Bayesian model comparison in the era of AI

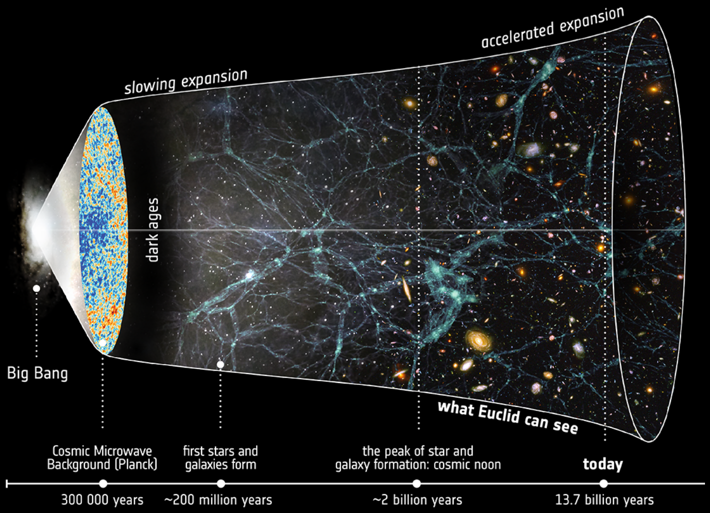
Jason D. McEwen
www.jasonmcewen.org

Scientific AI (SciAI) Group
Mullard Space Science Laboratory (MSSL), University College London (UCL)

Computational and Statistical Machine Learning in the Sciences, London Mathematical Society, 2024

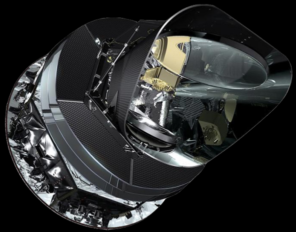
Motivation and introduction

Cosmic timeline

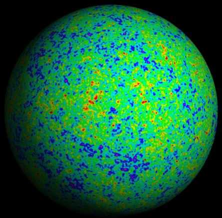


Cosmic microwave background (CMB) radiation

What is the origin of structure in our Universe?



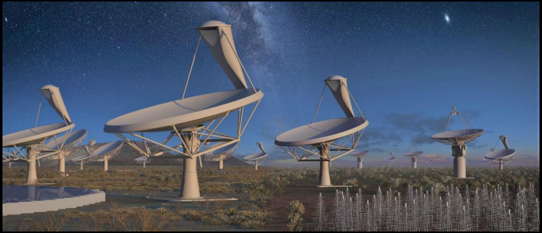
Planck satellite



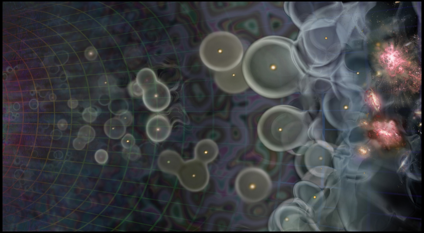
CMB

Epoch of reionisation

How did the first luminous objects in the Universe form?



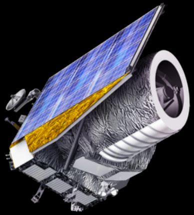
Square Kilometre Array (SKA)



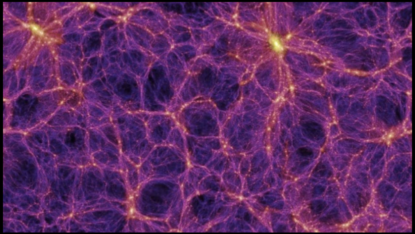
Ionised bubbles in neutral hydrogen

Large-scale structure of the Universe

What is the nature of dark energy?



Euclid satellite



Large-scale structure

Model selection in cosmology

These are questions of **model selection**.

These are questions of **model selection**.

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

Model selection in cosmology

These are questions of **model selection**.

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

$\rightsquigarrow$ Bayesian model selection

Bayesian inference: parameter estimation

Bayes' theorem

$$\underbrace{p(\theta \mid \boldsymbol{x}, M)}_{\text{posterior}} = \frac{\overbrace{p(\boldsymbol{x} \mid \theta, M)}^{\text{likelihood}} \; \overbrace{p(\theta \mid M)}^{\text{prior}}}{\underbrace{p(\boldsymbol{x} \mid M)}_{\text{evidence}}} = \frac{\overbrace{\mathcal{L}(\theta)}^{\text{likelihood}} \; \overbrace{\pi(\theta)}^{\text{prior}}}{\underbrace{z}_{\text{evidence}}},$$

for parameters θ, model M and observed data $\boldsymbol{x}$.

Bayes' theorem

$$\underbrace{p(\theta \mid \mathbf{x}, M)}_{\text{posterior}} = \frac{\overbrace{p(\mathbf{x} \mid \theta, M)}^{\text{likelihood}}\ \overbrace{p(\theta \mid M)}^{\text{prior}}}{\underbrace{p(\mathbf{x} \mid M)}_{\text{evidence}}} = \frac{\overbrace{\mathcal{L}(\theta)}^{\text{likelihood}}\ \overbrace{\pi(\theta)}^{\text{prior}}}{\underbrace{Z}_{\text{evidence}}},$$

for parameters θ, model M and observed data $\mathbf{x}$.

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

Bayesian inference: model comparison

By Bayes' theorem for model M_j:

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

Bayesian inference: model comparison

By Bayes' theorem for model M_j:

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

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

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

Bayesian inference: model comparison

By Bayes' theorem for model M_j:

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

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

$$\frac{p(M_1 \,|\, \mathbf{x})}{p(M_2 \,|\, \mathbf{x})} = \frac{p(\mathbf{x} \,|\, M_1)}{p(\mathbf{x} \,|\, M_2)} \times \frac{p(M_1)}{p(M_2)} .$$

posterior odds — Bayes factor — prior odds

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

$$z = p(\mathbf{x} \,|\, M) = \int \mathrm{d}\theta \, \mathcal{L}(\theta) \, \pi(\theta) .$$

Bayesian inference: model comparison

By Bayes' theorem for model M_j:

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

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

$$\underbrace{\frac{p(M_1 \,|\, \mathbf{x})}{p(M_2 \,|\, \mathbf{x})}}_{\text{posterior odds}} = \underbrace{\frac{p(\mathbf{x} \,|\, M_1)}{p(\mathbf{x} \,|\, M_2)}}_{\text{Bayes factor}} \times \underbrace{\frac{p(M_1)}{p(M_2)}}_{\text{prior odds}} \,.$$

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

$$z = p(\mathbf{x} \,|\, M) = \int \mathrm{d}\theta \, \mathcal{L}(\theta) \, \pi(\theta) \,.$$

$\rightsquigarrow$ Challenging computational problem.

SciAI UCL

Jason McEwen

Model scenarios and Bayesian consistency

$\mathcal{M}$**-closed scenario**	$\mathcal{M}$**-open scenario** (model misspecification)
True model is **in** set of models $\mathcal{M}$.	True model is **not in** set of models $\mathcal{M}$.

Model scenarios and Bayesian consistency

$\mathcal{M}$-closed scenario

True model is **in** set of models $\mathcal{M}$.

$\mathcal{M}$-open scenario (model misspecification)

True model is **not in** set of models $\mathcal{M}$.

Parameter estimation consistency

$\mathcal{M}$-closed: parameter point estimators (MMSE and MAP) converge to true parameters (Bernardo & Smith 1994).

$\mathcal{M}$-open: parameter point estimators (MMSE and MAP) converge to best-fit parameters of model considered (Bernardo & Smith 1994).

Model scenarios and Bayesian consistency

$\mathcal{M}$-closed scenario	$\mathcal{M}$-open scenario (model misspecification)
True model is **in** set of models $\mathcal{M}$.	True model is **not in** set of models $\mathcal{M}$.

Parameter estimation consistency

$\mathcal{M}$-closed: parameter point estimators (MMSE and MAP) converge to true parameters (Bernardo & Smith 1994).

$\mathcal{M}$-open: parameter point estimators (MMSE and MAP) converge to best-fit parameters of model considered (Bernardo & Smith 1994).

Model selection consistency

$\mathcal{M}$-closed: posterior model distribution concentrates on true model (Dawid 2011).

$\mathcal{M}$-open: posterior model distribution concentrates on model closest in KL divergence (Dawid 2011).

Model scenarios and Bayesian consistency

$\mathcal{M}$-closed scenario	**$\mathcal{M}$-open scenario** (model misspecification)
True model is **in** set of models $\mathcal{M}$.	True model is **not in** set of models $\mathcal{M}$.

Parameter estimation consistency

$\mathcal{M}$-closed: parameter point estimators (MMSE and MAP) converge to true parameters (Bernardo & Smith 1994).

$\mathcal{M}$-open: parameter point estimators (MMSE and MAP) converge to best-fit parameters of model considered (Bernardo & Smith 1994).

Model selection consistency

$\mathcal{M}$-closed: posterior model distribution concentrates on true model (Dawid 2011).

$\mathcal{M}$-open: posterior model distribution concentrates on model closest in KL divergence (Dawid 2011).

$\rightsquigarrow$ Bayesian parameter estimation and model selection are consistent.

Challenge of Bayesian model selection

Naive Monte Carlo integration to compute marginal likelihood not effective.

Challenge of Bayesian model selection

Naive Monte Carlo integration to compute marginal likelihood not effective.

Require **tailored computational techniques**, such as nested sampling (Skilling 2006).

Challenge of Bayesian model selection

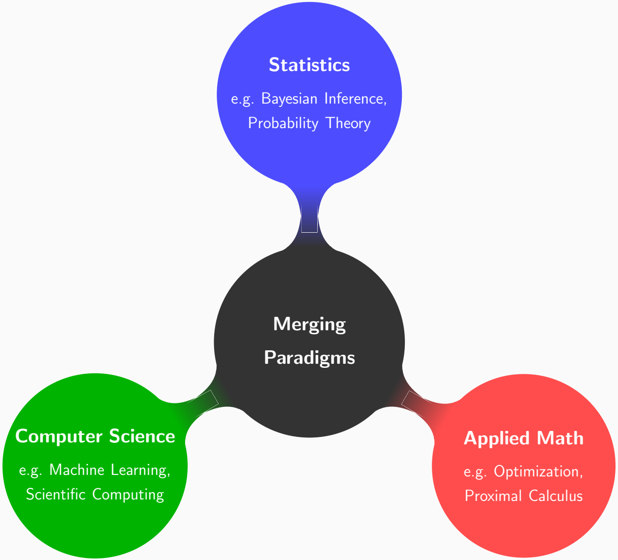
Naive Monte Carlo integration to compute marginal likelihood not effective.

Require **tailored computational techniques**, such as nested sampling (Skilling 2006).

Challenges:

- Support **general sampling** strategies.
- Support **simulation-based inference (SBI)** and **variational inference (VI)**.
- Scale to **high-dimensions**.
- Support **data-driven AI priors** (e.g. priors captured by generative models).

Merging paradigms



Outline

1. Motivation and introduction

2. AI-assisted Bayesian model comparison

3. AI data-driven priors in high-dimensions

AI-assisted Bayesian model comparison

Learning model posterior odds ratio

Leverage the **likelihood ratio trick** (Goodfellow *et al.* 2014, Cranmer *et al.* 2020) to learn model posterior odds ratio directly.

Train a classifier to distinguish models, *e.g.* with cross-entropy loss, which learns ratio

$$r(x) = \frac{p(M_1 \mid x)}{p(M_2 \mid x)}.$$

Numerous works considering this approach or variants (Radev *et al.* 2021, Spurio Mancini *et al.* McEwen 2023, Elsemüller *et al.* 2024, Jeffrey *et al.* 2024, Karchev *et al.* 2023).

Learning model posterior odds ratio

Leverage the **likelihood ratio trick** (Goodfellow *et al.* 2014, Cranmer *et al.* 2020) to learn model posterior odds ratio directly.

Train a classifier to distinguish models, *e.g.* with cross-entropy loss, which learns ratio

$$r(x) = \frac{p(M_1 \,|\, x)}{p(M_2 \,|\, x)}.$$

Numerous works considering this approach or variants (Radev *et al.* 2021, Spurio Mancini *et al.* McEwen 2023, Elsemüller *et al.* 2024, Jeffrey *et al.* 2024, Karchev *et al.* 2023).

$\rightsquigarrow$ No consistency guarantees for $\mathcal{M}$-open scenario.

Nested sampling: reparameterising the likelihood

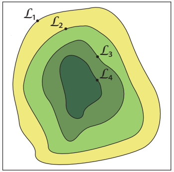
Nested sampling: ingenious approach to efficiently evaluate the evidence (Skilling 2006).

Group the parameter space Ω into a series of **nested subspaces**: $\Omega_{L^*} = \{x \mid \mathcal{L}(x) \geq L^*\}$. Define the prior volume ξ within Ω_{L^*} by

$$\xi(L^*) = \int_{\Omega_{L^*}} \pi(x)\mathrm{d}x.$$

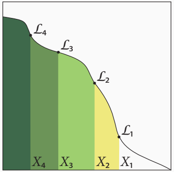
Evidence can then be rewritten as

$$z = \int_0^1 \mathcal{L}(\xi)\mathrm{d}\xi.$$



Feroz *et al.* (2013)

Nested subspaces



Feroz *et al.* (2013)

Reparameterised likelihood

SciAI UCL

Jason McEwen

Nested sampling: reparameterising the likelihood

Nested sampling: ingenious approach to efficiently evaluate the evidence (Skilling 2006).

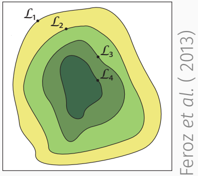
Group the parameter space Ω into a series of **nested subspaces**: $\Omega_{L^*} = \{x \,|\, \mathcal{L}(x) \geq L^*\}$. Define the prior volume ξ within Ω_{L^*} by

$$\xi(L^*) = \int_{\Omega_{L^*}} \pi(x) \mathrm{d}x.$$

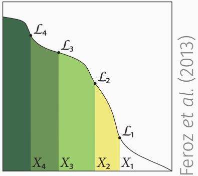
Evidence can then be rewritten as

$$z = \int_0^1 \mathcal{L}(\xi) \mathrm{d}\xi.$$

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



Nested subspaces



Reparameterised likelihood

The problem of nested sampling

Many highly effective nested sampling algorithms (for a review see Ashton *et al.* 2022).

Method of choice for the past almost two decades!

However, nested sampling has a **fundamental problem**...

The problem of nested sampling

Many highly effective nested sampling algorithms (for a review see Ashton *et al.* 2022).

Method of choice for the past almost two decades!

However, nested sampling has a fundamental problem...

Nested sampling tightly couples sampling strategy to marginal likelihood calculation.

As the name suggests, **one must sample in a nested manner**.

SciAI UCL

The problem of nested sampling

Many highly effective nested sampling algorithms (for a review see Ashton *et al.* 2022).

Method of choice for the past almost two decades!

However, nested sampling has a fundamental problem...

Nested sampling tightly couples sampling strategy to marginal likelihood calculation.

As the name suggests, **one must sample in a nested manner**.

- **Precludes** many alternative **accelerated sampling** strategies that scale to high-dimensions.
- **Precludes** use in many **simulation-based inference (SBI)** and **variational inference (VI)** settings, where one draws posterior samples directly.

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

$$\rho = \mathbb{E}_{p(\theta \,|\, \boldsymbol{x})}\left[\frac{1}{\mathcal{L}(\theta)}\right]$$

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

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

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

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

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

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

SciAI UCL
Jason McEwen

Harmonic mean relationship (Newton & Raftery 1994)

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

Original harmonic mean estimator (Newton & Raftery 1994)

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

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

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

Original harmonic mean estimator (Newton & Raftery 1994)

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

✔ Only requires posterior samples!

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

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

Original harmonic mean estimator (Newton & Raftery 1994)

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

✔ **Only requires posterior samples!**

✖ But **can fail catastrophically!** (Neal 1994)

SciAI UCL

Learned harmonic mean estimator

Propose the ***learned* harmonic mean estimator**, leveraging AI to solve the catastrophic failure of the original harmonic mean (McEwen, Wallis, Price, Spurio Mancini 2021; arXiv:2111.12720).

Chris Wallis

Matt Price

Alessio Spurio Mancini

SciAI UCL

Importance sampling interpretation of harmonic mean estimator

Alternative interpretation of harmonic mean relationship:

importance sampling

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

Importance sampling interpretation of harmonic mean estimator

Alternative interpretation of harmonic mean relationship:

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

Importance sampling interpretation:

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

Importance sampling interpretation of harmonic mean estimator

Alternative interpretation of harmonic mean relationship:

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

Importance sampling interpretation:

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

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

Importance sampling failure mode when sampling density is posterior and target is prior.

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

$$\rho = \mathbb{E}_{p(\theta \,|\, \boldsymbol{x})} \left[\frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \right] = \frac{1}{z}$$

Normalised distribution $\varphi(\theta)$ **now plays the role of the importance sampling target**
$\rightsquigarrow$ must **not** have fatter tails than posterior.

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

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

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: (McEwen *et al.* 2021)

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

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: (McEwen *et al.* 2021)

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

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

Learned harmonic mean estimator

Learn an approximation of the optimal target distribution:

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

SciAI UCL

Jason McEwen

Learned harmonic mean estimator

Learn an approximation of the optimal target distribution:

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

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

Constraining tails of target approach 1: bespoke optimisation problem

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

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

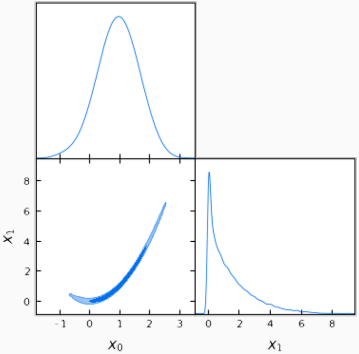
Solve by bespoke **mini-batch stochastic gradient descent**.

Cross-validation to select density estimation model and hyperparameters.

Rosenbrock example

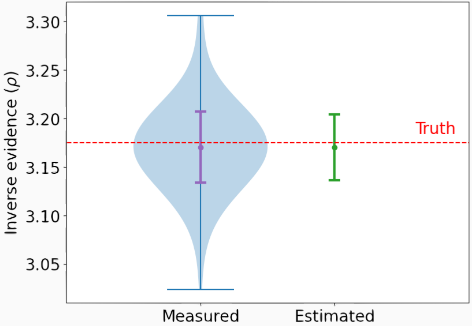
Rosenbrock function is the classical example of a **pronounced thin curving degeneracy**, with likelihood defined by

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



Posterior recovered by MCMC sampling.

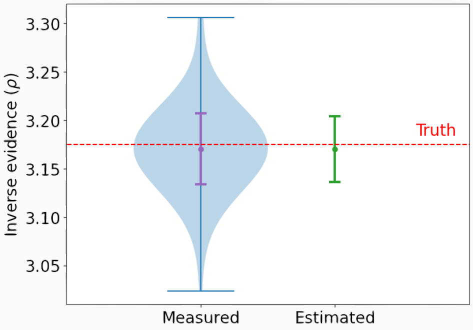
Rosenbrock example



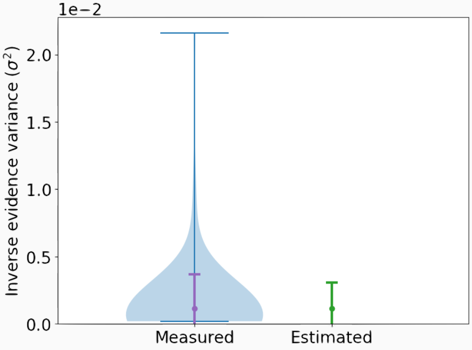
Reciprocal evidence

Accuracy of learnt harmonic mean estimator for Rosenbrock example.

Rosenbrock example



Reciprocal evidence



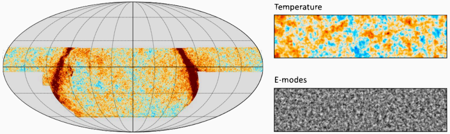
Variance of reciprocal evidence

Accuracy of learnt harmonic mean estimator for Rosenbrock example.

Atacama Cosmology Telescope (ACT) analysis

Compare ΛCDM (**Einstein's cosmological constant**) vs $w_0 w_a$CDM (**dynamical dark energy**) using learned harmonic mean (McEwen *et al.*2021) with ACT data (Aiola *et al.* 2020).

Atacama Cosmology Telescope (ACT)



CMB observations

7D vs 9D models:	ΛCDM	$w_0 w_a$CDM	$\log \mathrm{BF}_{\Lambda\mathrm{CDM}\text{-}w_0 w_a\mathrm{CDM}}$
Nested sampling	-168.92 ± 0.35	-169.38 ± 0.24	0.46 ± 0.42
Learned harmonic mean	-168.87 ± 0.29	-169.32 ± 0.25	0.45 ± 0.38

$\rightsquigarrow$ ΛCDM mildly favoured $\rightsquigarrow$ 3× **acceleration**

Constraining tails of target approach 2: normalizing flows

Learned harmonic mean with normalizing flows (Polanska *et al.* 2024; arXiv:2405.05969)

Elegant way to constrain tails of target distribution $\varphi(\theta)$.

Alicja Polanska

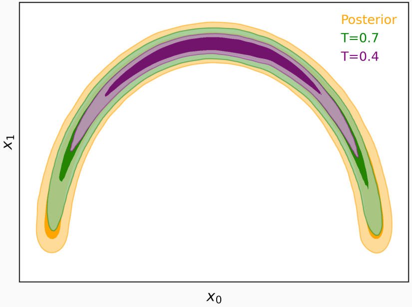
Matt Price

Davide Piras

Alessio Spurio Mancini

Constraining tails of target approach 2: normalizing flows

Concentrate probability of target by lowering temperature T (variance) of the base distribution.



- **Flexible**: no bespoke training; can vary T after training.
- **Robust**: only one hyperparameter T that does not require fine tuning.
- **Scalable**: flows scale to higher dimensions than classical density estimators.

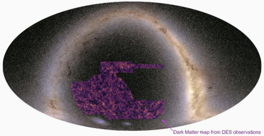
SciAI UCL

Jason McEwen

Dark Enery Survey (DES)-like analysis

Compare ΛCDM vs wCDM using learned harmonic mean with DES Year-1 lensing and clustering simulations (Polanska *et al.* 2024).

Dark Energy Survey (DES)



Lensing observations

20D vs 21D models:	$\log(z_{\Lambda\text{CDM}})$	$\log(z_{w\text{CDM}})$	$\log \text{BF}_{\Lambda\text{CDM-}w\text{CDM}}$	Computation time (64 CPU cores)
Nested sampling	-65.21 ± 0.32	-67.44 ± 0.32	2.23 ± 0.45	94 hours
Learned harmonic mean	$-65.262^{+0.011}_{-0.011}$	$-67.407^{0.009}_{-0.009}$	$2.145^{0.014}_{-0.014}$	16 hours

$\rightsquigarrow$ **6$\times$ acceleration**

Leveraging AI to accelerate Bayesian inference further

4 pillars of AI-accelerated Bayesian inference (Piras *et al.* 2024; arXiv:2405.12965).

Leveraging AI to accelerate Bayesian inference further

4 pillars of AI-accelerated Bayesian inference (Piras *et al.* 2024; arXiv:2405.12965).

1. **Emulation** to accelerate physical model encapsulated in likelihood,
   e.g. CosmoPower (Spurio Mancini *et al.* 2022, Piras & Spurio Mancini 2023)

4 pillars of AI-accelerated Bayesian inference (Piras *et al.* 2024; arXiv:2405.12965).

1. **Emulation** to accelerate physical model encapsulated in likelihood, *e.g.* CosmoPower (Spurio Mancini *et al.* 2022, Piras & Spurio Mancini 2023)
2. **Differentiable and probabilistic programming** to accelerate gradient calculations and development of statistical models, *e.g.* JAX, NumPyro

Leveraging AI to accelerate Bayesian inference further

4 pillars of AI-accelerated Bayesian inference (Piras *et al.* 2024; arXiv:2405.12965).

1. **Emulation** to accelerate physical model encapsulated in likelihood, *e.g.* CosmoPower (Spurio Mancini *et al.* 2022, Piras & Spurio Mancini 2023)
2. **Differentiable and probabilistic programming** to accelerate gradient calculations and development of statistical models, *e.g.* JAX, NumPyro
3. **Scalable (gradient-based) MCMC sampling** to accelerate sampling and parameter estimation, *e.g.* NUTS

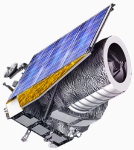
Leveraging AI to accelerate Bayesian inference further

4 pillars of AI-accelerated Bayesian inference (Piras *et al.* 2024; arXiv:2405.12965).

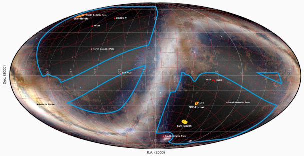
1. **Emulation** to accelerate physical model encapsulated in likelihood, *e.g.* CosmoPower (Spurio Mancini *et al.* 2022, Piras & Spurio Mancini 2023)
2. **Differentiable and probabilistic programming** to accelerate gradient calculations and development of statistical models, *e.g.* JAX, NumPyro
3. **Scalable (gradient-based) MCMC sampling** to accelerate sampling and parameter estimation, *e.g.* NUTS
4. **Scalable and decoupled marginal likelihood computation** to accelerate model selection, *e.g.* learned harmonic mean (McEwen *et al.*2021, Polanska *et al.* 2024)

Euclid (Stage IV survey)-like analysis

Compare ΛCDM vs $w_0 w_a$CDM leveraging **4 pillars of AI-acceleration** with Euclid-like lensing and clustering simulations (Piras *et al.* 2024).



Euclid satellite

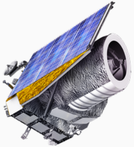


Observation field

37D vs 39D models:	$\log(z_{\Lambda\mathrm{CDM}})$	$\log(z_{w_0 w_a \mathrm{CDM}})$	$\log \mathrm{BF}_{\Lambda\mathrm{CDM}\text{-}w_0 w_a \mathrm{CDM}}$	Total computation time
Classical	-107.03 ± 0.27	-107.81 ± 0.74	0.78 ± 0.79	8 months (48 CPUs)
AI-accelerated (ours)	40956.55 ± 0.06	40955.03 ± 0.04	1.53 ± 0.07	2 days (12 GPUs)

$\rightsquigarrow$ $120\times$ **acceleration**

Extend to combined 3× Stage IV Survey-like lensing and clustering simulations (Piras *et al.* 2024).


Euclid satellite


Rubin observatory


Roman satellite

157D vs 159D models:	$\log(z_{\Lambda\mathrm{CDM}})$	$\log(z_{w_0 w_a\mathrm{CDM}})$	log BF	Total computation time
Classical	Unfeasible	Unfeasible	Unfeasible	12 years projected (48 CPUs)
AI-accelerated (ours)	$406689.6^{+0.5}_{-0.3}$	$406687.7^{+0.5}_{-0.3}$	$1.9^{+0.7}_{-0.5}$	8 days (24 GPUs)

$\rightsquigarrow$ **Opens up new analyses (550× acceleration)**

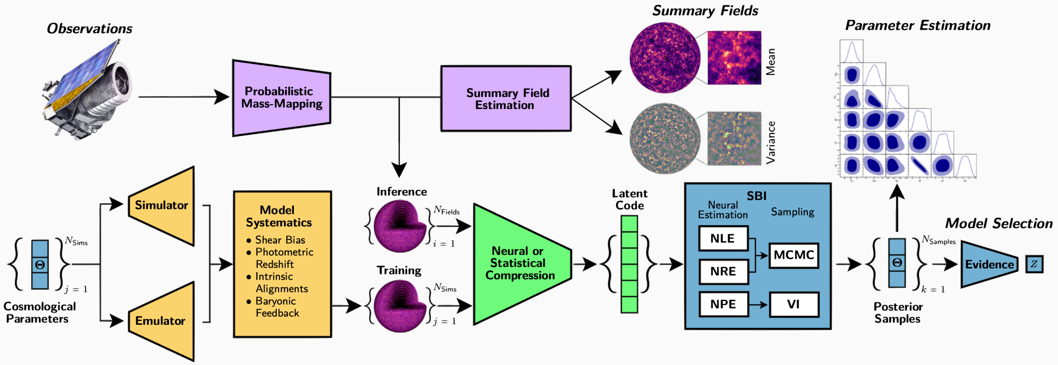
Simulation-based inference (SBI)

Simulation-based inference (aka. likelihood-free inference) seeks to perform Bayesian inference by **estimating the posterior** $p(\theta \,|\, \boldsymbol{x}_o, M)$ of **parameters** θ for **observed data** $\boldsymbol{x}_o$ using **simulations only**.

Key advantages:

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

Field-level SBI pipeline for Euclid cosmic shear



Could field-level SBI distinguish dynamical dark energy?

Recent results from DESI experiment provide tantalising hints of dynamical dark energy (Adame *et al.* 2024a, 2024b).

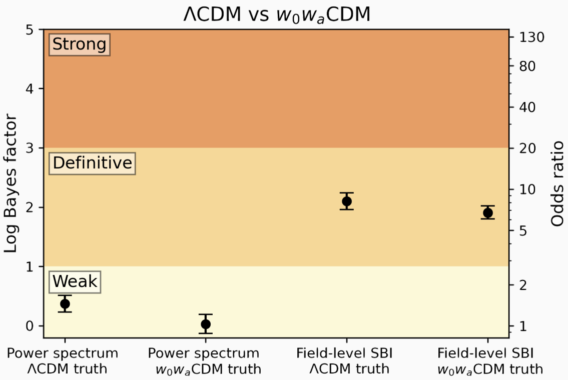
If these results reflected true underlying nature of the Universe, **could a field-level SBI analysis of a Stage IV survey distinguish dynamical dark energy definitively?**
(Spurio Mancini *et al.* 2024; arXiv:2410.10616)

Alessio Spurio Mancini

Kiyam Lin

Could field-level SBI distinguish dynamical dark energy?

If these results reflected true underlying nature of the Universe, **could a field-level SBI analysis of a Stage IV survey distinguish dynamical dark energy definitively?** (Spurio Mancini *et al.* 2024; arXiv:2410.10616)



AI data-driven priors in high-dimensions

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

Constrained sampling formulation

Consider case where likelihood and prior of the form

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

Likelihood Prior

where $g = -\log \mathcal{L}$ is convex lower semicontinuous function (prior need not be log-convex).

Constrained sampling formulation

Consider case where likelihood and prior of the form

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

Likelihood   Prior

where $g = -\log \mathcal{L}$ is convex lower semicontinuous function (prior need not be log-convex).

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

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

Constrained sampling formulation

Consider case where likelihood and prior of the form

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

Likelihood   Prior

where $g = -\log \mathcal{L}$ is convex lower semicontinuous function (prior need not be log-convex).

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

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

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

Constrained sampling formulation

Taking the logarithm, we can write

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

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

$$\mathcal{B}_\tau := \{x \mid -\log \mathcal{L}(x) < \tau\},$$

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

MCMC sampling with Langevin dynamics

Require MCMC sampling strategy that can scale to **high-dimensions**.

If target distribution $p(x)$ is differentiable can adopt **Langevin dynamics**.

Langevin diffusion process $x(t)$, with $p(x)$ as stationary distribution:

$$\mathrm{d}x(t) = \frac{1}{2}\mathrm{d}t + \mathrm{d}w(t),$$

where w is Brownian motion.

MCMC sampling with Langevin dynamics

Require MCMC sampling strategy that can scale to **high-dimensions**.

If target distribution $p(\boldsymbol{x})$ is differentiable can adopt **Langevin dynamics**.

Langevin diffusion process $\boldsymbol{x}(t)$, with $p(\boldsymbol{x})$ as stationary distribution:

$$\mathrm{d}\boldsymbol{x}(t) = \frac{1}{2}\nabla \log p\big(\boldsymbol{x}(t)\big)\mathrm{d}t + \mathrm{d}\boldsymbol{w}(t),$$

where $\boldsymbol{w}$ is Brownian motion.

Need gradients so **not directly applicable** $\Rightarrow$ adopt Morea-Yosida approximation.

MCMC sampling with Langevin dynamics

Require MCMC sampling strategy that can scale to **high-dimensions**.

If target distribution $p(x)$ is differentiable can adopt **Langevin dynamics**.

Langevin diffusion process $x(t)$, with $p(x)$ as stationary distribution:

$$\mathrm{d}x(t) = \frac{1}{2} \underbrace{\nabla \log p\big(x(t)\big)}_{\text{Gradient}} \mathrm{d}t + \mathrm{d}w(t),$$

where w is Brownian motion.

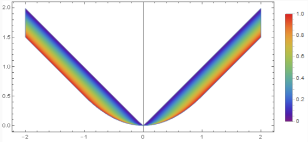
Need gradients so **not directly applicable** $\Rightarrow$ adopt Morea-Yosida approximation.

SciAI UCL

Morea-Yosida (M-Y) approximation

The Morea-Yosida approximation of a convex function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is given by the **infimal convolution**:

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



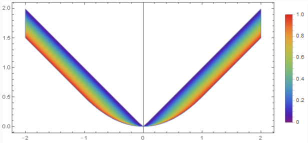
M-Y envelope of $|x|$ for varying λ.

Moreau-Yosida approximation

Morea-Yosida (M-Y) approximation

The Morea-Yosida approximation of a convex function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is given by the **infimal convolution**:

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



M-Y envelope of $|x|$ for varying λ.

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

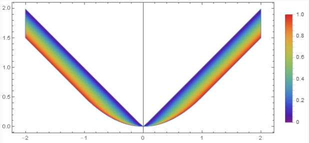
1. As $\lambda \rightarrow 0, f^\lambda(x) \rightarrow f(x)$
2. $\nabla f^\lambda(x) = (x - \text{prox}_f^\lambda(x))/\lambda$

Moreau-Yosida approximation

Morea-Yosida (M-Y) approximation

The Morea-Yosida approximation of a convex function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is given by the **infimal convolution**:

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



M-Y envelope of $|x|$ for varying λ.

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

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

- **Regularise** non-differentiable function (e.g. likelihood level-set constraint!)
- **Compute gradient** by prox.
- Leverage **gradient-based Bayesian computation**.

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:

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

Proximal nested sampling intuition

Recall proximal nested sampling Markov chain (from previous slide):

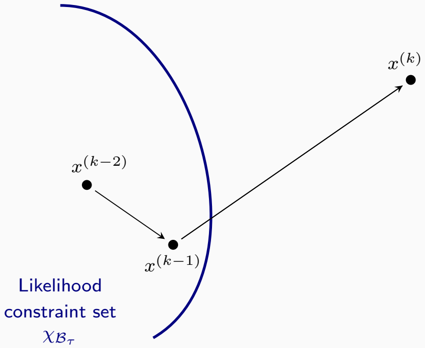
$$x^{(k+1)} = x^{(k)} + \frac{\delta}{2}\nabla \log \pi(x^{(k)}) - \frac{\delta}{2\lambda}\left[x^{(k)} - \text{prox}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right] + \sqrt{\delta}w^{(k+1)}.$$

Proximal nested sampling intuition

Recall proximal nested sampling Markov chain (from previous slide):

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

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.



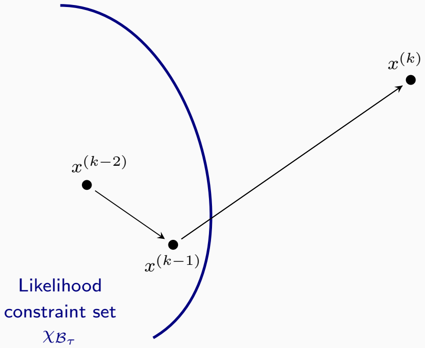
Proximal nested sampling intuition

Recall proximal nested sampling Markov chain (from previous slide):

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

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.

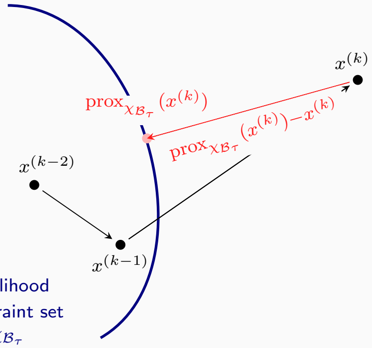


Proximal nested sampling intuition

Recall proximal nested sampling Markov chain (from previous slide):

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

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.



SciAI UCL
Jason McEwen

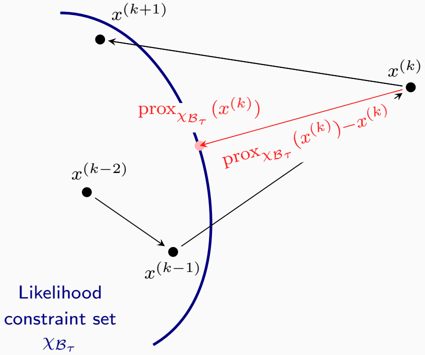
Proximal nested sampling intuition

Recall proximal nested sampling Markov chain (from previous slide):

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

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.



A subsequent Metropolis-Hastings step can be introduced to **guarantee hard likelihood constraint is satisfied**.

A subsequent Metropolis-Hastings step can be introduced to **guarantee hard likelihood constraint is satisfied**.

For sparsity-promoting non-differentiable priors $f(x)$ (e.g. $-\log \pi(\mathbf{x}) = \|\Psi^\dagger \mathbf{x}\|_1$), can also make Moreau-Yosida approximation $f^\lambda(\mathbf{x})$ and leverage prox to compute gradient ∇f^λ:

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

But **how do we compute the proximity operators?**

Explicit forms of proximal nested sampling

But **how do we compute the proximity operators?**

Consider common imaging problem as example:

$$-\log \pi(x) = \left\|\Psi^\dagger x\right\|_1 + \text{const.}$$

Prior

$$\text{prox}^{\lambda}_{-\log \pi}(x) = x + \Psi\left(\text{soft}_{\lambda\mu}(\Psi^\dagger x') - \Psi^\dagger x\right),$$

SciAI UCL

Explicit forms of proximal nested sampling

But **how do we compute the proximity operators?**

Consider common imaging problem as example:

$$-\log \mathcal{L}(x) = \|y - \mathbf{\Phi} x\|_2^2 + \text{const.}$$

Likelihood

$$-\log \pi(x) = \|\Psi^\dagger x\|_1 + \text{const.}$$

Prior

Straightforward when $\mathbf{\Phi}$ is identity.

Otherwise express as equivalent saddle-point problem and solve using primal-dual method.

$$\text{prox}^{\lambda}_{-\log \pi}(x) = x + \Psi\big(\text{soft}_{\lambda\mu}(\Psi^\dagger x') - \Psi^\dagger x\big),$$

SciAI UCL

Computing proximal operator for likelihood

Prox for the likelihood is equivalent to the saddle-point problem:

$$\min_{x \in \mathbb{R}^d} \max_{z \in \mathbb{C}^K} \left\{ z^\dagger \Phi x - \chi^*_{\mathcal{B}'_{\tau'}}(z) + \|x - x'\|_2^2/2 \right\}.$$

Solve iteratively by primal dual method:

1. $z^{(i+1)} = z^{(i)} + \delta_1 \Phi \bar{x}^{(i)} - \text{prox}_{\chi_{\mathcal{B}'_{\tau'}}}(z^{(i)} + \delta_1 \Phi \bar{x}^{(i)}),$

   where $\text{prox}_{\chi_{\mathcal{B}'_{\tau'}}}(z) = \text{proj}_{\mathcal{B}'_{\tau'}}(z) = \begin{cases} z, & \text{if } z \in \mathcal{B}'_{\tau'}, \\ \frac{z-y}{\|z-y\|_2}\sqrt{2\tau\sigma^2} + y, & \text{otherwise.} \end{cases}$

2. $x^{(i+1)} = (x' + x^{(i)} - \delta_2 \Phi^\dagger z^{(i+1)})/2$
3. $\bar{x}^{(i+1)} = x^{(i+1)} + \delta_3 (x^{(i+1)} - x^{(i)})$

Empirical Bayes: deep data-driven priors

Handcrafted priors (e.g. promoting sparsity in a wavelet basis) are **not expressive enough**.

Consider **empirical Bayes** approach with **data-driven priors** learned from training data.

Empirical Bayes: deep data-driven priors

Handcrafted priors (e.g. promoting sparsity in a wavelet basis) are **not expressive enough**.

Consider **empirical Bayes** approach with **data-driven priors** learned from training data.

Aim: **integrate learned deep data-driven priors** into proximal nested sampling.

Proximal nested sampling requires only likelihood to be convex, so **prior can be arbitrarily complex** (e.g. deep learned model).

Proximal nested sampling with deep data driven-priors

Proximal nested sampling with data driven-priors for physical scientists
(McEwen, Liaudat, Price, Cai & Pereyra 2023; arXiv:2307.00056)

Tobias Liaudat

Henry Aldridge

Matt Price

Xiaohao Cai

Marcelo Pereyra

Tweedie's formula

Consider noisy observations $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{x}, \sigma^2 I)$ of $\boldsymbol{x}$ sampled from some underlying prior.

Tweedie's formula gives the posterior expectation of $\boldsymbol{x}$ given $\boldsymbol{z}$ as

$$\mathbb{E}(\boldsymbol{x} \,|\, \boldsymbol{z}) = \boldsymbol{z} + \sigma^2 \nabla \log p(\boldsymbol{z}),$$

where $p(z)$ is the marginal distribution of z.

Tweedie's formula

Tweedie's formula

Consider noisy observations $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{x}, \sigma^2 I)$ of $\boldsymbol{x}$ sampled from some underlying prior.

Tweedie's formula gives the posterior expectation of $\boldsymbol{x}$ given $\boldsymbol{z}$ as

$$\mathbb{E}(\boldsymbol{x} \,|\, \boldsymbol{z}) = \boldsymbol{z} + \sigma^2 \nabla \log p(\boldsymbol{z}),$$

where $p(z)$ is the marginal distribution of z.

- Can be interpreted as a denoising strategy.
- Can be used to relate a denoiser (potentially a trained deep neural network) to the score $\nabla \log p(z)$.

Learning score of regularised prior

No guarantee that data-driven prior is well-suited for gradient-based Bayesian computation, *e.g.* it may not be differentiable or proper.

⇝ Consider **regularised prior** defined by Gaussian smoothing:

$$\pi_\epsilon(x) = (2\pi\epsilon)^{-d/2} \int \mathrm{d}x' \, \exp(| \, x - x' \|_2^2/(2\epsilon)) \, \pi(x').$$

SciAI UCL

Learning score of regularised prior

No guarantee that data-driven prior is well-suited for gradient-based Bayesian computation, *e.g.* it may not be differentiable or proper.

$\rightsquigarrow$ Consider **regularised prior** defined by Gaussian smoothing:

$$\pi_\epsilon(x) = (2\pi\epsilon)^{-d/2} \int \mathrm{d}x' \exp(\| x - x' \|_2^2/(2\epsilon)) \, \pi(x').$$

Consider **learned denoiser** D_ϵ trained to recover x from noisy observations $x_\epsilon \sim \mathcal{N}(x, \epsilon I)$.

By Tweedie's formula the score of the regualised prior related to the learned denoiser by

$$\nabla \log \pi_\epsilon(x) = \epsilon^{-1}(D_\epsilon(x) - x).$$

Proximal nested sampling with learned data-driven priors

Substituting the denoiser $\nabla \log \pi_\epsilon(x) = \epsilon^{-1}(D_\epsilon(x) - x)$ into the proximal nested sampling Markov chain update:

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

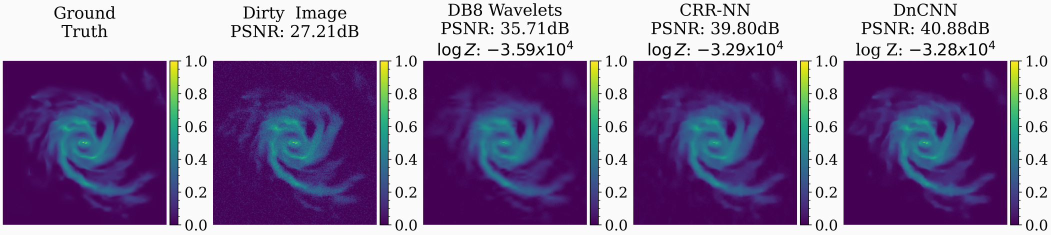
SciAI UCL

Hand-crafted vs data-driven priors

Consider simple Galaxy denoising inverse problem with:

- ▷ **hand-crafted prior** based on sparsity-promoting wavelet representation;
- ▷ **data-driven priors** based on a deep neural networks

  (Goujon et al. 2023, Ryu et al. 2019).



Which model best?

- ▷ SNR (require ground-truth) $\Rightarrow$ **data-driven priors best**;
- ▷ Bayesian evidence (no ground-truth knowledge) $\Rightarrow$ **data-driven priors best.**

Summary

Summary

- AI-assisted Bayesian model comparison
  - Learned harmonic mean (McEwen *et al.* 2021; arXiv:2111.12720)
  - Learned harmonic mean with normalizing flows (Polanska *et al.* 2024; arXiv:2405.05969)
  - 4 pillars of AI-accelerated Bayesian inference (Piras *et al.* 2024; arXiv:2405.12965)
  - Bayesian model comparison for SBI (Spurio Mancini *et al.* 2022; arXiv:2207.04037)
  - Field-level SBI model comparison (Spurio Mancini *et al.* 2024; arXiv:2410.10616)
- AI data-driven priors in high-dimensions
  - Proximal nested sampling (Cai *et al.* 2021; arXiv:2106.03646)
  - Learned proximal nested sampling (McEwen *et al.* 2023; arXiv:2307.00056)

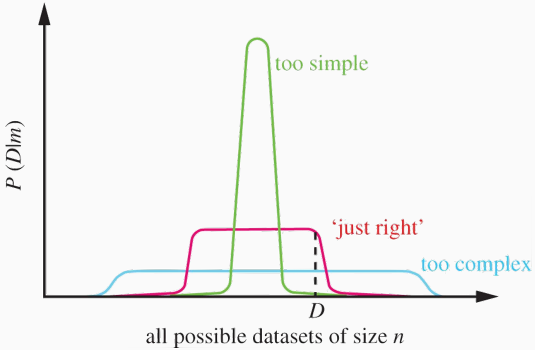
`harmonic` code

Extra slides

Occam's razor

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

- In Bayesian formalism models specified as probability distributions over datasets.
- Each model has limited "probability budget".
- Complex models can represent a wide range of datasets well, but spreads predictive probability.
- In doing so, model evidence of complex models penalised if complexity not required.



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 priors

  e.g. regularize by imposing sparsity in dictionary

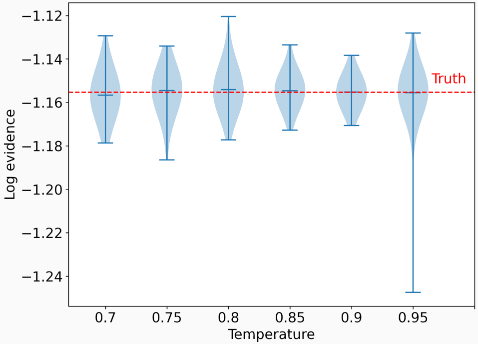
- Data-informed priors

  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)

Robustness to choice of temperature hyperparameter



Marginal likelihood estimates for Rosenbrock example with varying temperature (Polanska *et al.* 2024).

- Marginal likelihood estimates robust to choice of temperature.
- Temperature of $T = 0.90$ suitable for most cases.