Bayesian model selection with parallel-tempering

Steven Reyes

Bayes Theorem

$$p(\theta|M)\ p(D|\theta, M) = p(D|M)\, p(\theta|D, M)$$

$$\text{Prior} \times \text{Likelihood} = \text{Evidence} \times \text{Posterior}$$

$$\pi(\theta)\, \mathcal{L}(\theta) = \mathcal{EP}(\theta)$$

$$\text{Inputs} \rightarrow \text{Outputs}$$

- Adapted from: Ch1. of Bayesian Methods in Cosmology[1]

Bayes Theorem

$$p(\theta|M)\ p(D|\theta, M) = p(D|M)\, p(\theta|D, M)$$

$$\text{Prior} \times \text{Likelihood} = \text{Evidence} \times \text{Posterior}$$

$$\pi(\theta)\, \mathcal{L}(\theta) = \mathcal{E}\mathcal{P}(\theta)$$

$$\text{Inputs} \rightarrow \text{Outputs}$$

- Adapted from: Ch1. of Bayesian Methods in Cosmology[1]

What is the “evidence” and why does it matter?

- The more evidence your model has, the more credible that model is.
- The evidence is formally:

$$\mathcal{E} = \int \pi(\theta)\mathcal{L}(\theta)d\theta$$

- There are a LARGE number of possible models. We have to prioritize testing the most probable models.
- Can we numerically calculate or estimate the evidence for a model?

Numerical Methods for estimating the evidence

- Arithmetic Mean Estimator (AME): Sample from the prior, sum up the likelihood. This is the most straightforward evidence calculation.

- Thermodynamic Integration Estimator (TIE): Use parallel tempering to estimate the evidence. See refs [2-9].

- SteppingStone Estimator (SSE): Uses the same parallel tempering samples from TIE. Uses importance sampling for numerical integration method. See ref [5, 8].

Arithmetic Mean Estimator (AME) (1)

- The evidence is:

$$\mathcal{E} = \int \pi(\theta)\mathcal{L}(\theta)d\theta$$

- Estimate it by sampling from the prior and calculating:

$$\mathcal{E} \sim \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(\theta)$$

- Straightforward and easy to calculate.

Arithmetic Mean Estimator (AME) (2)

- The problems:
  - Tends to spend too much time across the prior-distribution where the likelihood is low.
  - Doesn’t spend enough time near the posterior-distribution where the likelihood is high.
- Characteristically underestimates the evidence.
- Choosing an estimator that samples only from the posterior-distribution will overestimate the evidence.
- We need an estimator that can efficiently calculate the evidence in a path from the prior-distribution to the posterior-distribution. This can be done using thermodynamic integration.

Calculating $\mathcal{E}$ through thermodynamic integration (1)

- Ordinarily a posterior is defined as:

$$\mathcal{P}(\theta) \propto \pi(\theta)\mathcal{L}(\theta)$$

- Consider a power-posterior defined as:

$$\mathcal{P}_\beta(\theta) \propto \pi(\theta)\,\mathcal{L}^\beta(\theta)$$

  for inverse temperature, β between 0 and 1.

- When β = 0, we sample from the prior. When β = 1, we sample from the posterior. By sampling continuously between 0 and 1, we can sample efficiently across the entire space.

Calculating $\mathcal{E}$ through thermodynamic integration (2)

- With some algebra and calculus we can find the log evidence as

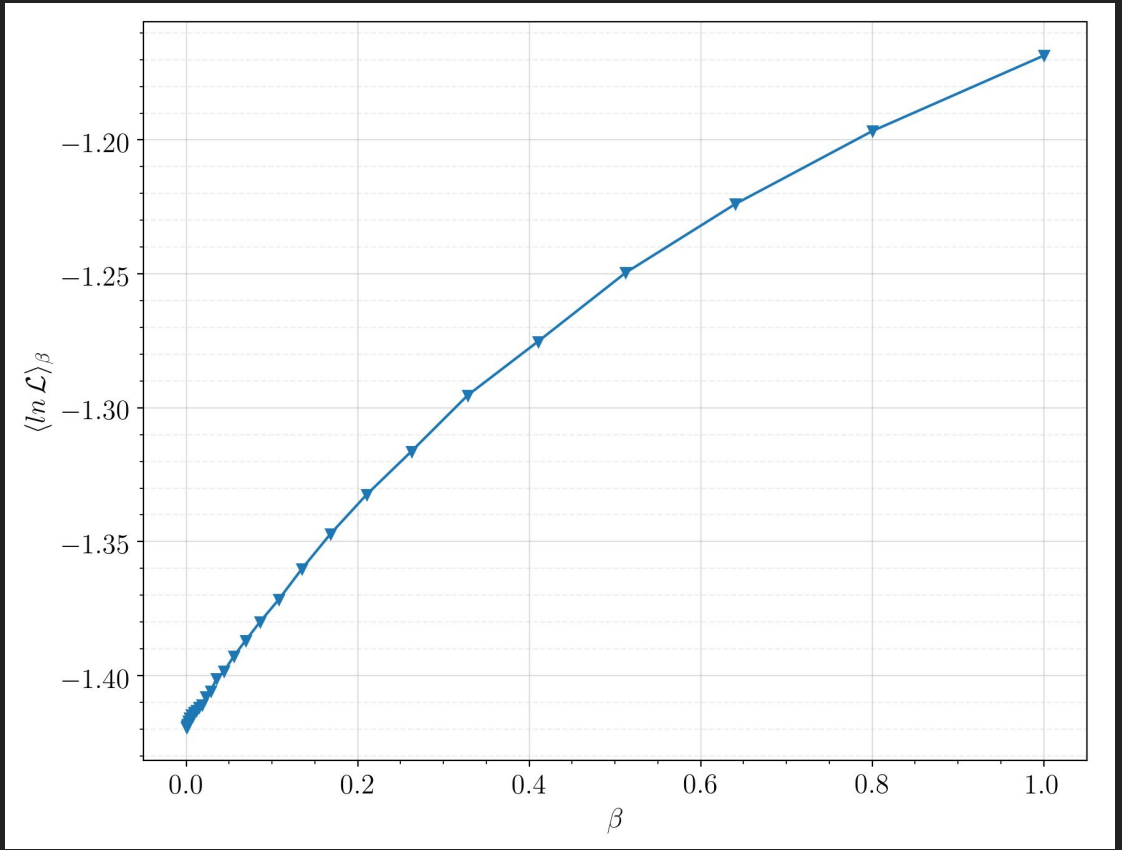
$$\int_0^1 \langle ln\,\mathcal{L}\rangle_\beta \, d\beta = \int_{-\infty}^0 \beta \, \langle ln\,\mathcal{L}\rangle_\beta \, d(ln\,\beta) = ln\,\mathcal{E}$$

- In practice we cannot sample from every β. We need to select a good enough set of β's that a numerical integrator can estimate the log evidence with minimal bias.

- Limiting the bias can be roughly estimated by plotting the integrand.

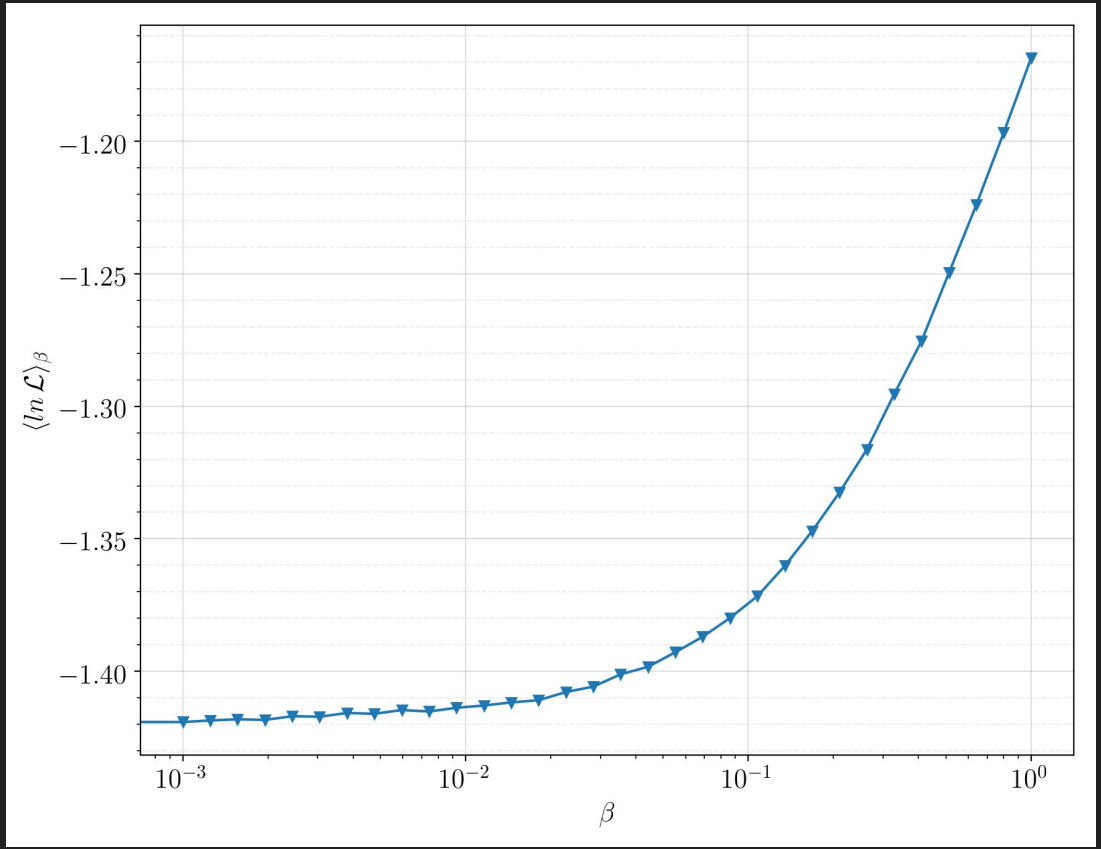
Investigating the TI Integrand (1)

- **Always** plot the TI integrand.
- Plot $\langle ln\,\mathcal{L} \rangle_\beta$ vs β on a linear scale.
- $\langle ln\,\mathcal{L} \rangle_\beta$ curve should be smooth and monotonic.



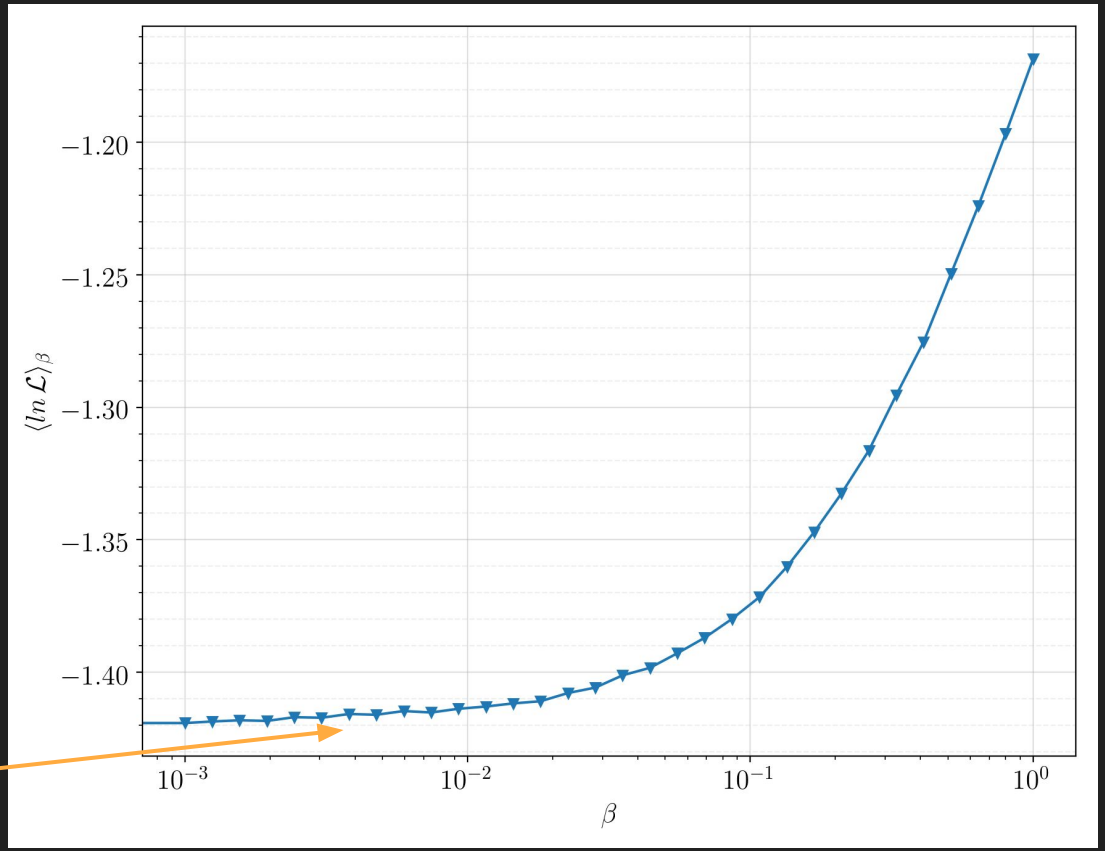
Investigating the TI Integrand (2)

- **Always** plot the TI integrand.
- Plot $\langle ln\,\mathcal{L} \rangle_\beta$ vs β on a log scale.
- $\langle ln\,\mathcal{L} \rangle_\beta$ curve should be smooth and monotonic.
- Sometimes see fluctuations of $\langle ln\,\mathcal{L} \rangle_\beta$, near small β.



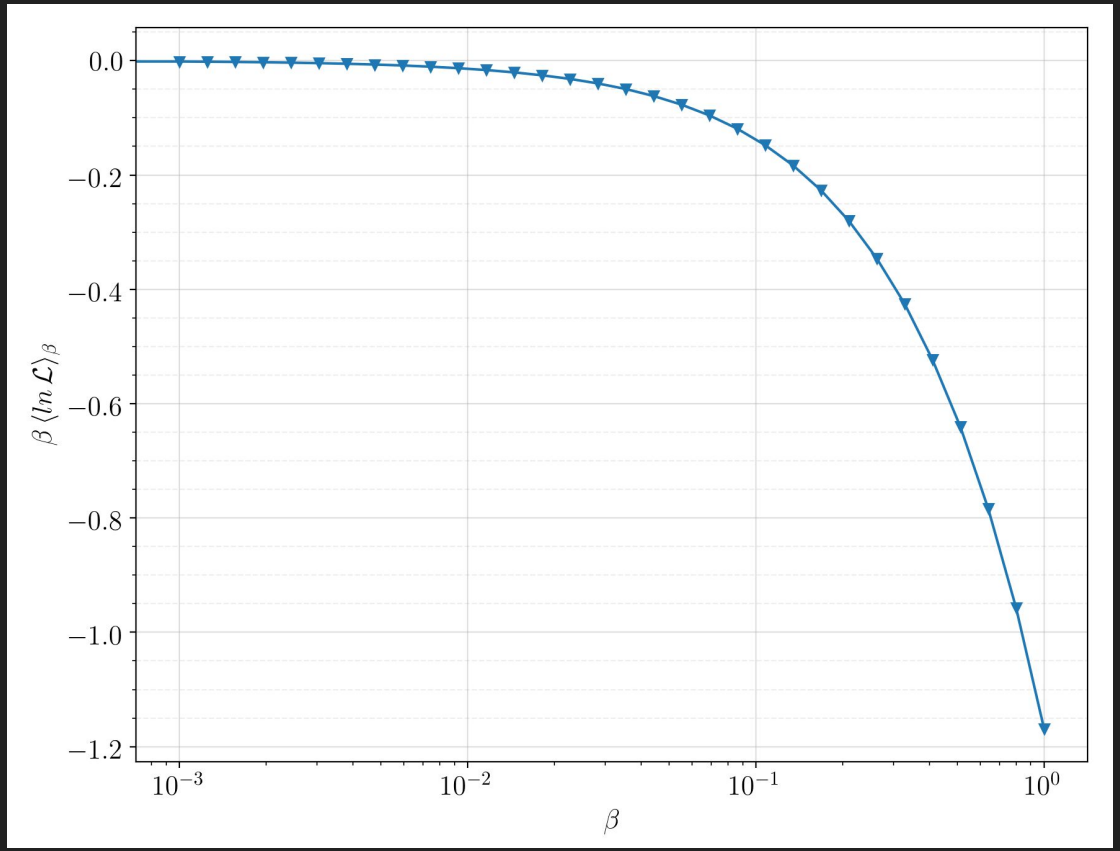
Investigating the TI Integrand (2)

- **Always** plot the TI integrand.
- Plot $\langle ln\,\mathcal{L} \rangle_\beta$ vs β on a log scale.
- $\langle ln\,\mathcal{L} \rangle_\beta$ curve should be smooth and monotonic.
- Sometimes see fluctuations of $\langle ln\,\mathcal{L} \rangle_\beta$, near small β.



Investigating the TI Integrand (3)

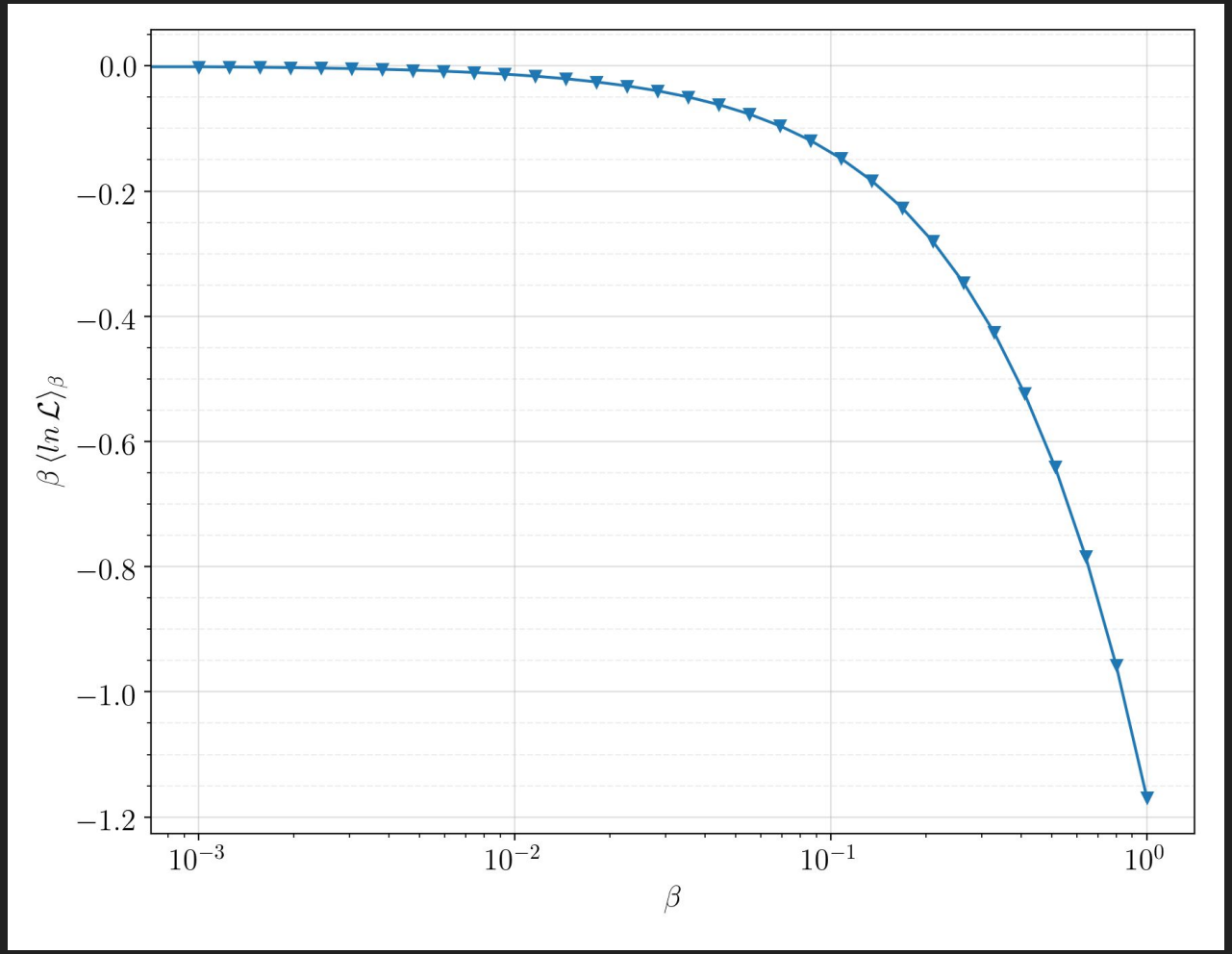
- **Always** plot the TI integrand.
- Plot $\beta \langle ln\, \mathcal{L} \rangle_\beta$ vs β on a log scale.
- $\beta \langle ln\, \mathcal{L} \rangle_\beta$ is monotonic and smooth.
- Curve looks good! Could use more dense sampling in $\beta \in (0.1, 1)$, or $\beta < 10^{-3}$



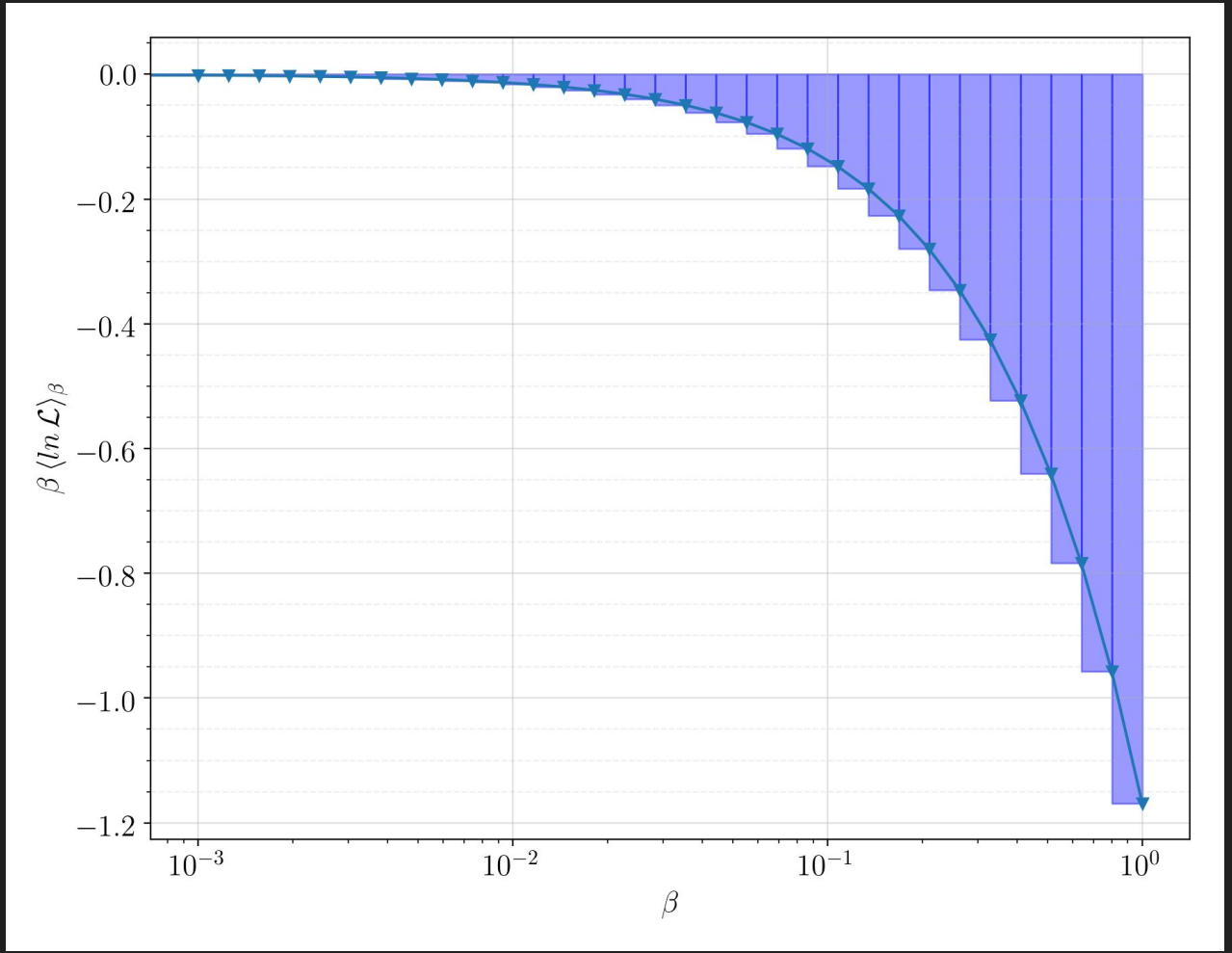
Designing a temperature ladder / schedule

- “Setting the [temperature] schedule is something of a black art.” - John Skilling[1]
- A geometric distribution of temperatures seems to work well as a first trial temperature placement[3]. No known analytic solution for placement[2-8].
- In practice, using parallel tempering means running your analysis twice (or more) to accurately calculate the evidence[8, 9].
- Inspect integrand plot, and look for regions that appear to be undersampled.
- Place temperatures by hand in undersampled regions[8, 9].

Calculating TI with Riemann Sums

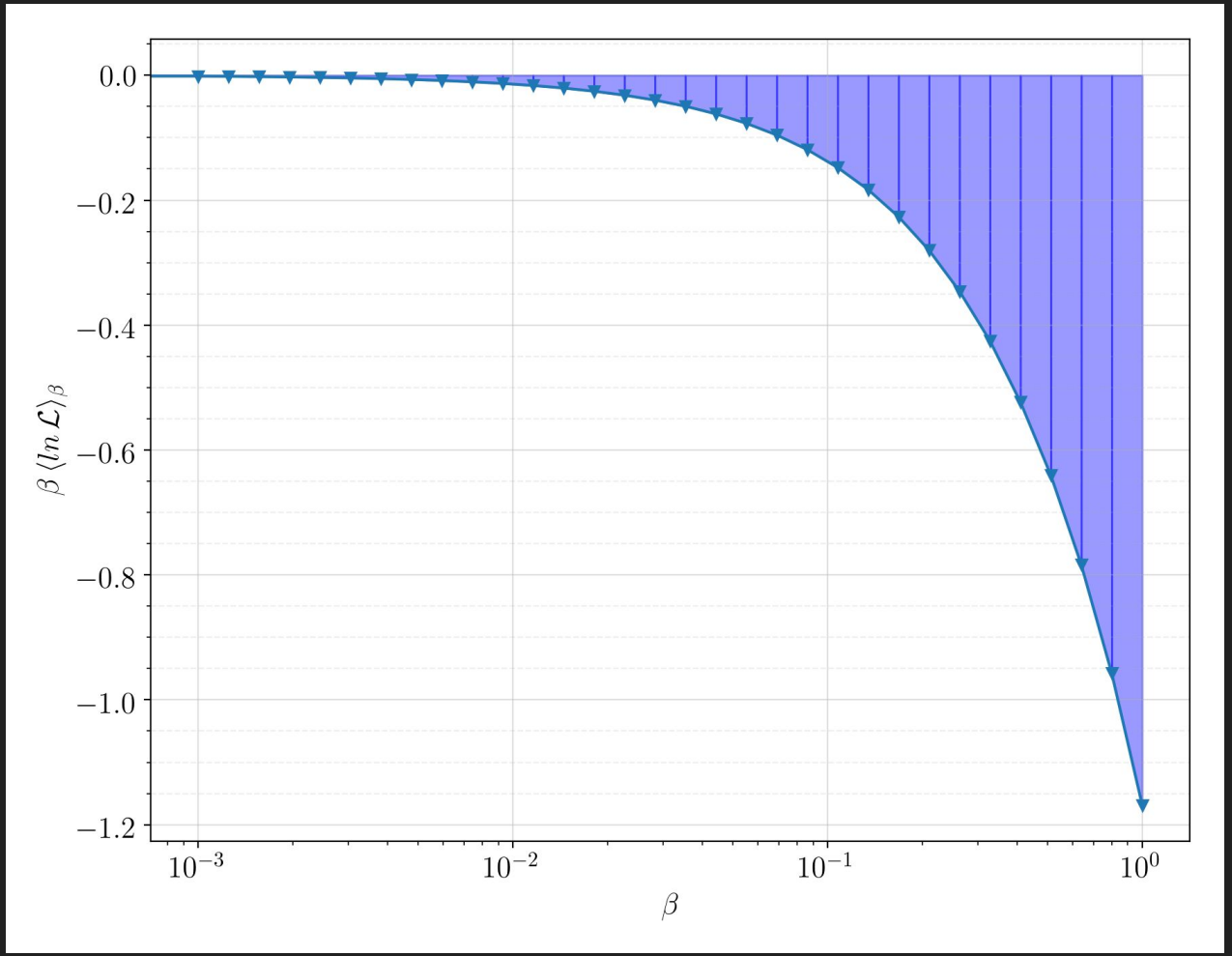


Calculating TI with Riemann Sums



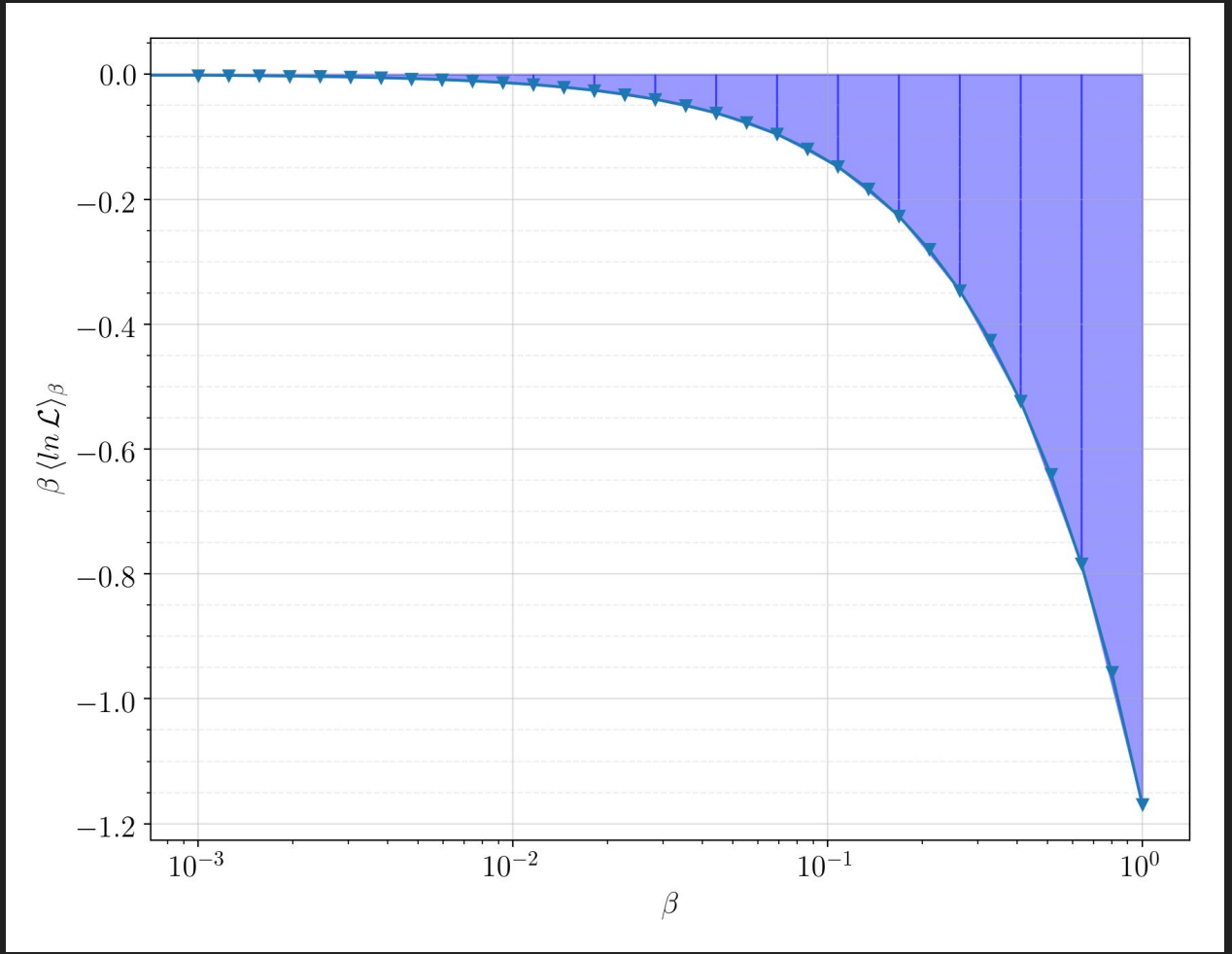
Calculating TI with Trapezoid Rule

- We can calculate composite trapezoid rule[2].
- We can also calculate the composite trapezoid rule with $\mathcal{O}(h^3)$ error corrective term[3,4]. (not pictured)



Calculating TI with Simpson's Rule

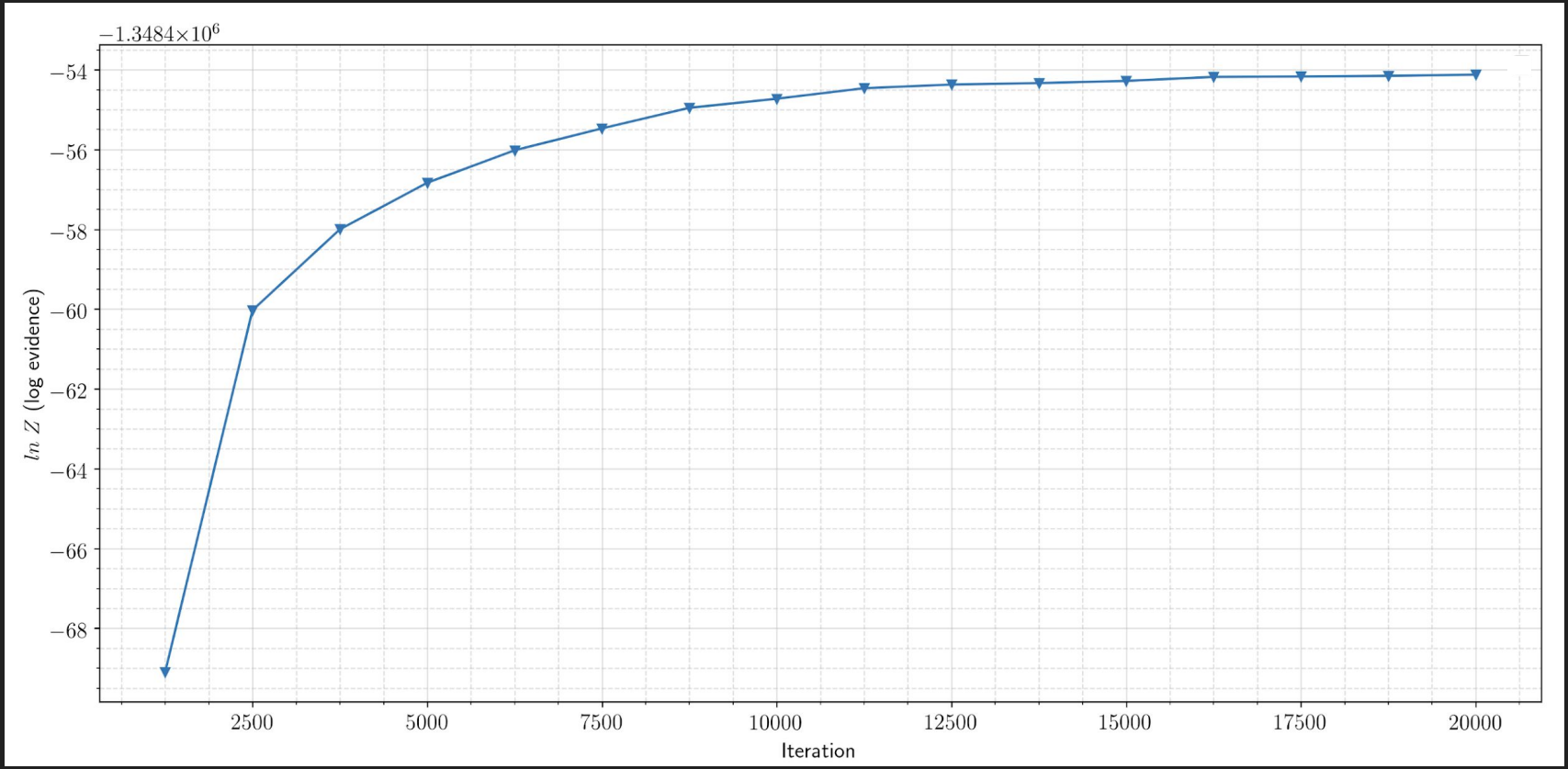
- Simpson's (⅓) rule for arbitrary β placement.
- Simpson's rule with $\mathcal{O}(h^4)$ corrective term: Coming soon!



Convergence Evidence Error Estimate (1)

- First issue: Did the evidence even converge?
- Track the evidence as a function of MCMC iteration, being careful to only draw independent samples.
- Investigate whether $\frac{d(ln\,\mathcal{E})}{d(\mathrm{Iteration})} \rightarrow 0$.
- Take the difference between last two sub-analysis chunks as the convergence error, σ_{conv}.

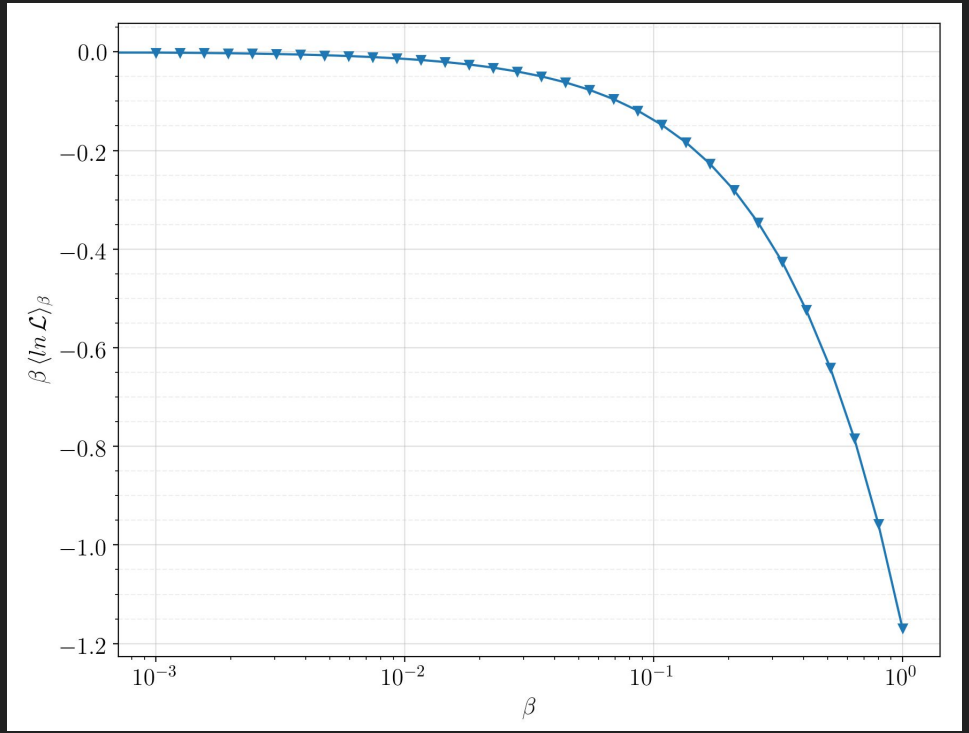
Convergence Evidence Error Estimate (2)



Statistical Evidence Error Estimate

- Second Issue: What's the statistical error?
- Monte Carlo Error Estimates[8]:
  - Can be done empirically from multiple re-runs. Expensive.
  - We can take all the samples and calculate a sample variance of TI integral for every sample.
  - From the sample variance, estimate the standard error (variance of mean).

$$\sigma^2_{\mathrm{MC}} \equiv \sigma^2_{\mathrm{mean}} = \frac{1}{N}\,\sigma^2_{\mathrm{samples}}$$



Combining the Error Estimates

- Add the MCMC error and convergence error in quadrature:

$$\sigma_{\text{error}} = \sqrt{\sigma_{MC}^2 + \sigma_{conv}^2}$$

- For a truly converged MCMC this double counts the statistical error.

- Better safe than sorry!

Systematic Error / Bias:

- Systematic error or bias may not be accounted for by previous method.
- Going to higher order numerical integration should account for some bias in thermodynamic integration estimator.
- If the various estimators don't agree, check the temperature ladder. More temperatures, better placement, or wait longer for convergence.
- Best defense is to use multiple estimators, not just parallel tempering, when possible, and ensure that the evidence estimates agree.

Feeding a temperature ladder to PyCBC

```
import numpy
import h5py
```

Generate a simple temperature ladder

```
betas = numpy.geomspace(0.001, 1., 32, endpoint=True)
betas = numpy.concatenate([[0], betas])
betas = numpy.sort(betas)

print betas
```

```
[0.         0.001      0.00124961 0.00156152 0.00195129 0.00243835
 0.00304699 0.00380755 0.00475794 0.00594557 0.00742964 0.00928415
 0.01160155 0.01449741 0.01811609 0.02263803 0.02828869 0.03534981
 0.04417345 0.05519954 0.06897785 0.08619536 0.10771051 0.13459603
 0.16819243 0.2101748  0.26263635 0.32819279 0.41011271 0.51248059
 0.64040043 0.80025023 1.        ]
```

```
f = h5py.File("geometric_dense_betas.hdf", "w")
f.attrs["betas"] = betas
f.close()
```

Feeding a temperature ladder to PyCBC

```
%cat priors.ini
```

```
[model]
name = test_normal

[sampler]
name = emcee_pt
inverse-temperatures-file = /home/steven.reyes/projects/pycbc_pe_workshop_model_selection/gaussian_example/temperature_ladde
rs/geometric_dense_betas.hdf
nwalkers = 500
niterations = 5000

[variable_params]
x =

[prior-x]
name = gaussian
x_mean = 0.0
x_var = 1.0
min-x = -10
max-x = 10
```

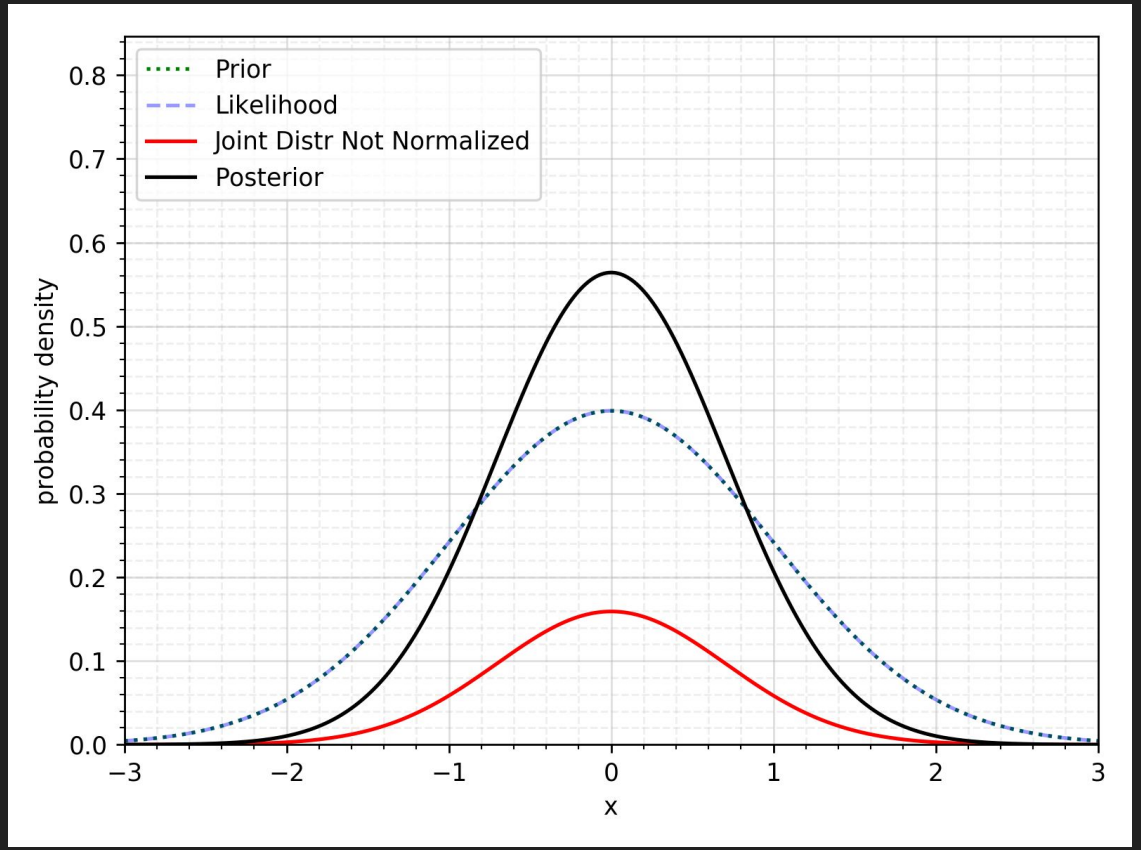
Feeding a temperature ladder to PyCBC

```
%cat run_pycbc_inf.sh
```

```
#!/bin/sh
pycbc_inference --verbose \
        --config-files priors.ini \
        --output-file thirty_three_temps_gauss.hdf \
        --nprocesses 11 \
        --seed 10 \
        --force
```

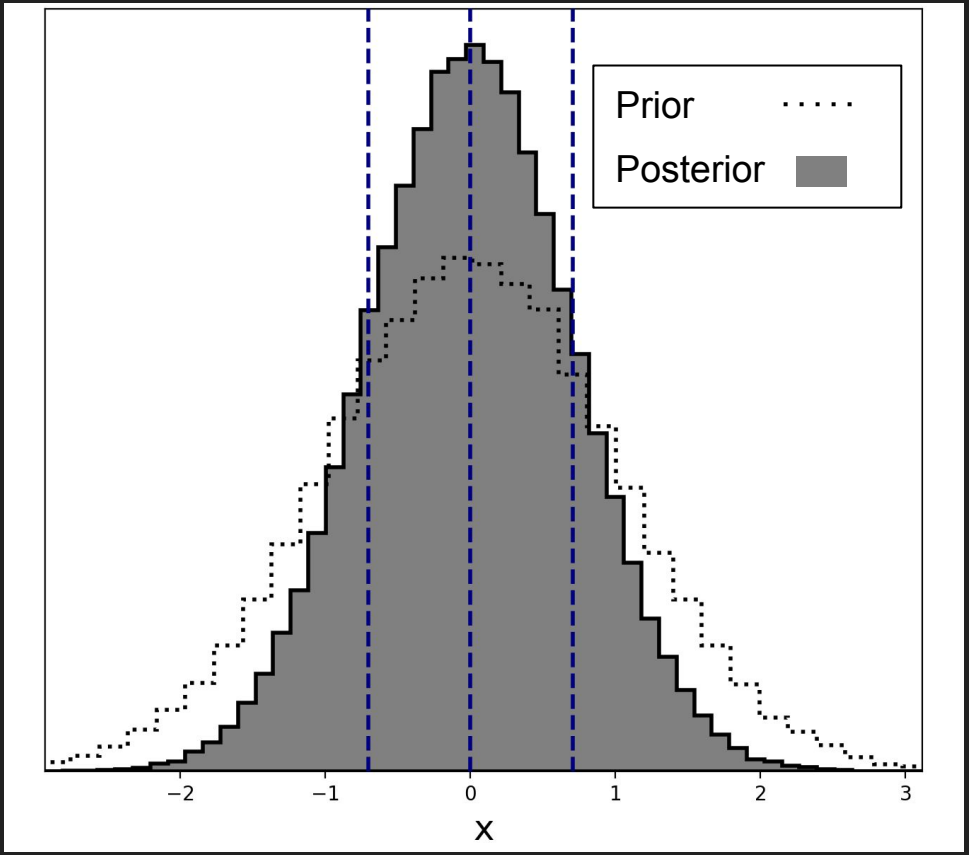
A Toy Model: Gauss Prior, Gauss Likelihood

- Prior is $N(\mu_f=0, \sigma_f^2=1)$
- Likelihood is $N(\mu_g=0, \sigma_g^2=1)$
- Joint Distribution = Prior x Likelihood
- Posterior = Joint / Evidence
- The normalizing constant or Evidence is $1/\mathrm{Sqrt}(4\pi) \sim 0.28209479177...$



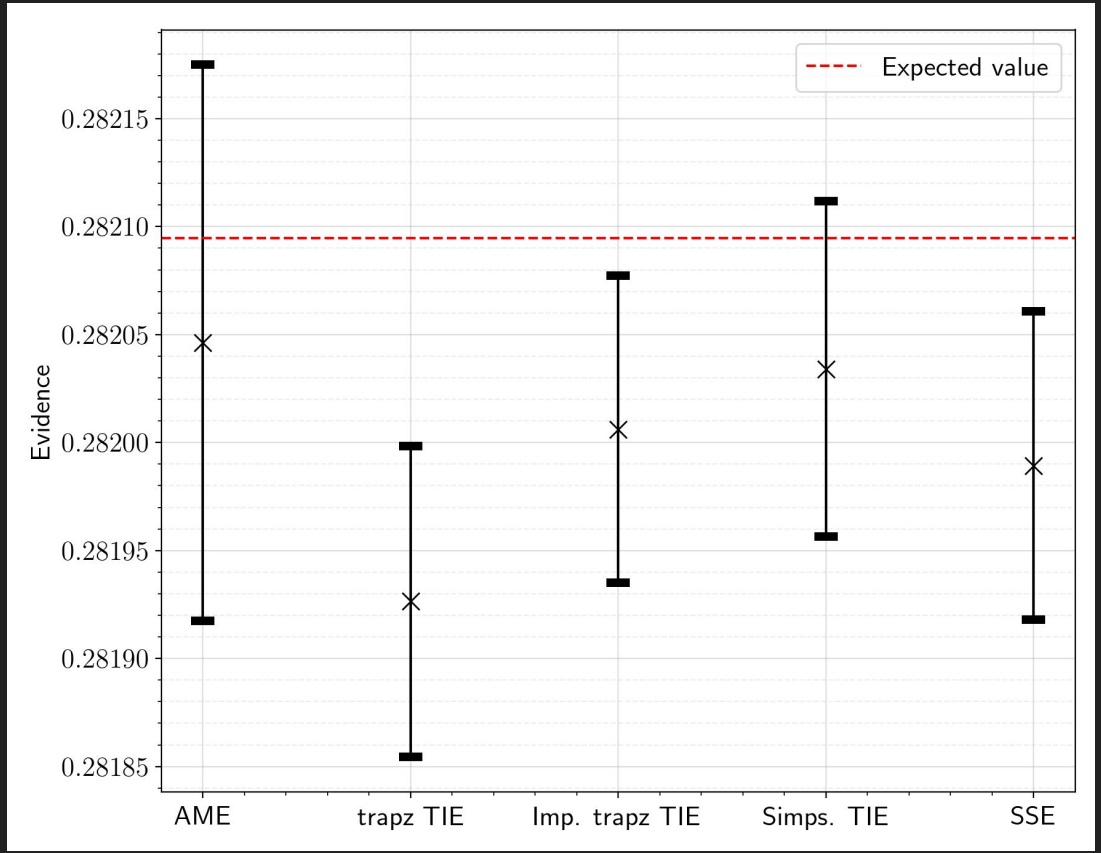
A Toy Model: Gauss Prior, Gauss Likelihood (2)

- PyCBC Inference did a pretty good job!
- It does this without calculating the evidence!
- What happens if we try to calculate the evidence?



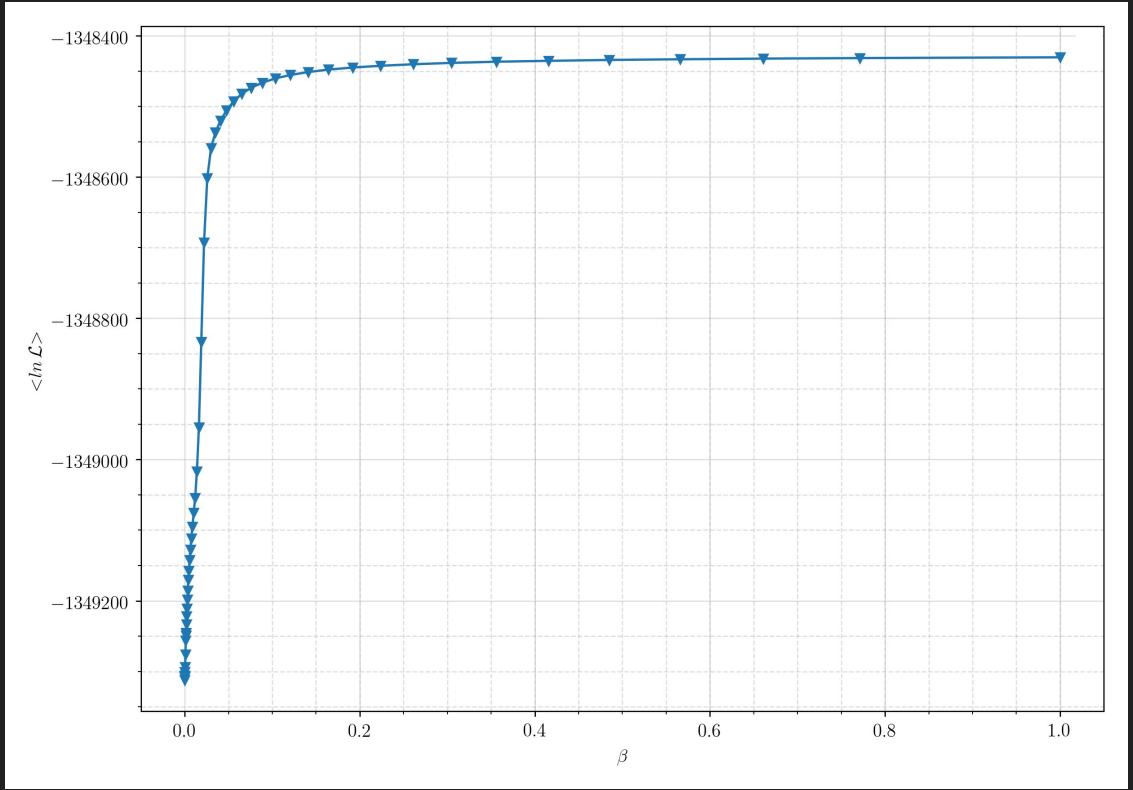
A Toy Model: Gauss Prior, Gauss Likelihood (3)

- True Evidence is 0.28209479...
- 1,250,000 independent samples.
- Use 33 geometrically spaced temperatures, $\beta \in 0$ & (0.001, 1).
- $\mathcal{O}$(<0.05%) bias. For simplicity, only show estimated statistical error bars.



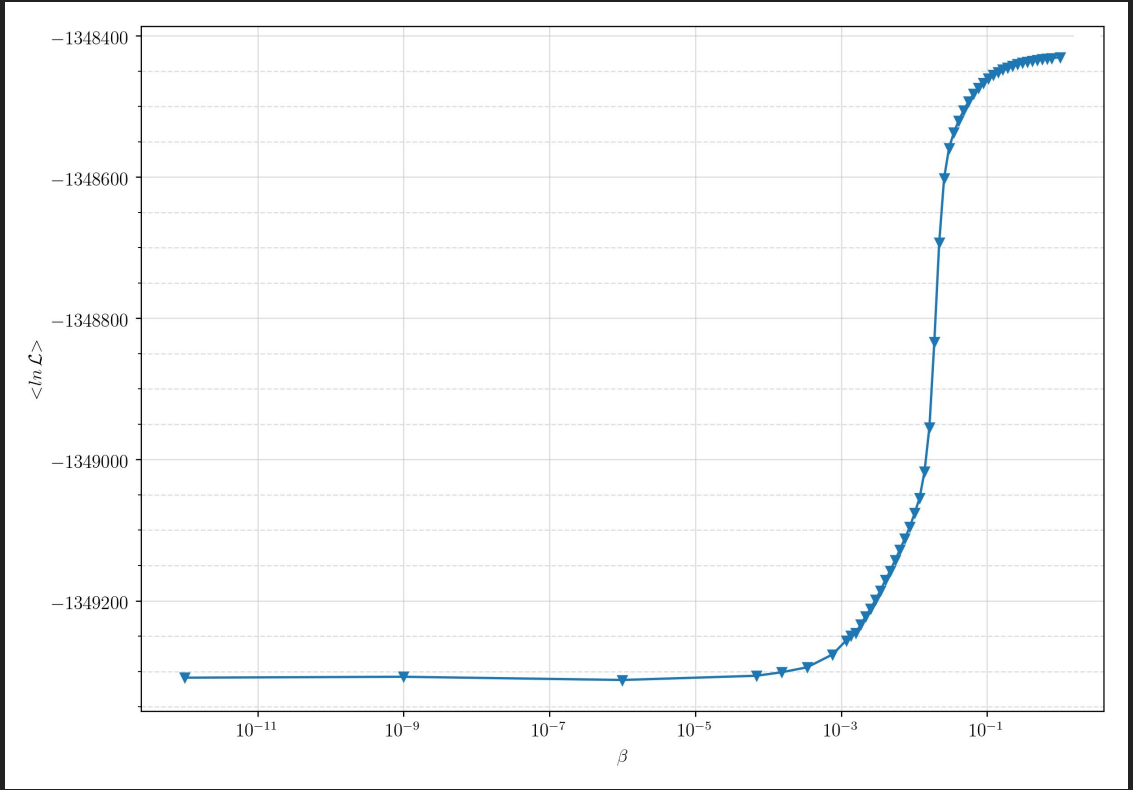
Real Model: GW170817 Common EOS[10]

- Use 51 linearly, logarithmically, and geometrically spaced temperatures, where $\beta \in (10^{-12}, 1)$.
- The $\langle ln\,\mathcal{L} \rangle_\beta$ vs β plot on a linear scale.



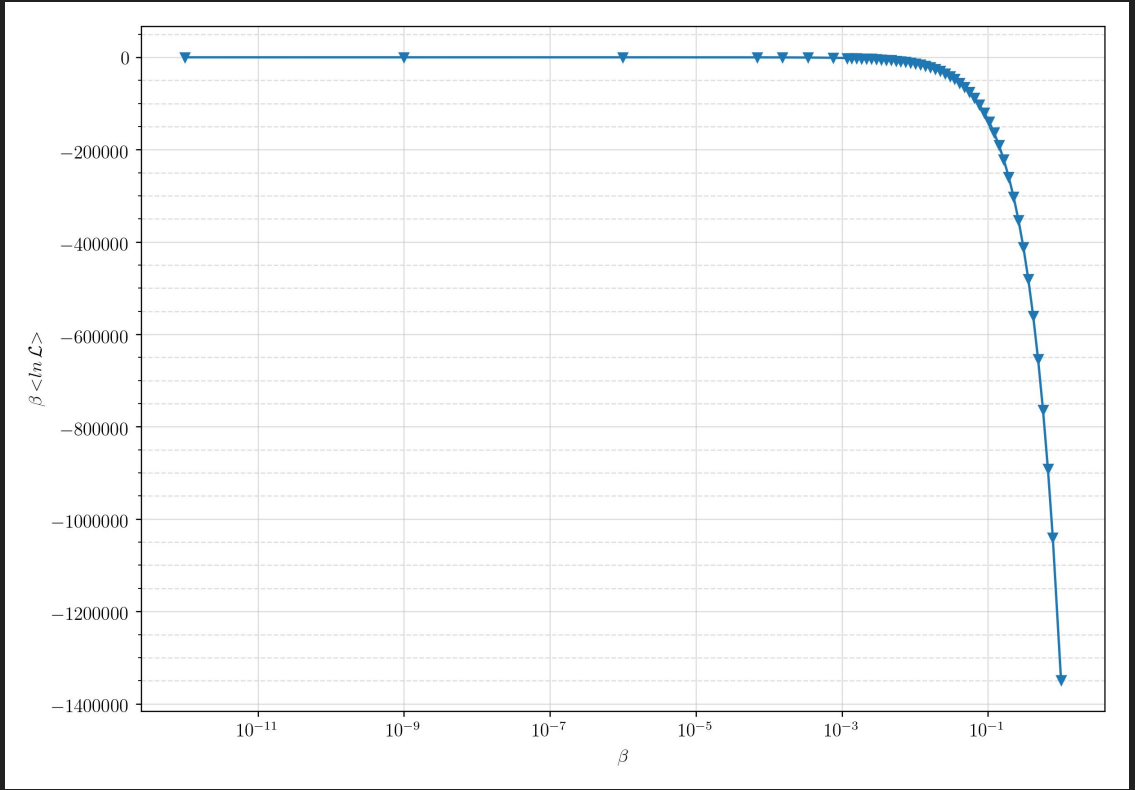
Real Model: GW170817 Common EOS[10]

- Use 51 linearly, logarithmically, and geometrically spaced temperatures, where $\beta \in (10^{-12}, 1)$.
- The $\langle ln\,\mathcal{L} \rangle_\beta$ vs β plot on a log scale.



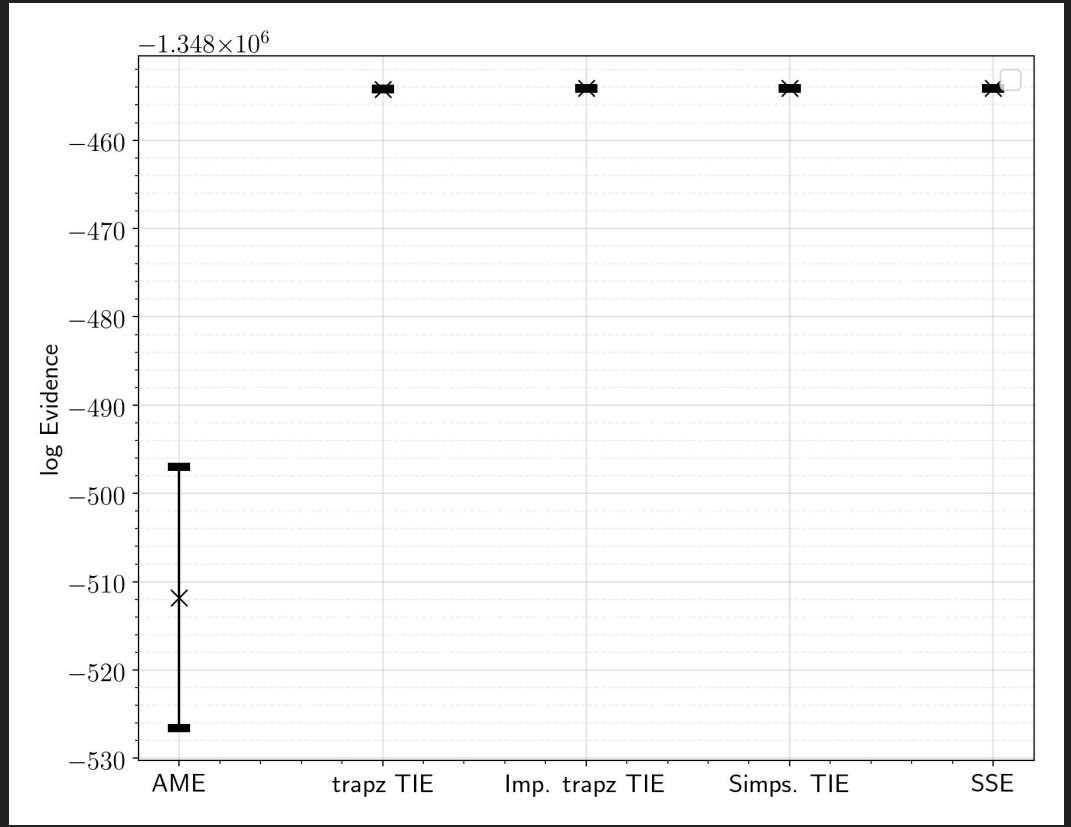
Real Model: GW170817 Common EOS[10]

- Use 51 linearly, logarithmically, and geometrically spaced temperatures, where $\beta \in (10^{-12}, 1)$.
- The $\beta \langle ln\, \mathcal{L} \rangle_\beta$ vs β plot on a log scale.



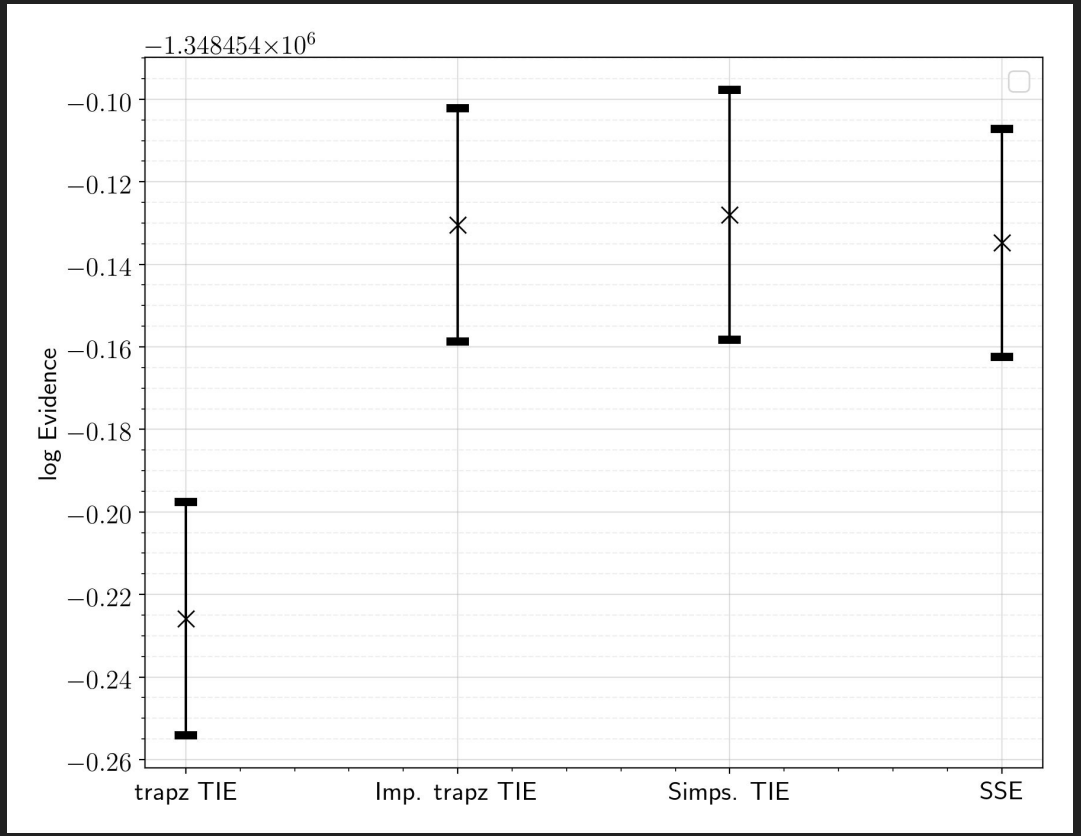
Real Model: GW170817 Common EOS[10]

- 5 methods to try to estimate the evidence. Only show MCMC error for simplicity.
- Arithmetic Mean Estimator (AME) using 4,000,000 samples
- Parallel tempering have 18,800 independent samples.



Real Model: GW170817 Common EOS[10]

- Error between integration methods is small.
- Relatively well-designed temperature ladder.
- Denser temperature placement could close the gap.



Calculating a Bayes factor

- The ratio of evidences is the Bayes Factor:

  Bayes Factor = (Evidence 1) / (Evidence 2)

- How much more support is there in the data for model 1 vs model 2?

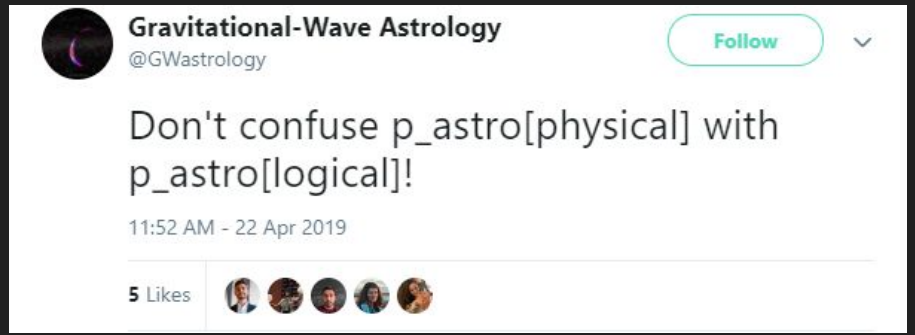
- Posterior Odds Ratio = Bayes Factor x Prior Odds Ratio

- If you think both prior model choices are equally likely models, set the prior odds ratio to 1.

- Think carefully about what prior distributions you want to choose, and what their prior odds are.

Reporting Bayes Factors

- Does the Bayes Factor make sense with respect to the log-likelihoods (signal-to-noise ratios) you measured?
- Generally, the bigger the Bayes Factor, the more obvious it should seem that one model wasn't a good fit to the data.
- Use your physics intuition when possible!



Future Development

- Add more methods into PyCBC.
- Comparisons with nested sampling and other techniques.
- Model selection with likelihood ratios between two models, straight to Bayes Factors, one run.
- Improve run-time.

References

1. *Hobson, M. P., Jaffe, A. H., Liddle, A. R., Mukherjee, P., & Parkinson, D. (Eds.). (2010). Bayesian methods in cosmology. Cambridge University Press.*

2. *Lartillot, N. and Philippe, H. 2006, Systematic biology, 55(2):195-207*

3. *Behrens, G., Friel, N., & Hurn, M. (2012). Tuning tempered transitions. Statistics and computing, 22(1), 65-78.*

4. *Friel, N., Hurn, M., & Wyse, J. 2014, Statistics and Computing, 24(5), 709-723.*

5. *Xie, W., Lewis, PO., Fan, Y., Kuo, L., Chen, M-H. 2011, Systematic biology, 60:150-160*

6. *Calderhead, B., & Girolami, M. (2009). Estimating Bayes factors via thermodynamic integration and population MCMC. Computational Statistics & Data Analysis, 53(12), 4028-4045.*

7. *Lefebvre, G., Steele, R., & Vandal, A. C. (2010). A path sampling identity for computing the Kullback–Leibler and J divergences. Computational Statistics & Data Analysis, 54(7), 1719-1731.*

8. *Annis, J., Evans, N. J., Miller, B. J., & Palmeri, T. J. (2019). Thermodynamic integration and steppingstone sampling methods for estimating Bayes factors: A tutorial. Journal of mathematical psychology, 89, 67-86.*

9. *Liu, P., Elshall, A. S., Ye, M., Beerli, P., Zeng, X., Lu, D., & Tao, Y. (2016). Evaluating marginal likelihood with thermodynamic integration method and comparison with several other numerical methods. Water Resources Research, 52(2), 734-758.*

10. *De, S., Finstad, D., Lattimer, J.M., et al. 2018, Phys. Rev. Lett., 121, 091102*

Thanks!

*Thanks to Chaitanya Afle, Duncan Brown, Derek Davis, Soumi De, Daniel Finstad, and Varun Srivastava for helpful comments and suggestions on this talk.

Computational work was supported by Syracuse University and National Science Foundation grant OAC-1541396

Steven Reyes is supported by National Science Foundation grant PHY1707954.

This talk is based on data provided by the Gravitational Wave Open Science Center.

Future Development Detailed

- Add more Evidence methods and Information Criterion model selection methods into PyCBC.
- Could use method of minimizing information gain (Kullback-Leibler divergence) between power-posteriors for temperature ladder selection[3,4,6,7].
- Could use Geweke statistic for convergence estimates.
- Comparisons with nested sampling and other techniques.
- Model selection with likelihood ratios between two models, straight to Bayes Factors, one run.
- Improve run-time.
- Model selection without computation of evidence?? RJMCMC??

Helpful Resources for learning Bayesian Inference

1. Trotta, R. (2008). Bayes in the sky: Bayesian inference and model selection in cosmology. Contemporary Physics, 49(2), 71-104.
   [Great Introduction]

2. Hobson, M. P., Jaffe, A. H., Liddle, A. R., Mukherjee, P., & Parkinson, D. (Eds.). (2010). Bayesian methods in cosmology. Cambridge University Press.
   [Seems somewhat outdated on Thermodynamic Integration, but still good.]

3. Annis, J., Evans, N. J., Miller, B. J., & Palmeri, T. J. (2019). Thermodynamic integration and steppingstone sampling methods for estimating Bayes factors: A tutorial. Journal of mathematical psychology, 89, 67-86.
   [Very approachable read on thermodynamic integration and steppingstone with all the necessary references within.]

Derivation of the log $\mathcal{E}$ for TI (1)

- Consider a power-posterior defined as:

$$\mathcal{P}_\beta(\theta) \propto \pi(\theta)\,\mathcal{L}^\beta(\theta)$$

  for inverse temperature, β between 0 and 1.

- The evidence for a power-posterior is:

$$\mathcal{E}_\beta = \int p(\theta)\mathcal{L}^\beta(\theta)d\theta$$

Derivation of the log $\mathcal{E}$ for TI (2)

- Consider the following expression from the 2nd Fundamental Theorem of Calculus:

$$\int_0^1 \frac{d(ln\,\mathcal{E}_\beta)}{d\beta} d\beta = ln\,\mathcal{E}_{\beta=1} - ln\,\mathcal{E}_{\beta=0}$$

- This simplifies to:

$$\int_0^1 \frac{d(ln\,\mathcal{E}_\beta)}{d\beta} d\beta = ln\,\mathcal{E}$$

$$\int_0^1 \frac{1}{\mathcal{E}_\beta} \frac{d\mathcal{E}_\beta}{d\beta} d\beta = ln\,\mathcal{E}$$

Derivation of the log $\mathcal{E}$ for TI (3)

- Which becomes:

$$\int_0^1 \frac{\int \frac{d}{d\beta}[\pi(\theta)\mathcal{L}^\beta]d\theta}{\int \pi(\theta)\mathcal{L}^\beta d\theta} d\beta = ln\,\mathcal{E}$$

$$\int_0^1 \frac{\int [ln\,\mathcal{L}(\theta)]\,\pi(\theta)\mathcal{L}^\beta d\theta}{\int \pi(\theta)\mathcal{L}^\beta d\theta} d\beta = ln\,\mathcal{E}$$

- And voila:

$$\int_0^1 \langle ln\,\mathcal{L}\rangle_\beta d\beta = ln\,\mathcal{E}$$

Numerical techniques for calculating TIE

- Riemann sums ($\langle ln\,\mathcal{L}\rangle_\beta$ is arithmetic mean value at β):

$$ln\,\mathcal{E} \sim \sum_{i=1}^{N_\beta - 1} (\beta_{i+1} - \beta_i)\langle ln\,\mathcal{L}\rangle_{\beta_i}$$

- Trapezoid rule[2]:

$$ln\,\mathcal{E} \sim \sum_{i=1}^{N_\beta - 1} \frac{(\beta_{i+1} - \beta_i)}{2} \left(\langle ln\,\mathcal{L}\rangle_{\beta_{i+1}} + \langle ln\,\mathcal{L}\rangle_{\beta_i}\right)$$

- Trapezoid rule with $\mathcal{O}(h^3)$ error corrective term[3, 4]:

$$ln\,\mathcal{E} \sim \mathrm{trapz} - \sum_{i=1}^{N_\beta - 1} \frac{(\beta_{i+1} - \beta_i)^2}{12} \left(\mathrm{Var}[ln\,\mathcal{L}_{\beta_{i+1}}] - \mathrm{Var}[ln\,\mathcal{L}_{\beta_i}]\right)$$

Numerical techniques for calculating TIE

- Simpson's rule for arbitrary β placement: See Scipy! Algorithm looks roughly like this, but Scipy implementation is smarter with edge cases...

$$ln\,\mathcal{E} \sim \sum_{i=1,\, i\neq \text{odd}}^{N_\beta - 2} \frac{(h_i + h_{i+1})}{6} \left[\frac{(2h_i - h_{i+1})}{h_i} \langle ln\,\mathcal{L} \rangle_{\beta_i} + \frac{(h_i + h_{i+1})^2}{h_i\, h_{i+1}} \langle ln\,\mathcal{L} \rangle_{\beta_{i+1}} + \frac{(2h_{i+1} - h_i)}{h_{i+1}} \langle ln\,\mathcal{L} \rangle_{\beta_{i+2}} \right]$$

for:

$$h_i = \beta_{i+1} - \beta i$$

Above expression adapted from: Easa, S. M. (1988). Area of irregular region with unequal intervals. Journal of Surveying Engineering, 114(2), 50-58.

- Simpson's rule with $\mathcal{O}(h^4)$ corrective term: Coming soon!

  See slide 21 of talk by Merrilee Hurn, https://warwick.ac.uk/fac/sci/statistics/crism/workshops/estimatingconstants/hurn.pdf

Tracking Error more Robustly

- Good method for estimating error is the Root Mean Square Error (RMSE):

$$\mathrm{RMSE} = \sqrt{E[(ln\,\mathcal{E} - ln\,\mathcal{E}_{\mathrm{truth}})]^2}$$

$$\mathrm{RMSE} = \sqrt{\mathrm{Var}(ln\,\mathcal{E}) + (E[ln\,\mathcal{E}] - ln\,\mathcal{E}_{\mathrm{truth}})^2}$$

- Only possible when you know what the true log evidence is.

- If E[ln $\mathcal{E}$] - ln $\mathcal{E}_{\mathrm{truth}}$ = 0, the estimator is said to be unbiased.

Derivation of toy model evidence (1)

- Prior is $\pi(x) = N(\mu_f=0, \sigma_f^2=1)$

$$\pi(x) = \frac{1}{\sqrt{2\pi\sigma_f^2}} \exp\left[-\frac{(x-\mu_f)^2}{2\sigma_f^2}\right]$$

- Likelihood is $\mathscr{L}(x) = N(\mu_g=0, \sigma_g^2=1)$

$$\mathcal{L}(x) = \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp\left[-\frac{(x-\mu_g)^2}{2\sigma_g^2}\right]$$

- Joint distribution is the product of two Gaussians.

Derivation of toy model evidence (2)

- Joint distribution is the product of two Gaussians, see (http://www.tina-vision.net/docs/memos/2003-003.pdf) for derivation:

$$\pi(x)\,\mathcal{L}(x) = \frac{1}{\sqrt{2\pi\sigma_{fg}^2}} \times \left[\frac{1}{\sqrt{2\pi\left(\sigma_f^2+\sigma_g^2\right)}}\exp\left[-\frac{-(\mu_f-\mu_g)^2}{2\sigma_{fg}^2}\right]\right] \exp\left[-\frac{(x-\mu_{fg})^2}{2\sigma_{fg}^2}\right]$$

for:

$$\mu_{fg} = \frac{\mu_f\,\sigma_g^2+\mu_g\sigma_f^2}{\sigma_f^2+\sigma_g^2} \qquad \sigma_{fg} = \sqrt{\frac{\sigma_f^2\sigma_g^2}{\sigma_f^2+\sigma_g^2}}$$

Derivation of toy model evidence (3)

- $\mu_f = \mu_g = \mu_{fg} = 0$

  $\sigma^2_f = \sigma^2_g = 1$, $\sigma^2_{fg} = (½)$,

$$\pi(x)\,\mathcal{L}(x) = \frac{1}{\sqrt{2\pi(1/2)}} \times \left[\frac{1}{\sqrt{2\pi\,(1+1)}}\exp\left[-\frac{-(\quad 0)^2}{2(1/2)}\right]\right] \exp\left[-\frac{(x-0)^2}{2(1/2)}\right]$$

$$\pi(x)\,\mathcal{L}(x) = \frac{1}{2\pi}\exp\left[-x^2\right]$$

- Evidence is the normalizing factor for the Joint distribution

Derivation of toy model evidence (4)

- Evidence is a Gaussian integral

$$\mathcal{E} = \int f(x)\, g(x) dx = \int_{-\infty}^{\infty} \frac{1}{2\pi} \exp\left[-x^2\right] dx$$

$$\mathcal{E} = \frac{\sqrt{\pi}}{2\pi} = \frac{1}{\sqrt{4\pi}}$$