayesLine` was on or off. The number of chains refers to the number of discrete β values at which the evidence is being estimated. As is to be expected, more chains leads to a more accurate estimate of how well a model matches the signal. The number of iterations is how many times each sampling chain moves within parameter space as it samples. `BayesLine` can be toggled on and off for runs, and turned out to effect the comparison between the accuracy of TI and SS. Later, the realization of the random background noise was also varied.

My first large set of runs included runs with 2, 5, 7, 10, 12, 15, 20, and 30 chains. For each chain value, runs were conducted with 1, 2, 3, and 4 million iterations. This set of runs was done with `BayesLine` both on and off, and used both TI and SS.

For the runs that did not use `BayesLine`, the number of chains used versus the calculated evidence estimates is given by figure 4. The cool colored dots are evidences found using TI, and the warm X's are evidences found using SS. The color bars indicate the number of iterations used to obtain each evidence. For thermodynamic integration, we have estimates for the standard deviation of the evidence estimates, so error bars are included in the plot. Currently, the standard deviation from SS is not calculated. Figures 5 and 6 show the standard deviation on the TI evidence estimates depending on the amount of sampling chains used.

From figure 4, we can see that TI estimates are poor when a small amount of chains are used, especially when run with few iterations. As the number of chains increases, the estimates of the evidence value quickly converge to a very accurate value. In the case of TI, the standard deviation drops to a value very close to zero. From figures 7 and 8, we can see that both TI and SS provide very accurate estimates when 10 or more chains are used. However, the deviation from this value at low chains is much larger using TI than it is for SS. This behavior agrees with figure 9[7], from a 2018 paper also exploring the comparison between SS and TI in the context of gravitational wave physics.

Interestingly, behavior is different when `BayesLine` is

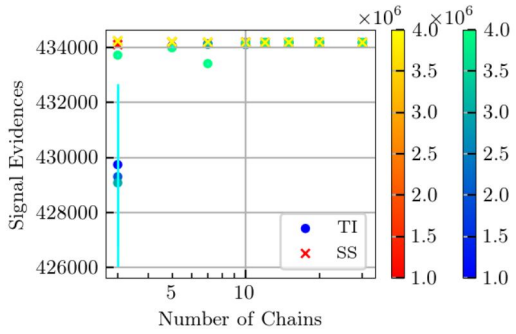

FIG. 4: Number of chains used for sampling vs calculated evidence values (`BayesLine` off)

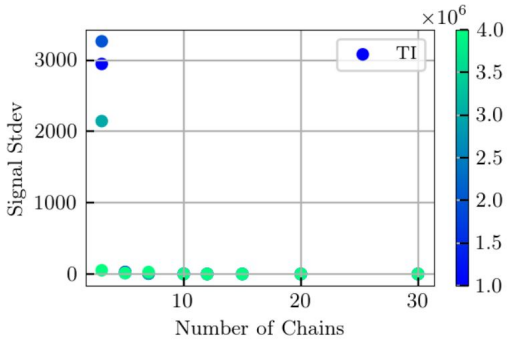

FIG. 5: Number of chains used for sampling vs standard deviation of the calculated TI evidences

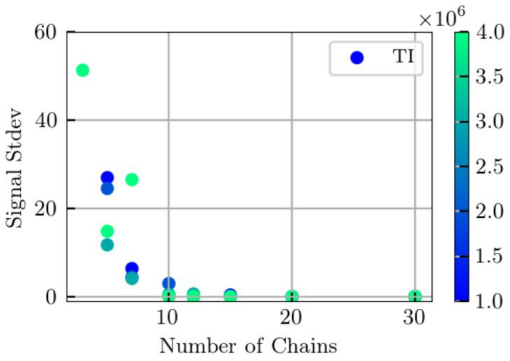

FIG. 6: Close-up of figure 5

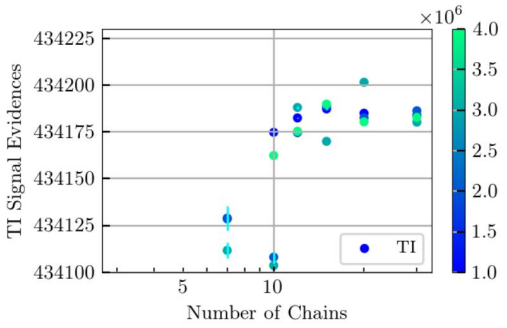

FIG. 7: Thermodynamic integration evidence values (`BayesLine` off)

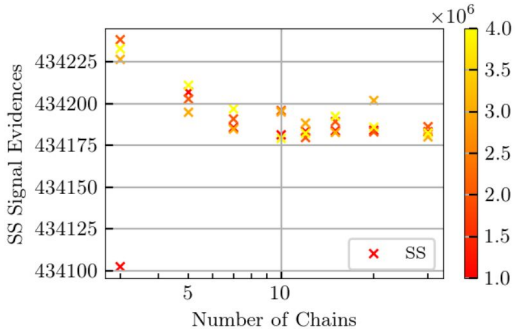

FIG. 8: Stepping-Stone evidence values (`BayesLine` off)

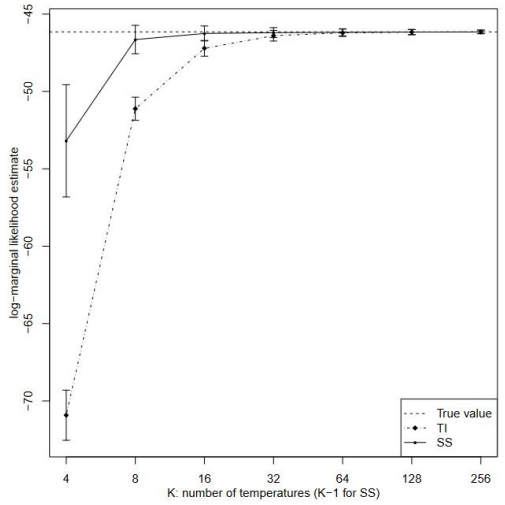

FIG. 9: TI and SS evidence comparison from "The stepping-stone sampling algorithm for calculating the evidence of gravitational wave models", [7]

turned on. Here, the accuracy of evidences at low chains is comparable between SS and TI. One does not seem to be a better choice than the other. As you can see in figure 10, at low numbers of chains SS overestimates the evidence by about the same amount that TI underestimates it. Once again, after about ten chains are used, both become very accurate. It is interesting that `BayesLine`

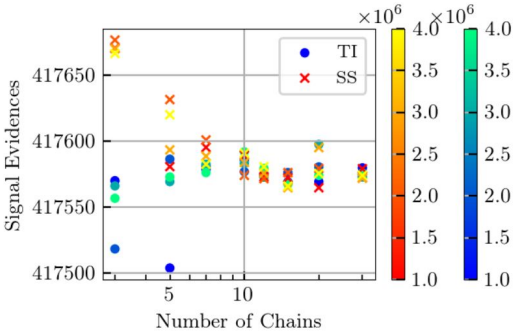

FIG. 10: Number of chains used for sampling vs the ratio between signal evidences and noise

seems to remove the advantage that SS had over TI during runs with `BayesLine` turned off. A question that arises here is whether TI and SS are producing the same accuracy of evidences when a large number of chains is used. In the future, a method for calculating SS's standard deviation will likely be incorporated.

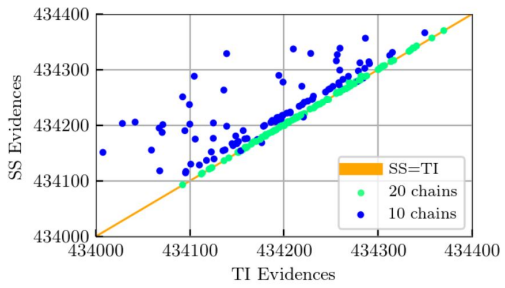

FIG. 11: TI evidences versus SS evidences (`BayesLine` off)

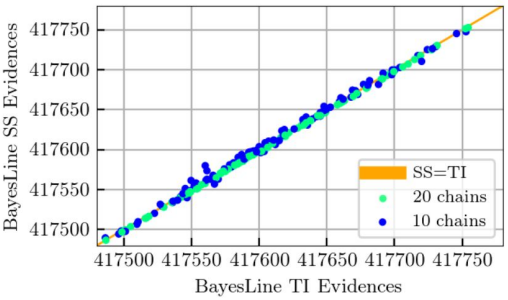

FIG. 12: TI evidences versus SS evidences (`BayesLine` on)

To see how well estimates evidences from TI and SS correlated, the next step was to do many runs with varying realizations of random noise. We simulate gaussian noise with different random seeds to obtain many noise realizations. With the same injected signal, runs at both 10 and 20 chains were conducted, with `BayesLine` both on and off. Then, plotting TI evidences on the X axis

and SS evidences on the Y axis provides a comparison between the two, where points will fall along the diagonal if the evidences are equivalent. When `BayesLine` is off, the results are exhibited by figure 11. At ten chains, as both sampling methods begin to converge at an accurate estimate, SS evidences are larger than TI evidences. This makes sense, as at low chains SS tends to overestimate evidences and TI tends to underestimate. However, when 20 chains are used, the evidences fall right along the diagonal, meaning that both SS and TI are producing accurate estimates, and either can be used safely for signal matching. Interestingly, when `BayesLine` is on, both TI and SS cluster close along the diagonal, as seen in 12. Even at low chains, they are well matched in accuracy, and both produce good results.

VII. CONCLUSIONS

The results of this project have shown that when `BayesLine` is off, the Stepping-Stone sampling algorithm is able to provide more accurate evidence estimates than Thermodynamic Integration with less computational work. However, it seems as though when `BayesLine` is turned on, the two methods are on par with one another. This brings to rise questions about under what circumstances SS is more accurate than TI, and when are the two equally effective. In future studies, it could be beneficial to construct more complicated run and compare the two. For example, runs in this project were done assuming a single signal or glitch within the data. Runs in which we assume that both a signal and a glitch are present will require more computational labor, and may present a different comparison.

It is likely that the Stepping-Stone branch of `BayesWave` will be incorporated into the main branch. It has been shown that the SS branch runs smoothly and contains no major bugs, and incorporation will make SS a more accessible tool that can be used by any `BayesWave` users for their runs. It would be beneficial in the future to be able to calculate the standard deviation on SS estimates. This will give us a better idea of how error estimation compares between SS and TI.

VIII. ACKNOWLEDGEMENTS

I'd like to thank Sophie Hourihane and Katerina Chatziioannou for their tremendous support in this project, as well as Meg Millhouse for her work on the stepping-stone branch of `BayesWave`. I also gratefully acknowledge the support from the National Science Foundation Research Experience for Undergraduates (NSF REU) program, the California Institute of Technology, and the LIGO Summer Undergraduate Research Fellowship."

[1] LIGO- A Gravitational-Wave Interferometer.
[2] t. K. C. R. A. e. a. The LIGO Scientific Collaboration, the Virgo Collaboration, GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo During the Second Part of the Third Observing Run, (2021).
[3] N. J. Cornish, T. B. Littenberg, B. Bécsy, K. Chatziioannou, J. A. Clark, S. Ghonge, and M. Millhouse, BayesWave analysis pipeline in the era of gravitational wave observations, Phys. Rev. D **103**, 044006 (2021), arXiv:2011.09494 [gr-qc].
[4] N. Lartillot and H. Phillipe, Computing Bayes' Factors Using Thermodynamic Integration (2006).
[5] J. Annis, Thermodynamic Integration and Steppingstone Sampling Methods for Estimating Bayes Factors: A Tutorial, Journal of mathematical psychology **89** (2019).
[6] J. S. Speagle, A Conceptual Introduction to Markov Chain Monte Carlo Methods (2020).
[7] P. Maturana-Russel, R. Meyer, J. Veitch, and N. Christensen, Stepping-stone sampling algorithm for calculating the evidence of gravitational wave models, Phys. Rev. D **99**, 084006 (2019), arXiv:1810.04488 [physics.data-an].
[8] W. Xie, P. Lewis, Y. Fan, L. Kuo, and M.-H. Chen, Improving Marginal Likelihood Estimation for Bayesian Phylogenetic Model Selection", url =.