Bayesian model selection for likelihood-based and simulation-based inference

Jason D. McEwen
www.jasonmcewen.org

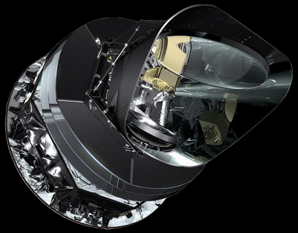
Mullard Space Science Laboratory (MSSL), University College London (UCL)

IAU-IAA Astrostats & Astroinfo seminar, October 2022

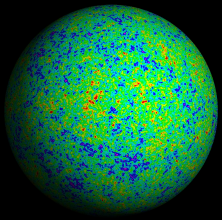
Observable Universe

Cosmic microwave background (CMB) radiation

What is the origin of structure in our Universe?



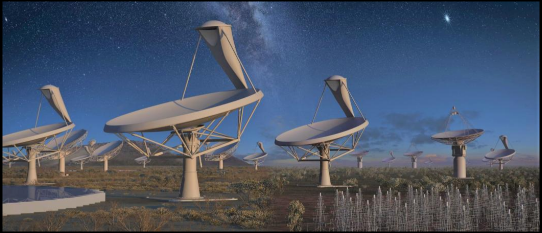
Planck satellite



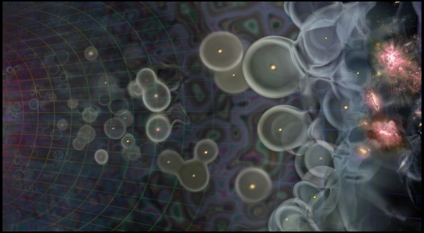
CMB

Epoch of reionisation

How did the first luminous objects in the Universe form?



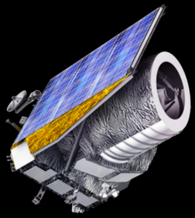
Square Kilometre Array (SKA)



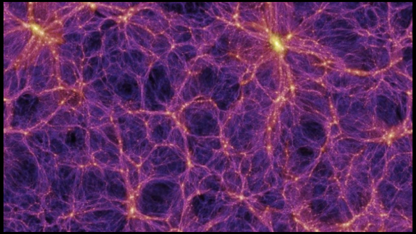
Ionised bubbles in neutral hydrogen

Large-scale structure of the Universe

What is the nature of dark energy?



Euclid satellite

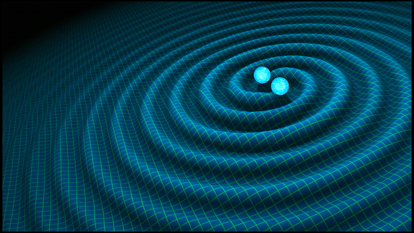


Large-scale structure

Gravitational waves

What are the physical processes responsible for an observed gravitational wave signal?

LIGO



Merging black holes

Model selection in cosmology and astrophysics

These are questions of **model selection**.

Model selection in cosmology and astrophysics

These are questions of **model selection**.

In cosmology we **cannot perform experiments** but just have **one Universe** to observe.

Model selection in cosmology and astrophysics

These are questions of **model selection**.

In cosmology we **cannot perform experiments** but just have **one Universe** to observe.

In astrophysics, we again cannot perform experiments, but may have a small number of observations of similar processes.

Model selection in cosmology and astrophysics

These are questions of **model selection**.

In cosmology we **cannot perform experiments** but just have **one Universe** to observe.

In astrophysics, we again cannot perform experiments, but may have a small number of observations of similar processes.

$\Rightarrow$ Bayesian model selection

Outline

1. Bayesian model selection
2. Learnt harmonic mean estimator for likelihood-based model selection
3. Learnt harmonic mean estimator for simulation-based model selection
4. Proximal nested sampling for high-dimensional model selection

Bayesian model selection

Bayesian inference: parameter estimation

Bayes' theorem

$$\underbrace{\boxed{p(\theta \mid y, M)}}_{\text{posterior}} = \frac{\overbrace{\boxed{p(y \mid \theta, M)}}^{\text{likelihood}} \; \overbrace{\boxed{p(\theta \mid M)}}^{\text{prior}}}{\underbrace{\boxed{p(y \mid M)}}_{\text{evidence}}}$$

for parameters θ, model M and observed data y.

Bayesian inference: parameter estimation

Bayes' theorem

$$\underbrace{\boxed{p(\theta \,|\, y, M)}}_{\text{posterior}} = \frac{\overbrace{\boxed{p(y \,|\, \theta, M)}}^{\text{likelihood}} \overbrace{\boxed{p(\theta \,|\, M)}}^{\text{prior}}}{\underbrace{\boxed{p(y \,|\, M)}}_{\text{evidence}}} = \frac{\overbrace{\boxed{\mathcal{L}(\theta)}}^{\text{likelihood}} \overbrace{\boxed{\pi(\theta)}}^{\text{prior}}}{\underbrace{\boxed{z}}_{\text{evidence}}},$$

for parameters θ, model M and observed data y.

Bayesian inference: parameter estimation

Bayes' theorem

$$\underbrace{\boxed{p(\theta \,|\, y, M)}}_{\text{posterior}} = \frac{\overbrace{\boxed{p(y \,|\, \theta, M)}}^{\text{likelihood}} \overbrace{\boxed{p(\theta \,|\, M)}}^{\text{prior}}}{\underbrace{\boxed{p(y \,|\, M)}}_{\text{evidence}}} = \frac{\overbrace{\boxed{\mathcal{L}(\theta)}}^{\text{likelihood}} \overbrace{\boxed{\pi(\theta)}}^{\text{prior}}}{\underbrace{\boxed{z}}_{\text{evidence}}},$$

for parameters θ, model M and observed data y.

For **parameter estimation**, typically draw samples from the posterior by *Markov chain Monte Carlo (MCMC)* sampling.

Bayesian inference: model selection

By Bayes' theorem for model M_j:

$$p(M_j \,|\, y) = \frac{p(y \,|\, M_j) p(M_j)}{\sum_j p(y \,|\, M_j) p(M_j)} \,.$$

Bayesian inference: model selection

By Bayes' theorem for model M_j:

$$p(M_j \mid y) = \frac{p(y \mid M_j) p(M_j)}{\sum_j p(y \mid M_j) p(M_j)} .$$

For **model selection**, consider posterior model odds:

$$\underbrace{\boxed{\frac{p(M_1 \mid y)}{p(M_2 \mid y)}}}_{\text{posterior odds}} = \underbrace{\boxed{\frac{p(y \mid M_1)}{p(y \mid M_2)}}}_{\text{Bayes factor}} \times \underbrace{\boxed{\frac{p(M_1)}{p(M_2)}}}_{\text{prior odds}} .$$

Bayesian inference: model selection

By Bayes' theorem for model M_j:

$$p(M_j \mid y) = \frac{p(y \mid M_j)p(M_j)}{\sum_j p(y \mid M_j)p(M_j)} .$$

For **model selection**, consider posterior model odds:

$$\underbrace{\boxed{\frac{p(M_1 \mid y)}{p(M_2 \mid y)}}}_{\text{posterior odds}} = \underbrace{\boxed{\frac{p(y \mid M_1)}{p(y \mid M_2)}}}_{\text{Bayes factor}} \times \underbrace{\boxed{\frac{p(M_1)}{p(M_2)}}}_{\text{prior odds}} .$$

Must compute the **Bayesian model evidence** or **marginal likelihood** given by the normalising constant

$$\boxed{z = p(y \mid M) = \int \mathrm{d}\theta \, \mathcal{L}(\theta) \, \pi(\theta)} .$$

Bayesian inference: model selection

By Bayes' theorem for model M_j:

$$p(M_j \mid y) = \frac{p(y \mid M_j) p(M_j)}{\sum_j p(y \mid M_j) p(M_j)} .$$

For **model selection**, consider posterior model odds:

$$\underbrace{\boxed{\frac{p(M_1 \mid y)}{p(M_2 \mid y)}}}_{\text{posterior odds}} = \underbrace{\boxed{\frac{p(y \mid M_1)}{p(y \mid M_2)}}}_{\text{Bayes factor}} \times \underbrace{\boxed{\frac{p(M_1)}{p(M_2)}}}_{\text{prior odds}} .$$

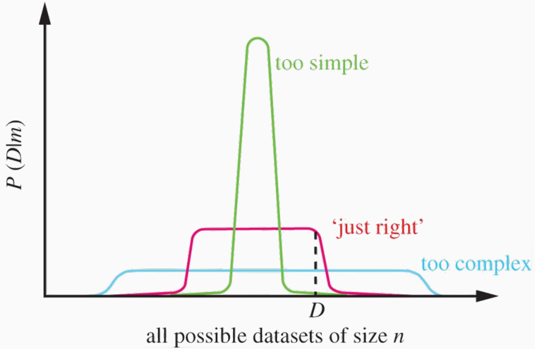
Must compute the **Bayesian model evidence** or **marginal likelihood** given by the normalising constant

$$\boxed{z = p(y \mid M) = \int \mathrm{d}\theta \, \mathcal{L}(\theta) \, \pi(\theta)} .$$

$\rightarrow$ Extremely challenging computational problem in high-dimensions.

Occam's razor

The Bayesian model evidence **naturally incorporates Occam's razor**, trading off model complexity and goodness of fit.



Ghahramani (2013); MacKay (1991)

On priors

- Physics-informed priors

  e.g. mass constrained to be positive

- Uninformative prior

  e.g. invariance to symmetry transformations

- Informative prior

  e.g. regularize by imposing sparsity in dictionary

- Data-informed priors

  e.g. prior $\sim$ old data, likelihood $\sim$ new data, posterior $\sim$ old and new data

- Data-driven priors

  e.g. empirical Bayes (estimate prior from data), learn by machine learning (generative models)

Challenge of Bayesian model selection

Naive Monte Carlo integration can be used to compute the marginal likelihood in principle.

However, the resulting estimator has very large variance, rendering it **ineffective in practice** (even in relatively low dimensional settings).

Challenge of Bayesian model selection

Naive Monte Carlo integration can be used to compute the marginal likelihood in principle.

However, the resulting estimator has very large variance, rendering it **ineffective in practice** (even in relatively low dimensional settings).

Require techniques **tailored** to the computation of the marginal likelihood.

Challenge of Bayesian model selection

Naive Monte Carlo integration can be used to compute the marginal likelihood in principle.

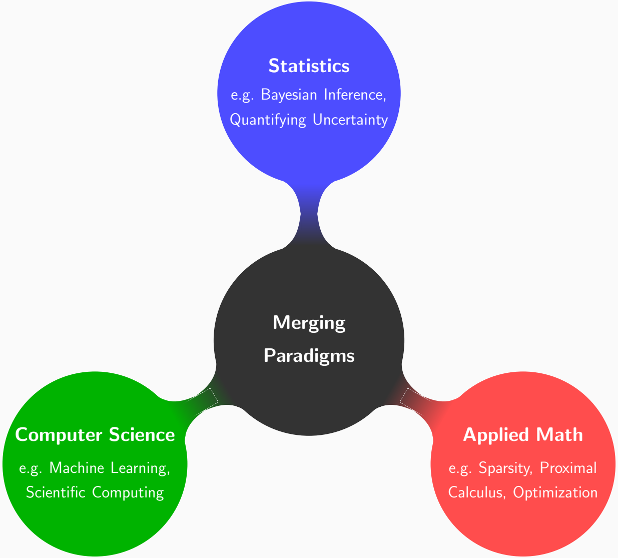
However, the resulting estimator has very large variance, rendering it **ineffective in practice** (even in relatively low dimensional settings).

Require techniques **tailored** to the computation of the marginal likelihood.

Challenges:

- Extending to general sampling strategies.
- Extending to simulation-based inference (likelihood-free inference).
- Scaling to high-dimensions.

Merging paradigms



Learnt harmonic mean estimator for likelihood-based model selection

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

$$\rho = \mathbb{E}_{p(\theta \mid y)} \left[\frac{1}{\mathcal{L}(\theta)} \right]$$

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

$$\rho = \mathbb{E}_{p(\theta \,|\, y)} \left[\frac{1}{\mathcal{L}(\theta)} \right] = \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} p(\theta \,|\, y)$$

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

$$\rho = \mathbb{E}_{p(\theta \,|\, y)}\left[\frac{1}{\mathcal{L}(\theta)}\right] = \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} p(\theta \,|\, y)$$

$$= \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} \frac{\mathcal{L}(\theta)\pi(\theta)}{z}$$

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

$$\begin{aligned}
\rho = \mathbb{E}_{p(\theta \,|\, y)}\left[\frac{1}{\mathcal{L}(\theta)}\right] &= \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} p(\theta \,|\, y) \\
&= \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} \frac{\mathcal{L}(\theta)\pi(\theta)}{z} \\
&= \frac{1}{z}
\end{aligned}$$

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

$$\rho = \mathbb{E}_{p(\theta \,|\, y)}\left[\frac{1}{\mathcal{L}(\theta)}\right] = \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} p(\theta \,|\, y) = \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} \frac{\mathcal{L}(\theta)\pi(\theta)}{z} = \frac{1}{z}$$

Original harmonic mean estimator (Newton & Raftery 1994)

$$\hat{\rho} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\mathcal{L}(\theta_i)} , \quad \theta_i \sim p(\theta \,|\, y)$$

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

$$\rho = \mathbb{E}_{p(\theta \,|\, y)} \left[\frac{1}{\mathcal{L}(\theta)} \right] = \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} p(\theta \,|\, y)$$
$$= \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} \frac{\mathcal{L}(\theta)\pi(\theta)}{z}$$
$$= \frac{1}{z}$$

Original harmonic mean estimator (Newton & Raftery 1994)

$$\hat{\rho} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\mathcal{L}(\theta_i)}, \quad \theta_i \sim p(\theta \,|\, y)$$

Very simple approach but **can fail catastrophically** (Neal 1994).

Importance sampling interpretation of harmonic mean estimator

Alternative interpretation of harmonic mean relationship:

$$\rho = \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} p(\theta \,|\, y) = \frac{1}{z} \int \mathrm{d}\theta \frac{\pi(\theta)}{p(\theta \,|\, y)} p(\theta \,|\, y) \, .$$

Importance sampling interpretation of harmonic mean estimator

Alternative interpretation of harmonic mean relationship:

$$\rho = \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} p(\theta \,|\, y) = \frac{1}{z} \underbrace{\boxed{\int \mathrm{d}\theta \frac{\pi(\theta)}{p(\theta \,|\, y)} p(\theta \,|\, y)}}^{\text{importance sampling}} .$$

Importance sampling interpretation of harmonic mean estimator

Alternative interpretation of harmonic mean relationship:

$$\rho = \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} p(\theta \,|\, y) = \frac{1}{z} \underbrace{\boxed{\int \mathrm{d}\theta \frac{\pi(\theta)}{p(\theta \,|\, y)} p(\theta \,|\, y)}}^{\text{importance sampling}} .$$

Importance sampling interpretation:

- Importance **sampling target distribution is prior** $\pi(\theta)$.
- Importance **sampling density is posterior** $p(\theta \,|\, y)$.

Importance sampling interpretation of harmonic mean estimator

Alternative interpretation of harmonic mean relationship:

$$\rho = \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} p(\theta \,|\, y) = \frac{1}{z} \underbrace{\boxed{\int \mathrm{d}\theta \frac{\pi(\theta)}{p(\theta \,|\, y)} p(\theta \,|\, y)}}^{\text{importance sampling}} .$$

Importance sampling interpretation:

- Importance **sampling target distribution is prior** $\pi(\theta)$.
- Importance **sampling density is posterior** $p(\theta \,|\, y)$.

For importance sampling, want sampling density to have fatter tails than target.

Importance sampling interpretation of harmonic mean estimator

Alternative interpretation of harmonic mean relationship:

$$\rho = \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} p(\theta \,|\, y) = \frac{1}{z} \underbrace{\boxed{\int \mathrm{d}\theta \frac{\pi(\theta)}{p(\theta \,|\, y)} p(\theta \,|\, y)}}^{\text{importance sampling}} .$$

Importance sampling interpretation:

- Importance **sampling target distribution is prior** $\pi(\theta)$.
- Importance **sampling density is posterior** $p(\theta \,|\, y)$.

For importance sampling, want sampling density to have fatter tails than target.

Not the case when importance sampling density is posterior and target is the prior.

Re-targeted harmonic mean estimator

Introduce an arbitrary importance sampling target $\varphi(\theta)$ (which must be normalised).

Re-targeted harmonic mean estimator

Introduce an arbitrary importance sampling target $\varphi(\theta)$ (which must be normalised).

Re-targeted **harmonic mean relationship** (Gelfand & Dey 1994)

$$\rho = \mathbb{E}_{p(\theta \mid y)}\left[\frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)}\right]$$

Re-targeted harmonic mean estimator

Introduce an arbitrary importance sampling target $\varphi(\theta)$ (which must be normalised).

Re-targeted **harmonic mean relationship** (Gelfand & Dey 1994)

$$\rho = \mathbb{E}_{p(\theta \mid y)}\left[\frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)}\right] = \int \mathrm{d}\theta \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} p(\theta \mid y)$$

Re-targeted harmonic mean estimator

Introduce an arbitrary importance sampling target $\varphi(\theta)$ (which must be normalised).

Re-targeted **harmonic mean relationship** (Gelfand & Dey 1994)

$$\rho = \mathbb{E}_{p(\theta \,|\, y)}\left[\frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)}\right] = \int \mathrm{d}\theta \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} p(\theta \,|\, y)$$

$$= \int \mathrm{d}\theta \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \frac{\mathcal{L}(\theta)\pi(\theta)}{z}$$

Re-targeted harmonic mean estimator

Introduce an arbitrary importance sampling target $\varphi(\theta)$ (which must be normalised).

Re-targeted harmonic mean relationship (Gelfand & Dey 1994)

$$\rho = \mathbb{E}_{p(\theta \,|\, y)}\left[\frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)}\right] = \int \mathrm{d}\theta \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} p(\theta \,|\, y)$$

$$= \int \mathrm{d}\theta \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \frac{\mathcal{L}(\theta)\pi(\theta)}{z}$$

$$= \frac{1}{z}$$

Re-targeted harmonic mean estimator

Introduce an arbitrary importance sampling target $\varphi(\theta)$ (which must be normalised).

Re-targeted **harmonic mean relationship** (Gelfand & Dey 1994)

$$\begin{aligned}\rho = \mathbb{E}_{p(\theta \,|\, y)}\left[\frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)}\right] &= \int \mathrm{d}\theta \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} p(\theta \,|\, y) \\ &= \int \mathrm{d}\theta \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \frac{\mathcal{L}(\theta)\pi(\theta)}{z} \\ &= \frac{1}{z}\end{aligned}$$

Re-targeted **harmonic mean estimator** (Gelfand & Dey 1994)

$$\hat{\rho} = \frac{1}{N} \sum_{i=1}^{N} \frac{\varphi(\theta_i)}{\mathcal{L}(\theta_i)\pi(\theta_i)}, \quad \theta_i \sim p(\theta \,|\, y)$$

Re-targeted harmonic mean estimator

Importance sampling interpretation:

$$\rho = \int \mathrm{d}\theta \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} p(\theta \,|\, y) = \frac{1}{z} \int \mathrm{d}\theta \frac{\varphi(\theta)}{p(\theta \,|\, y)} p(\theta \,|\, y) \,.$$

Re-targeted harmonic mean estimator

Importance sampling interpretation:

$$\rho = \int \mathrm{d}\theta \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} p(\theta \,|\, y) = \frac{1}{z} \int \mathrm{d}\theta \frac{\varphi(\theta)}{p(\theta \,|\, y)} p(\theta \,|\, y) \,.$$

Ensure importance sampling target $\varphi(\theta)$ does **not** have fatter tails than posterior $p(\theta \,|\, y)$ (importance sampling density).

Re-targeted harmonic mean estimator

Importance sampling interpretation:

$$\rho = \int \mathrm{d}\theta \, \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} p(\theta \,|\, y) = \frac{1}{z} \int \mathrm{d}\theta \, \frac{\varphi(\theta)}{p(\theta \,|\, y)} p(\theta \,|\, y) \, .$$

Ensure importance sampling target $\varphi(\theta)$ does **not** have fatter tails than posterior $p(\theta \,|\, y)$ (importance sampling density).

$\rightarrow$ How set importance sampling target distribution $\varphi(\theta)$?

How set importance sampling target distribution $\varphi(\theta)$?

Variety of cases been considered:

- Multi-variate Gaussian (*e.g.* Chib 1995)
- Indicator functions (*e.g.* Robert & Wraith 2009, van Haasteren 2009)

How set importance sampling target distribution $\varphi(\theta)$?

Variety of cases been considered:

- Multi-variate Gaussian (*e.g.* Chib 1995)
- Indicator functions (*e.g.* Robert & Wraith 2009, van Haasteren 2009)

Optimal target:

$$\boxed{\varphi^{\text{optimal}}(\theta) = \frac{\mathcal{L}(\theta)\pi(\theta)}{z}}$$

(resulting estimator has zero variance).

How set importance sampling target distribution $\varphi(\theta)$?

Variety of cases been considered:

- Multi-variate Gaussian (*e.g.* Chib 1995)
- Indicator functions (*e.g.* Robert & Wraith 2009, van Haasteren 2009)

Optimal target:

$$\boxed{\varphi^{\text{optimal}}(\theta) = \frac{\mathcal{L}(\theta)\pi(\theta)}{z}}$$

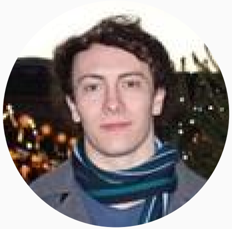
(resulting estimator has zero variance).

But clearly **not feasible** since requires knowledge of the evidence z (recall the target must be normalised) $\rightarrow$ requires problem to have been solved already!

Learnt harmonic mean estimator

Propose the ***learnt* harmonic mean estimator**
(McEwen, Wallis, Price, Docherty 2021; arXiv:2111.12720).

Chris Wallis



Matt Price

Matthew Docherty

Learn an approximation of the optimal target distribution:

$$\boxed{\varphi(\theta) \overset{\text{ML}}{\simeq} \varphi^{\text{optimal}}(\theta) = \frac{\mathcal{L}(\theta)\pi(\theta)}{z}}\,.$$

Learnt harmonic mean estimator

Learn an approximation of the optimal target distribution:

$$\boxed{\varphi(\theta) \overset{\text{ML}}{\simeq} \varphi^{\text{optimal}}(\theta) = \frac{\mathcal{L}(\theta)\pi(\theta)}{z}}\ .$$

- Approximation not required to be highly accurate.
- Must not have fatter tails than posterior.

Learnt harmonic mean estimator

Learn an approximation of the optimal target distribution:

$$\boxed{\varphi(\theta) \overset{\text{ML}}{\simeq} \varphi^{\text{optimal}}(\theta) = \frac{\mathcal{L}(\theta)\pi(\theta)}{z}}\ .$$

- Approximation not required to be highly accurate.
- Must not have fatter tails than posterior.

Also develop strategy to estimate the variance of the estimator, its variance, and other sanity checks.

Learning the target distribution

Consider a **variety of machine learning approaches**:

- Uniform hyper-ellipsoid
- Kernel Density Estimation (KDE)
- Modified Gaussian mixture model (MGMM)
- (Normalising flows coming...)

Fit model by **minimising variance of resulting estimator**, while ensuring unbiased, with possible regularisation:

$$\min \ \hat{\sigma}^2 + \lambda R \quad \text{subject to} \quad \hat{\rho} = \hat{\mu}_1$$

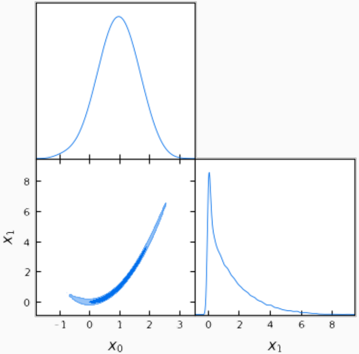
Solve by bespoke **mini-batch stochastic gradient descent**.

Cross-validation to select machine learning model and hyperparameters.

Rosenbrock example

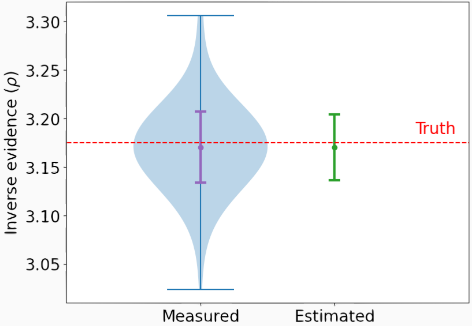
Rosenbrock function is the classical example of a **pronounced thin curving degeneracy**, with likelihood defined by

$$f(\theta) = \sum_{i=1}^{n-1} \left[(a - \theta_i)^2 + b(\theta_{i+1} - \theta_i^2)^2 \right] , \qquad \log(\mathcal{L}(\theta)) = -f(\theta) .$$



Posterior recovered by MCMC sampling.

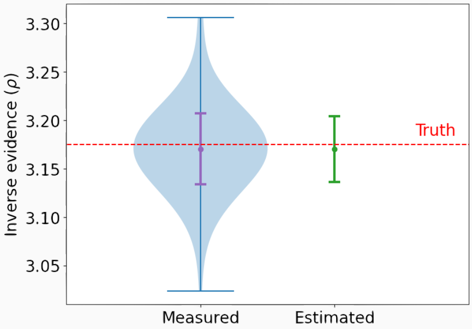
Rosenbrock example



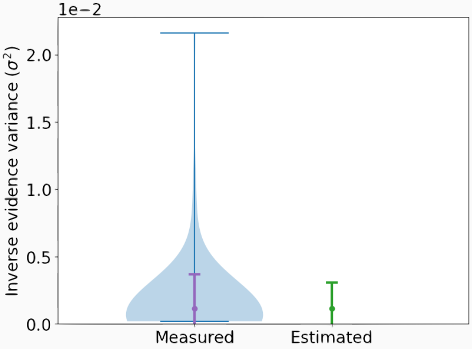
Reciprocal evidence

Accuracy of learnt harmonic mean estimator for Rosenbrock example.

Rosenbrock example



Reciprocal evidence



Variance of reciprocal evidence

Accuracy of learnt harmonic mean estimator for Rosenbrock example.

Normal-Gamma example

Pathological example (Friel & Wyse 2012) where original harmonic mean estimator fails.

Normal-Gamma example

Pathological example (Friel & Wyse 2012) where original harmonic mean estimator fails.

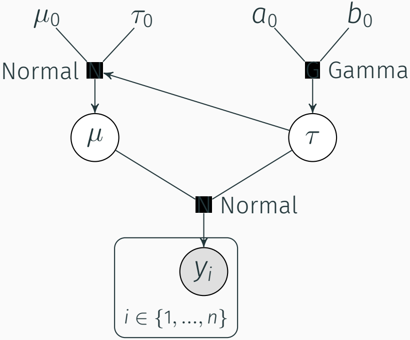
Data model:

$$y_i \sim \mathrm{N}(\mu, \tau^{-1})$$

Prior model:

$$\text{Mean: } \mu \sim \mathrm{N}\big(\mu_0, (\tau_0 \tau)^{-1}\big)$$

$$\text{Precision: } \tau \sim \mathrm{Ga}(a_0, b_0)$$



Hierarchical Bayesian model of Normal-Gamma example.

Normal-Gamma example

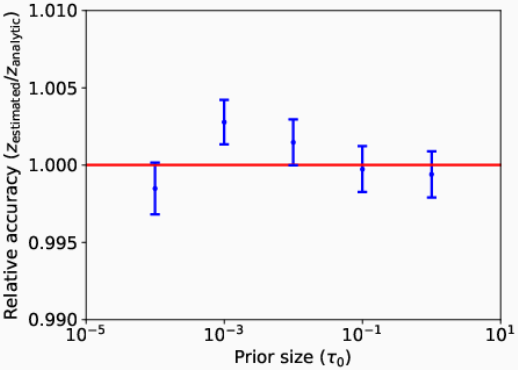
Analytic evidence:

$$\boxed{z = (2\pi)^{-n/2} \frac{\Gamma(a_n)}{\Gamma(a_0)} \frac{b_0^{a_0}}{b_n^{a_n}} \left(\frac{\tau_0}{\tau_n}\right)^{1/2}}$$

where

$$\tau_n = \tau_0 + n \, , \quad a_n = a_0 + n/2 \, , \quad b_n = b_0 + \frac{1}{2}\sum_{i=1}^{n}(y_i - \bar{y})^2 + \frac{\tau_0 n(\bar{y} - \mu_0)^2}{2(\tau_0 + n)} \, .$$

Normal-Gamma example



Comparison of marginal likelihood values computed to truth for varying prior.

Normal-Gamma example

Marginal likelihood values for Normal-Gamma example with varying prior.

τ_0	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}
Analytic $\log(z)$	-144.5530	-143.4017	-142.2505	-141.0999	-139.9552
Estimated $\log(\hat{z})$	-144.5545	-143.3990	-142.2490	-141.1001	-139.9558
Error (learnt harmonic mean)	-0.0015	0.0027	0.0015	-0.0011	-0.0006
Error (original harmonic mean)	12.2100	—	9.7900	8.5000	7.1000

Radiata pine example

Radiata pine data-set has become **classical benchmark** for evaluating evidence estimators:

- maximum compression strength parallel to grain y_i,
- density x_i,
- density adjust for resin content z_i,

for $i \in \{1, \dots, n\}$ where $n = 42$ specimens.

Is **density** or **resin-adjusted density** a better predictor of compression strength?

Radiata pine example

Gaussian linear models:

$$M_1: \quad y_i = \alpha + \underbrace{\boxed{\beta(x_i - \bar{x})}}_{\text{density}} + \epsilon_i \, , \quad \epsilon_i \sim \mathrm{N}(0, \tau^{-1}) \, .$$

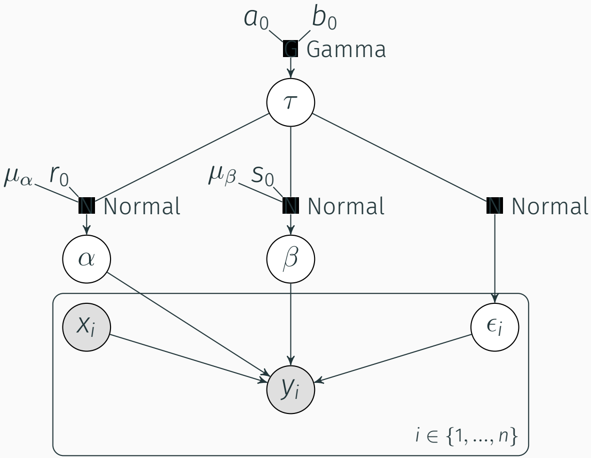
$$M_2: \quad y_i = \gamma + \underbrace{\boxed{\delta(z_i - \bar{z})}}_{\text{resin-adjusted density}} + \eta_i \, , \quad \eta_i \sim \mathrm{N}(0, \lambda^{-1}) \, .$$

Priors for model 1 (similar for model 2):

$$\alpha \sim \mathrm{N}\big(\mu_\alpha, (r_0\tau)^{-1}\big) \, ,$$
$$\beta \sim \mathrm{N}\big(\mu_\beta, (s_0\tau)^{-1}\big) \, ,$$
$$\tau \sim \mathrm{Ga}(a_0, b_0) \, ,$$

($\mu_\alpha = 3000$, $\mu_\beta = 185$, $r_0 = 0.06$, $s_0 = 6$, $a_0 = 3$, $b_0 = 2 \times 300^2$).

Radiata pine example



Hierarchical Bayesian model for Radiata pine example (for model 1; model 2 is similar).

Analytic evidence:

$$\boxed{z = \pi^{-n/2} b_0^{a_0} \frac{\Gamma(a_0 + n/2)}{\Gamma(a_0)} \frac{|Q_0|^{1/2}}{|M|^{1/2}} \left(\boldsymbol{y}^\mathsf{T}\boldsymbol{y} + \boldsymbol{\mu_0}^\mathsf{T} Q_0 \boldsymbol{\mu_0} - \boldsymbol{\nu_0}^\mathsf{T} M \boldsymbol{\nu_0} + 2b_0\right)^{-a_0-n/2}}$$

where $\boldsymbol{\mu_0} = (\mu_\alpha, \mu_\beta)^\mathsf{T}$, $Q_0 = \mathrm{diag}(r_0, s_0)$, and $M = X^\mathsf{T} X + Q_0$.

Radiata pine example

Marginal likelihood values for Radiata Pine example.

	Model M_1 $\log(z_1)$	Model M_2 $\log(z_2)$	$\log \mathrm{BF}_{21}$ $= \log(z_2) - \log(z_1)$
Analytic	-310.12829	-301.70460	8.42368
Estimated	-310.12807	-301.70413	8.42394
	$\pm$ 0.00072	$\pm$ 0.00074	$\pm$ 0.00145
Error (learnt harmonic mean)	0.00022	0.00047	0.00026
Error (original harmonic mean)	–	–	-0.17372

GitHub harmonic | Tests passing | docs passing | codecov 96% | pypi package 1.0.3 | License GPL | arXiv 2111.12720

Harmonic

`harmonic` is an open source, well tested and documented Python implementation of the *learnt harmonic mean estimator* (McEwen et al. 2021) to compute the marginal likelihood (Bayesian evidence), required for Bayesian model selection.

Github: `https://github.com/astro-informatics/harmonic`

Docs: `https://astro-informatics.github.io/harmonic`

(Seamless integration with emcee.)

Learnt harmonic mean estimator for simulation-based model selection

Simulation-based inference (SBI)

Consider situation where the likelihood $p(y \,|\, \theta, M)$ is unknown or intractable.

Simulation-based inference (likelihood-free inference) seeks to perform parameter inference by estimating the posterior $p(\theta \,|\, y_o, M)$ for observed data y_o using **simulations only**.

Simulation-based inference (SBI)

Consider situation where the likelihood $p(y \mid \theta, M)$ is unknown or intractable.

Simulation-based inference (likelihood-free inference) seeks to perform parameter inference by estimating the posterior $p(\theta \mid y_o, M)$ for observed data y_o using **simulations only**.

Advantages:

- Forward modelling of complex physics, contamination, observational process.
- No assumptions on the form of the likelihood.

Flavours of neural density estimation

1. **Neural posterior estimation** (NPE)
   (Papamakarios & Murray 2016; Lueckmann *et al.* 2017; Greenberg *et al.* 2019)
2. **Neural likelihood estimation** (NLE)
   (Papamakarios *et al.* 2019)
3. **Neural ratio estimation** (NRE)
   (Hermans *et al.* 2019; Durkan *et al.* 2020)

Neural posterior estimation (NPE)

Construct training data $\{(\theta_i, y_i)\}$ where parameter drawn from proposal prior $\theta_i \sim \tilde{p}(\theta \,|\, M)$ and then generate simulation $y_i \sim p(y \,|\, \theta_i)$.

Learn posterior

$$\boxed{q_\phi^{\text{NPE}}(\theta \,|\, y, M) \simeq p(\theta \,|\, y, M)} \, ,$$

where ϕ are the parameters of the learned model.

(Papamakarios & Murray 2016; Lueckmann *et al.* 2017; Greenberg *et al.* 2019)

Neural likelihood estimation (NLE)

Learn likelihood

$$\boxed{q_\phi^{\text{NLE}}(y \,|\, \theta, M) \simeq p(y \,|\, \theta, M)} \,,$$

where ϕ are the parameters of the learned model.

(Papamakarios *et al.* 2019)

Neural ratio estimation (NRE)

Learn density ratio proportional to the likelihood

$$\boxed{r_{\boldsymbol{\phi}}(y, \boldsymbol{\theta}) = \frac{p(y, \boldsymbol{\theta})}{p(y)p(\boldsymbol{\theta})} = \frac{p(y|\boldsymbol{\theta})}{p(y)} = \frac{p(\boldsymbol{\theta}|y)}{p(\boldsymbol{\theta})}},$$

where ϕ are the parameters of the learned model.

(Hermans *et al.* 2019; Durkan *et al.* 2020)

Amortised vs sequential

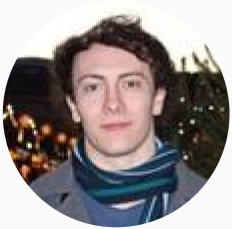
Amortized approach: Amortise training of the density estimator, allowing offline inference to be run on multiple different observations.

Sequential approach: Focus on specific observation and train in *runs* where the proposal prior matches the intermediate learned posterior.

Bayesian model comparison for simulation-based inference
(Spurio Mancini, Docherty, Price, McEwen 2022; arXiv:2207.04037).

Alessio Spurio Mancini

Matthew Docherty



Matt Price

Naive model evidence computation

Recall NPE and NLE:

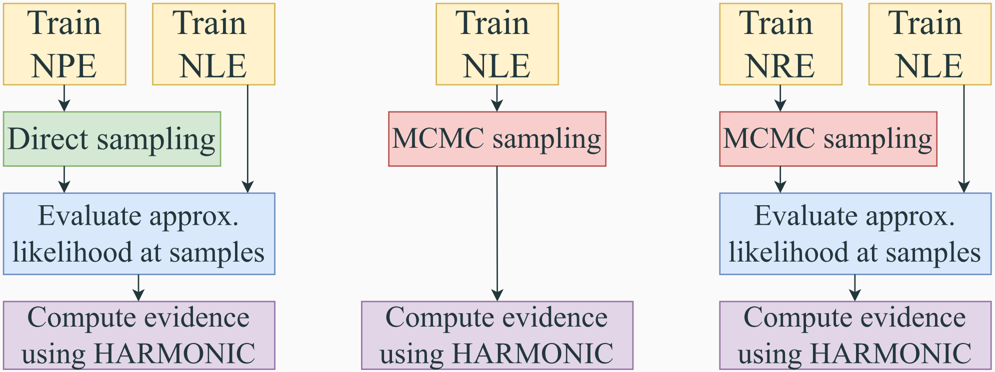
$$q_\phi^{\mathrm{NPE}}(\theta \,|\, y, M) \simeq p(\theta \,|\, y, M); \qquad q_\psi^{\mathrm{NLE}}(y \,|\, \theta, M) \simeq p(y \,|\, \theta, M)\,.$$

Naive estimate of the model evidence:

$$\boxed{\hat{z} = \frac{1}{N} \sum_i \frac{q_\psi^{\mathrm{NLE}}(y_o \,|\, \theta_i, M)\, p(\theta_i \,|\, M)}{q_\phi^{\mathrm{NPE}}(\theta_i \,|\, y_o, M)}}\,.$$

Ratio of two approximate quantities, hence approximation errors compound.

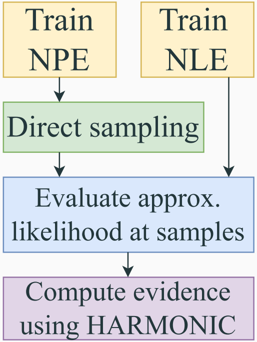
SBI evidence computation methodologies



Neural posterior estimation (NPE)

Neural likelihood estimation (NLE)

Neural ratio estimation (NRE)

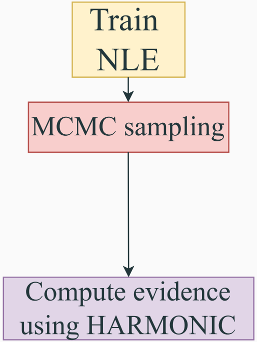


Neural posterior estimation (NPE)

Learnt harmonic mean estimator for NPE:

$$\boxed{\hat{\rho} = \frac{1}{N}\sum_{i=1}^{N} \frac{\varphi(\boldsymbol{\theta}_i)}{q_{\boldsymbol{\phi}}^{\text{NLE}}(y|\boldsymbol{\theta}_i)\,p(\boldsymbol{\theta}_i)}, \quad \boldsymbol{\theta}_i \overset{\text{direct}}{\sim} q_{\boldsymbol{\psi}}^{\text{NPE}}(\boldsymbol{\theta}|y)}\ .$$

- Samples can be generated directly using surrogate posterior (avoiding MCMC sampling) $\rightarrow$ highly efficient, computed in parallel.
- Only possible since learnt harmonic mean estimator agnostic to sampling strategy.
- Avoids compounding approximation errors.
- In a likelihood-based setting, can also be applied to accelerate evidence computation.

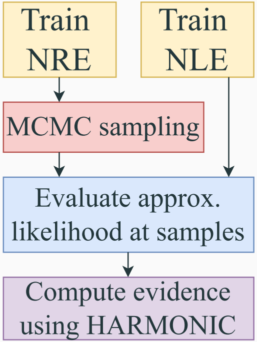


Neural likelihood estimation (NLE)

Learnt harmonic mean estimator for NLE:

$$\boxed{\hat{\rho} = \frac{1}{N}\sum_{i=1}^{N} \frac{\varphi(\boldsymbol{\theta}_i)}{q_{\boldsymbol{\phi}}^{\mathrm{NLE}}(y|\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)}, \quad \boldsymbol{\theta}_i \overset{\mathrm{MCMC}}{\sim} q_{\boldsymbol{\phi}}^{\mathrm{NLE}}(y|\boldsymbol{\theta})p(\boldsymbol{\theta})}\ .$$

- Samples generated by MCMC sampling.
- Avoids compounding approximation errors.
- Only need to train one model.



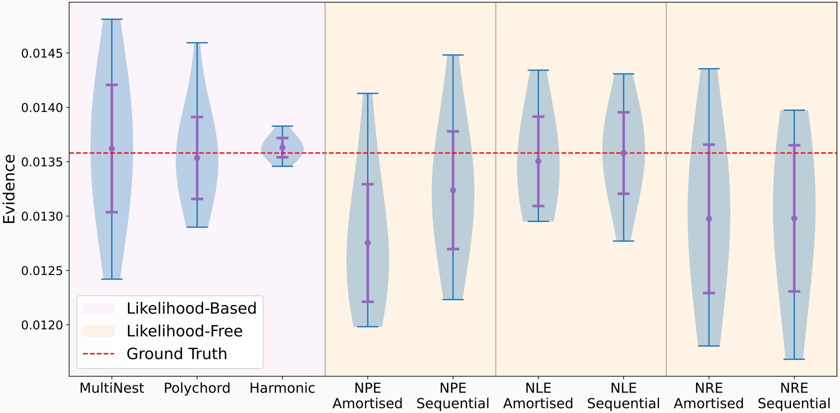
Neural ratio estimation (NRE)

Learnt harmonic mean estimator for NRE:

$$\boxed{\hat{\rho} = \frac{1}{N}\sum_{i=1}^{N} \frac{\varphi(\boldsymbol{\theta}_i)}{q_{\boldsymbol{\phi}}^{\text{NLE}}(y|\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)}, \quad \boldsymbol{\theta}_i \overset{\text{MCMC}}{\sim} r_{\boldsymbol{\psi}}^{\text{NRE}}(y,\boldsymbol{\theta})p(\boldsymbol{\theta})} \; .$$

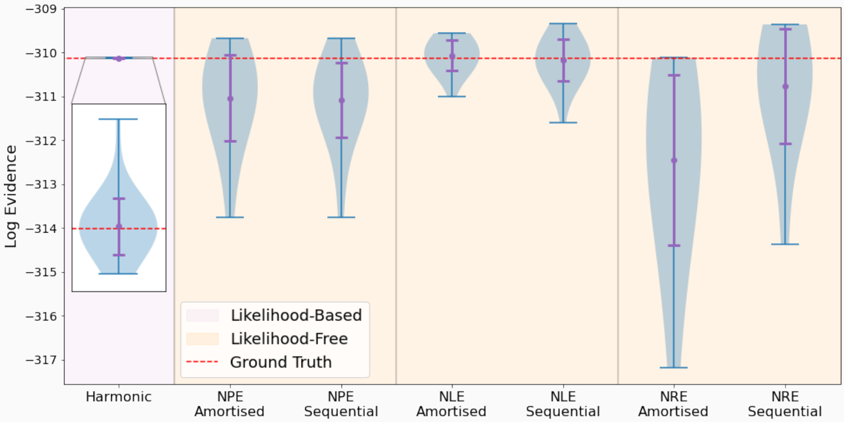
- Samples generated by MCMC sampling.
- Avoids compounding approximation errors.

Linear Gaussian example



Model evidence computed in likelihood-based and simulation-based settings.

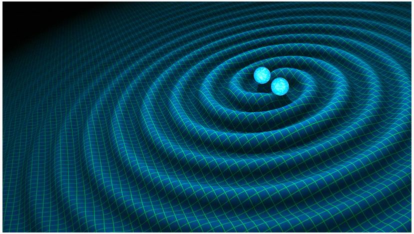
Radiata pine example



Model evidence computed in likelihood-based and simulation-based settings.

Gravitational wave example

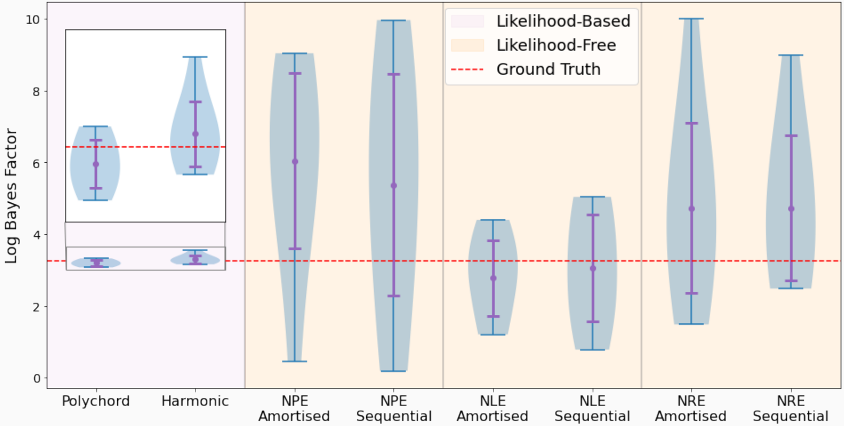
Simulate a **black-hole, black-hole merger** as observed by an interferometer (*e.g.* LIGO).



Perform model comparison for two models:

1. *Spin-Precessing Effective-One-Body Numerical Relativity*
2. *Inspiral Ringdown Merger*

Likelihood available for validation.



Model evidence computed in likelihood-based and simulation-based settings.

Proximal nested sampling for high-dimensional model selection

Nested sampling: reparameterising the likelihood

Nested sampling is a clever approach to efficiently evalute the evidence (Skilling 2006).

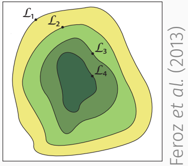
Consider $\Omega_{L^*} = \{x | \mathcal{L}(x) \geq L^*\}$, which groups the parameter space Ω into a series of **nested subspaces**.

Define the prior volume ξ within Ω_{L^*} by $\xi(L^*) = \int_{\Omega_{L^*}} \pi(x) \mathrm{d}x$.

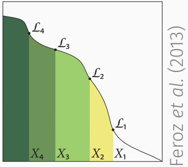
The marginal likelihood integral can then be rewritten as

$$\mathcal{Z} = \int_0^1 \mathcal{L}(\xi) \mathrm{d}\xi,$$

which is a **one-dimensional integral** over the prior volume ξ.



Nested subspaces



Reparameterised likelihood

Nested sampling: constrained sampling

Require strategy to compute likelihood level-sets (iso-contours) L_i and corresponding prior volumes $0 < \xi_i \leq 1$.

Nested sampling (Skilling 2006)

Nested sampling: constrained sampling

Require strategy to compute likelihood level-sets (iso-contours) L_i and corresponding prior volumes $0 < \xi_i \leq 1$.

Nested sampling (Skilling 2006)

1. Draw N_{live} *live* samples from prior, with prior volume $\xi_0 = 1$.

Nested sampling: constrained sampling

Require strategy to compute likelihood level-sets (iso-contours) L_i and corresponding prior volumes $0 < \xi_i \leq 1$.

Nested sampling (Skilling 2006)

1. Draw N_{live} *live* samples from prior, with prior volume $\xi_0 = 1$.
2. Remove sample with smallest likelihood, say L_i.

Nested sampling: constrained sampling

Require strategy to compute likelihood level-sets (iso-contours) L_i and corresponding prior volumes $0 < \xi_i \leq 1$.

Nested sampling (Skilling 2006)

1. Draw N_{live} *live* samples from prior, with prior volume $\xi_0 = 1$.
2. Remove sample with smallest likelihood, say L_i.
3. Replace removed sample with new sample from the prior but constrained to a higher likelihood than L_i.

Nested sampling: constrained sampling

Require strategy to compute likelihood level-sets (iso-contours) L_i and corresponding prior volumes $0 < \xi_i \leq 1$.

Nested sampling (Skilling 2006)

1. Draw N_{live} *live* samples from prior, with prior volume $\xi_0 = 1$.
2. Remove sample with smallest likelihood, say L_i.
3. Replace removed sample with new sample from the prior but constrained to a higher likelihood than L_i.
4. Estimate (stochastically) prior volume ξ_i enclosed by likelihood level-set L_i.

Nested sampling: constrained sampling

Require strategy to compute likelihood level-sets (iso-contours) L_i and corresponding prior volumes $0 < \xi_i \leq 1$.

Nested sampling (Skilling 2006)

1. Draw N_{live} *live* samples from prior, with prior volume $\xi_0 = 1$.
2. Remove sample with smallest likelihood, say L_i.
3. Replace removed sample with new sample from the prior but constrained to a higher likelihood than L_i.
4. Estimate (stochastically) prior volume ξ_i enclosed by likelihood level-set L_i.
5. Repeat 2–5.

Nested sampling: estimating enclosed prior volume stochastically

Enclosed prior volume decreases exponentially at each step: $\xi_{i+1} = t_{i+1}\xi_i$.

Shrinkage ratio can be estimated stochastically since $\mathbb{E}(\log t) = -1/N_{\text{live}}$.

The enclosed prior volume can then be estimated by

$$\xi_{i+1} = \exp(-i/N_{\text{live}}) \, .$$

Nested sampling: evidence estimation and posterior inference

Given the sequence of decreasing prior volumes $\{\xi_i\}_{i=0}^{N}$ and corresponding likelihoods $L_i = \mathcal{L}(\xi_i)$, the **model evidence** can be computed numerically using standard quadrature:

$$\boxed{\mathcal{Z} = \sum_{i=1}^{N} L_i w_i},$$

for quadrature weight w_i (e.g. the trapezium rule with $w_i = (\xi_{i-1} + \xi_{i+1})/2$).

Nested sampling: evidence estimation and posterior inference

Given the sequence of decreasing prior volumes $\{\xi_i\}_{i=0}^{N}$ and corresponding likelihoods $L_i = \mathcal{L}(\xi_i)$, the **model evidence** can be computed numerically using standard quadrature:

$$\boxed{\mathcal{Z} = \sum_{i=1}^{N} L_i w_i} \, ,$$

for quadrature weight w_i (e.g. the trapezium rule with $w_i = (\xi_{i-1} + \xi_{i+1})/2$).

Posterior inferences can also be computed by assigning importances weights

$$\boxed{p_i = \frac{L_i w_i}{\mathcal{Z}}} \, .$$

Nested sampling: challenge

Recall: to compute the marginal likelihood by nested sampling require strategy to generate likelihoods L_i and associated prior volumes ξ_i.

Achieved by **sampling from the prior, subject the likelihood iso-contour constraint**, *i.e.* sampling from the prior $\pi(x)$, such that $\mathcal{L}(x) > L^*$.

Nested sampling: challenge

Recall: to compute the marginal likelihood by nested sampling require strategy to generate likelihoods L_i and associated prior volumes ξ_i.

Achieved by **sampling from the prior, subject the likelihood iso-contour constraint**, *i.e.* sampling from the prior $\pi(x)$, such that $\mathcal{L}(x) > L^*$.

This is the **main difficulty** in applying nested sampling to high-dimensional problems.

Exploit common structure

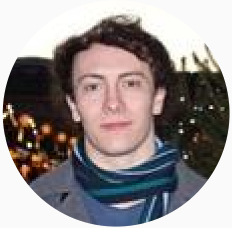
Many high-dimensional inverse problems are **log-convex**, *e.g.* inverse imaging problems with Gaussian data fidelity and sparsity-promoting prior.

Exploit structure (log convexity) of the problem.

⇒ Proximal nested sampling (Cai, McEwen & Pereyra 2022; arXiv:2106.03646)

Xiaohao Cai

Marcelo Pereyra



Matt Price

Constrained sampling formulation

Consider case where prior and likelihood of form

$$\boxed{\pi(x) = \exp(-f(x))}, \qquad \boxed{\mathcal{L}(x) = \exp(-g(x))},$$

prior likelihood

where f and g are convex lower semicontinuous functions on Ω.

Let $\iota_{L^*}(x)$ and $\chi_{L^*}(x)$ be the indicator and characteristic functions:

$$\iota_{L^*}(x) = \begin{cases} 1, & \mathcal{L}(x) > L^*, \\ 0, & \text{otherwise}, \end{cases} \quad \text{and} \quad \chi_{L^*}(x) = \begin{cases} 0, & \mathcal{L}(x) > L^*, \\ +\infty, & \text{otherwise}. \end{cases} \tag{1}$$

Then let $\boxed{\pi_{L^*}(x) = \pi(x)\iota_{L^*}(x)}$ represent the prior distribution with the hard likelihood constraint.

Constrained sampling formulation

Taking the logarithm, we can write

$$\boxed{-\log \pi_{L^*}(x) = f(x) + \chi_{\mathcal{B}_\tau}(x)},$$

where $\chi_{\mathcal{B}_\tau}(x)$ is the characteristic function associated with the convex set

$$\mathcal{B}_\tau := \{x | g(x) < \tau\},$$

for $\tau = -\log L^*$.

MCMC sampling with Langevin dynamics

Consider posteriors of the following form:

$$p(x \mid y) = \pi(x) \propto \exp\big(-p(x)\big).$$

If $p(x)$ differentiable can adopt Langevin dynamics.

Based on Langevin diffusion process $\mathcal{L}(t)$, with π as stationary distribution:

$$\mathrm{d}\mathcal{L}(t) = \frac{1}{2}\nabla \log \pi\big(\mathcal{L}(t)\big)\mathrm{d}t + \mathrm{d}\mathcal{W}(t), \quad \mathcal{L}(0) = l_0$$

where $\mathcal{W}$ is Brownian motion.

MCMC sampling with Langevin dynamics

Consider posteriors of the following form:

$$p(x \mid y) = \pi(x) \propto \exp\big(-p(x)\big).$$

If $p(x)$ differentiable can adopt Langevin dynamics.

Based on Langevin diffusion process $\mathcal{L}(t)$, with π as stationary distribution:

$$\mathrm{d}\mathcal{L}(t) = \frac{1}{2} \underbrace{\boxed{\nabla \log \pi\big(\mathcal{L}(t)\big)}}_{\text{gradient}} \mathrm{d}t + \mathrm{d}\mathcal{W}(t), \quad \mathcal{L}(0) = l_0$$

where $\mathcal{W}$ is Brownian motion.

Need gradients so **not directly applicable**.

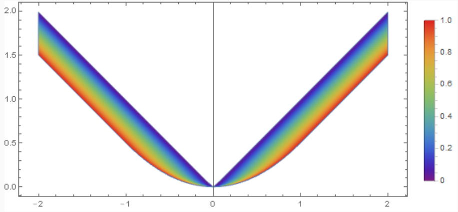
Moreau-Yosida approximation

Moreau-Yosida approximation (envelope) of f:

$$f^{\lambda}(x) = \inf_{u \in \mathbb{R}^N} f(u) + \frac{\|u - x\|^2}{2\lambda}$$

Important properties of $f^{\lambda}(x)$:

1. As $\lambda \rightarrow 0, f^{\lambda}(x) \rightarrow f(x)$
2. $\nabla f^{\lambda}(x) = (x - \mathrm{prox}_f^{\lambda}(x))/\lambda$



Moreau-Yosida envelope of $|x|$ for varying λ

[Credit: Stack exchange (ubpdqn)].

Proximal nested sampling

Proximal nested sampling (Cai, McEwen & Pereyra 2021; arXiv:2106.03646)

- Constrained sampling formulation
- Langevin MCMC sampling
- Moreau-Yosida approximation of constraint (and any non-differentiable prior)

Proximal nested sampling

Proximal nested sampling (Cai, McEwen & Pereyra 2021; arXiv:2106.03646)

- Constrained sampling formulation
- Langevin MCMC sampling
- Moreau-Yosida approximation of constraint (and any non-differentiable prior)

Proximal nested sampling Markov chain:

$$\boxed{x^{(k+1)} = x^{(k)} - \frac{\delta}{2}\nabla f(x^{(k)}) - \frac{\delta}{2\lambda}\left[x^{(k)} - \mathrm{prox}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right] + \sqrt{\delta}\, w^{(k+1)}}\,.$$

Proximal nested sampling intuition

Recall proximal nested sampling Markov chain:

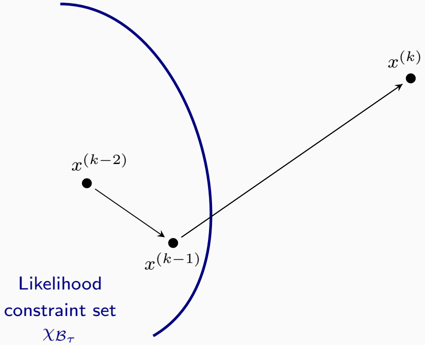
$$x^{(k+1)} = x^{(k)} - \frac{\delta}{2}\nabla f(x^{(k)}) - \frac{\delta}{2\lambda}\boxed{\left[x^{(k)} - \mathrm{prox}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right]} + \sqrt{\delta}w^{(k+1)}.$$

Proximal nested sampling intuition

Recall proximal nested sampling Markov chain:

$$x^{(k+1)} = x^{(k)} - \frac{\delta}{2}\nabla f(x^{(k)}) - \frac{\delta}{2\lambda}\boxed{\left[x^{(k)} - \mathrm{prox}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right]} + \sqrt{\delta}w^{(k+1)}.$$

1. $x^{(k)}$ **is already in** $\mathcal{B}_\tau$: term $\left[x^{(k)} - \mathrm{prox}^{\lambda}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right]$ disappears and recover usual Langevin MCMC.



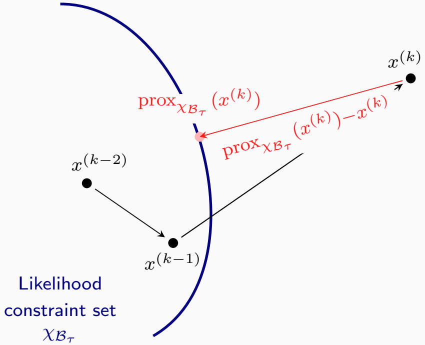
Proximal nested sampling intuition

Recall proximal nested sampling Markov chain:

$$x^{(k+1)} = x^{(k)} - \frac{\delta}{2}\nabla f(x^{(k)}) - \frac{\delta}{2\lambda}\boxed{\left[x^{(k)} - \mathrm{prox}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right]} + \sqrt{\delta}w^{(k+1)}.$$

1. $x^{(k)}$ **is already in** $\mathcal{B}_\tau$: term $\left[x^{(k)} - \mathrm{prox}^{\lambda}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right]$ disappears and recover usual Langevin MCMC.

2. $x^{(k)}$ **is not in** $\mathcal{B}_\tau$: a step is also taken in the direction $-\left[x^{(k)} - \mathrm{prox}^{\lambda}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right]$, which moves the next iteration in the direction of the projection of $x^{(k)}$ onto the convex set $\mathcal{B}_\tau$. Acts to push the Markov chain back into the constraint set $\mathcal{B}_\tau$ if it wanders outside of it.



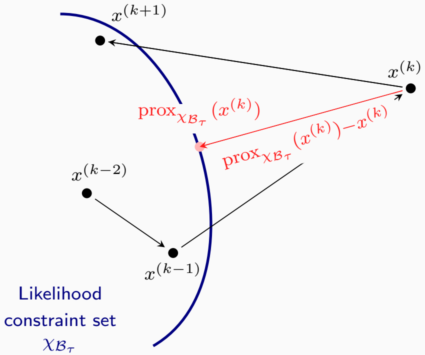
Proximal nested sampling intuition

Recall proximal nested sampling Markov chain:

$$x^{(k+1)} = x^{(k)} - \frac{\delta}{2}\nabla f(x^{(k)}) - \frac{\delta}{2\lambda}\boxed{\left[x^{(k)} - \text{prox}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right]} + \sqrt{\delta} w^{(k+1)}.$$

1. $x^{(k)}$ **is already in** $\mathcal{B}_\tau$: term $\left[x^{(k)} - \text{prox}^{\lambda}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right]$ disappears and recover usual Langevin MCMC.

2. $x^{(k)}$ **is not in** $\mathcal{B}_\tau$: a step is also taken in the direction $-\left[x^{(k)} - \text{prox}^{\lambda}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right]$, which moves the next iteration in the direction of the projection of $x^{(k)}$ onto the convex set $\mathcal{B}_\tau$. Acts to push the Markov chain back into the constraint set $\mathcal{B}_\tau$ if it wanders outside of it.



Proximal nested sampling

A subsequent **Metropolis-Hastings** step guarantees hard likelihood constraint is satisfied.

Proximal nested sampling

A subsequent **Metropolis-Hastings** step guarantees hard likelihood constraint is satisfied.

In practice need to compute $\text{prox}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})$, including measurement operator.

For sparsity-promoting non-differentiable priors $f(x)$, can also make Moreau-Yosida approximation $f^\lambda(x)$ and leverage prox to compute gradient ∇f^λ.

Proximal nested sampling

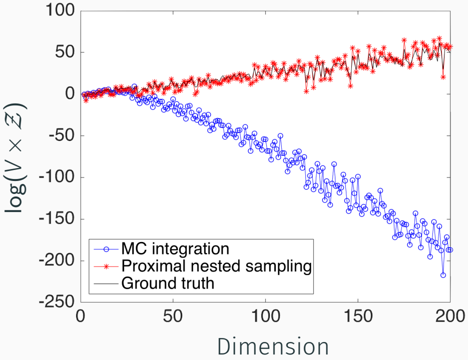
A subsequent **Metropolis-Hastings** step guarantees hard likelihood constraint is satisfied.

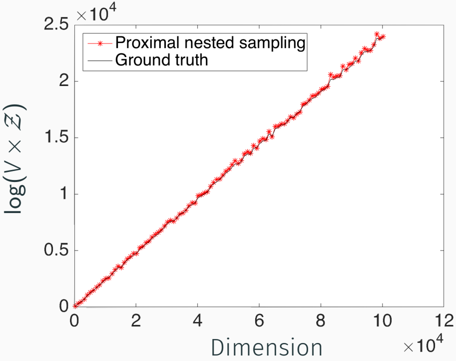
In practice need to compute $\mathrm{prox}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})$, including measurement operator.

For sparsity-promoting non-differentiable priors $f(x)$, can also make Moreau-Yosida approximation $f^\lambda(x)$ and leverage prox to compute gradient ∇f^λ.

Many further details regarding **explicit forms of proximal nested sampling** for common priors and likelihoods and how to compute proximity operators efficiently (Cai, McEwen & Pereyra 2022; arXiv:2106.03646).

Validation on Gaussian problem





Comparison of proximal nested sampling (red), naive MC integration (blue) and ground truth (black).

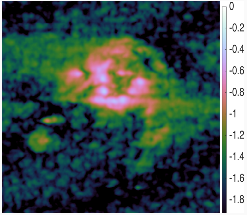
Measurement model misspecification experiment

Consider ground truth model $\mathbf{\Phi} = \mathsf{M}_{\text{truth}}\mathsf{F}$ to simulate observational data $\boldsymbol{y}$.

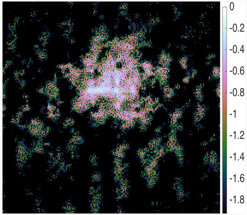
However, when solving the inverse problem consider misspecified models M_{γ}, where $\gamma > 0$ encodes the level of misspecification (mimics incorrectly specified wavelength).

Compute the model evidence using **proximal nested sampling**, using evidence to distinguish correct model.

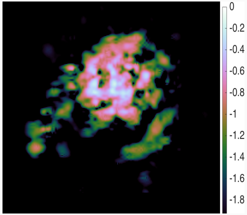
Measurement model misspecification experiment



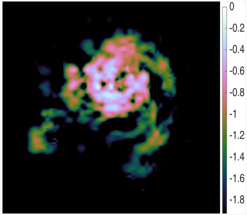
Dirty map



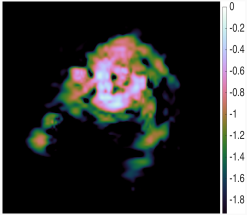
$\mathbf{\Phi} = \mathsf{M}_{0.12}\mathsf{F}$



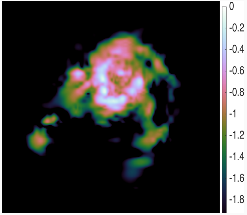
$\mathbf{\Phi} = \mathsf{M}_{0.09}\mathsf{F}$



$\mathbf{\Phi} = \mathsf{M}_{0.06}\mathsf{F}$



$\mathbf{\Phi} = \mathsf{M}_{0.03}\mathsf{F}$



$\mathbf{\Phi} = \mathsf{M}_{\mathrm{truth}}\mathsf{F}$

Measurement model misspecification experiment

Model	$\log \mathcal{Z}$	RMSE (Requires ground truth)
$\mathbf{\Phi} = \mathsf{M}_{\text{truth}}\mathsf{F}$	$-4.47 \times 10^3 \pm 0.08$	3.40
$\mathbf{\Phi} = \mathsf{M}_{0.03}\mathsf{F}$	$-4.88 \times 10^3 \pm 0.08$	7.85
$\mathbf{\Phi} = \mathsf{M}_{0.06}\mathsf{F}$	$-5.63 \times 10^3 \pm 0.08$	12.01
$\mathbf{\Phi} = \mathsf{M}_{0.09}\mathsf{F}$	$-9.21 \times 10^3 \pm 0.07$	15.71
$\mathbf{\Phi} = \mathsf{M}_{0.12}\mathsf{F}$	$-1.44 \times 10^4 \pm 0.08$	18.08

Evidence computed by proximal nested sampling correctly classifies models.

Proxnest code

GitHub ProxNest | Tests passing | Docs passing | codecov 95% | pypi package 0.0.7 | License GPL | arXiv 2106.03646

Proximal nested sampling for high-dimensional Bayesian model selection

`ProxNest` is an open source, well tested and documented Python implementation of the *proximal nested sampling* framework (Cai et al. 2022) to compute the Bayesian model evidence or marginal likelihood in high-dimensional log-convex settings. Furthermore, non-smooth sparsity-promoting priors are also supported.

Github: `https://github.com/astro-informatics/proxnest`

Docs: `https://astro-informatics.github.io/proxnest`

Summary

Summary

Many science questions are **questions of model comparison**
$\Rightarrow$ **Bayesian model comparison**.

Many outstanding challenges:

- Extending to general sampling strategies.
- Extending to simulation-based inference (likelihood-free inference).
- Scaling to high-dimensions.
- Learned data-driven priors.

1. Learnt harmonic mean estimator for Bayesian model comparison
   (McEwen, Wallis, Price & Docherty 2021; arXiv:2111.12720)
2. Bayesian model comparison for simulation-based inference
   (Spurio Mancini, Docherty, Price & McEwen 2022; arXiv:2207.04037).
3. Proximal nested sampling for high-dimensional Bayesian model comparison
   (Cai, McEwen & Pereyra 2022; arXiv:2106.03646)