Bayesian model comparison with data-driven priors

Learned proximal nested sampling

Jason D. McEwen
www.jasonmcewen.org

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

Workshop on Recent Advances in Learned Regularisation, Maths 4 Deep Learning (M4DL), 2025

Goal

Bayesian parameter estimation and model selection
for inverse imaging problems...

Goal

Bayesian parameter estimation and model selection
for inverse imaging problems...

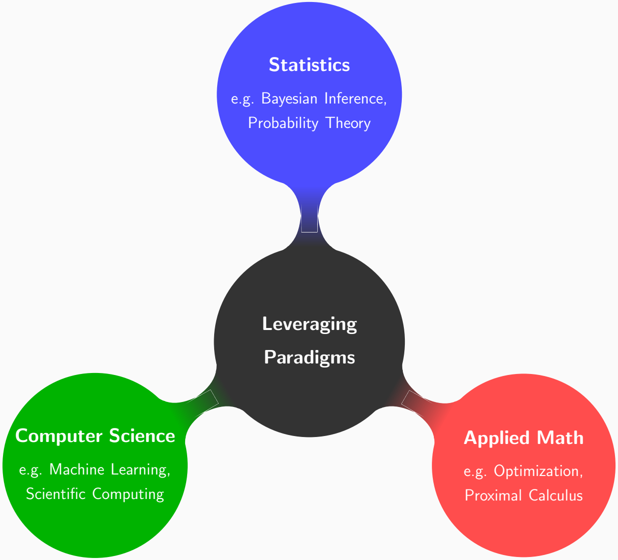
with data-driven priors (learned regularisation).

Questions that can be addressed by model selection

- What is the best forward model?
- (How set regularisation strength?)
- What is the best learned data-driven prior (regulariser)?
- What is the best training data-set?
- ...

Address these questions using the data itself... not by, *e.g.*, cross-validation.

Leveraging paradigms



Outline

1. Nested sampling
2. Proximal nested sampling
3. Learned data-driven priors

Nested sampling

Bayesian inference: parameter estimation

First, let's set the notation (and colour code)...

Bayesian inference: parameter estimation

First, let's set the notation (and colour code)...

Bayes' theorem

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

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

SciAI UCL
Jason McEwen

Bayesian inference: parameter estimation

First, let's set the notation (and colour code)...

Bayes' theorem

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

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

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:

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

Bayesian inference: model selection

By Bayes' theorem for model M_j:

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

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

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

Bayesian inference: model selection

By Bayes' theorem for model M_j:

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

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

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

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

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

Jason McEwen

Bayesian inference: model selection

By Bayes' theorem for model M_j:

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

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

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

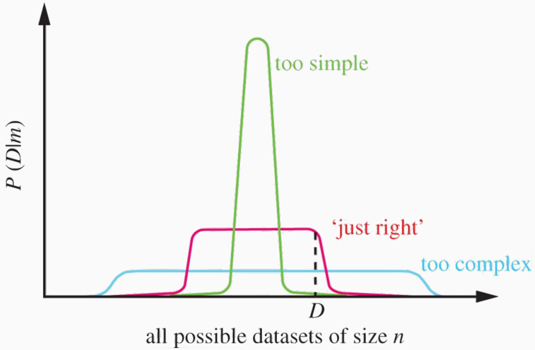
Must compute the **Bayesian model evidence** or **marginal likelihood** given by the normalising constant

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

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

The marginal likelihood **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, marginal likelihood 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. objective Bayes, 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)

Nested sampling: reparameterising the likelihood

Nested sampling is ingenious approach to evaluate the marginal likelihood (Skilling 2006).

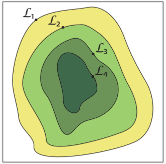
Consider $\Omega_{L^*} = \{x | \mathcal{L}(x) \geq L^*\}$, which groups the parameter space Ω into a series of **nested subspaces**.

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

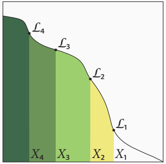
$$z = \int_0^1 \mathcal{L}(\xi) \mathrm{d}\xi,$$

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



Feroz *et al.* (2013)

Nested subspaces



Feroz *et al.* (2013)

Reparameterised likelihood

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: 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 **marginal likelihood** can be computed numerically using standard quadrature:

$$z = \sum_{i=1}^{N} L_i w_i \ ,$$

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 **marginal likelihood** can be computed numerically using standard quadrature:

$$z = \sum_{i=1}^{N} L_i w_i \ ,$$

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

$$p_i = \frac{L_i w_i}{z} \ .$$

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.

Crux: **sample from the prior, subject to the likelihood level-set constraint**, *i.e.* sample from the prior $\pi(\boldsymbol{x})$, such that $\mathcal{L}(\boldsymbol{x}) > L^*$.

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

Proximal nested sampling

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.

$\Rightarrow$ Proximal nested sampling (Cai, McEwen & Pereyra 2022; arXiv:2106.03646)

Constrained sampling formulation

Consider case where likelihood and prior of the form

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

Likelihood | Prior

where g is convex lower semicontinuous function (prior need not be log-convex).

Constrained sampling formulation

Consider case where likelihood and prior of the form

$$\underbrace{\mathcal{L}(x) \propto \exp(-g(x))}_{\text{Likelihood}}, \qquad \underbrace{\pi(x) \propto \exp(-f(x))}_{\text{Prior}},$$

where g 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

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

Likelihood   Prior

where g 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.

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

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**.

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.

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.

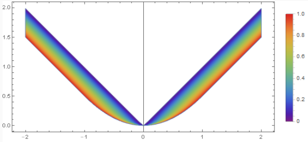
Need gradients so **not directly applicable** $\Rightarrow$ adopt Morea-Yosida approximation.

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**:

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



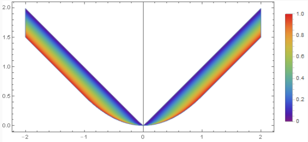
M-Y envelope of $|x|$ for varying λ.

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 $f^\lambda(x)$:

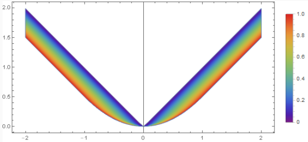
1. As $\lambda \rightarrow 0, f^\lambda(x) \rightarrow f(x)$
2. $\nabla f^\lambda(x) = (x - \text{prox}_f^\lambda(x))/\lambda$

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):

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

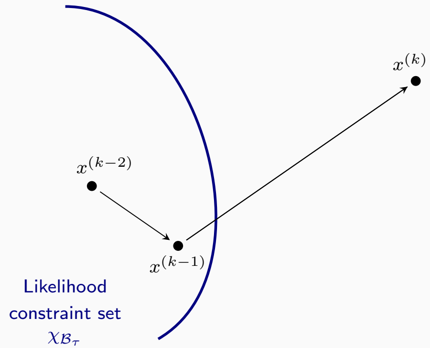
Jason McEwen

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.

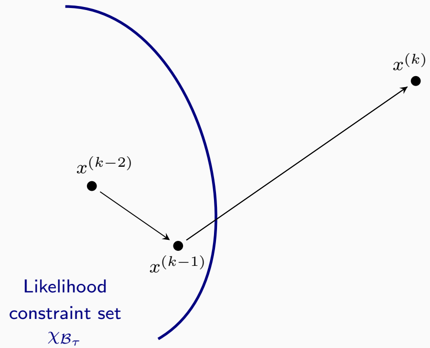


Proximal nested sampling intuition

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.

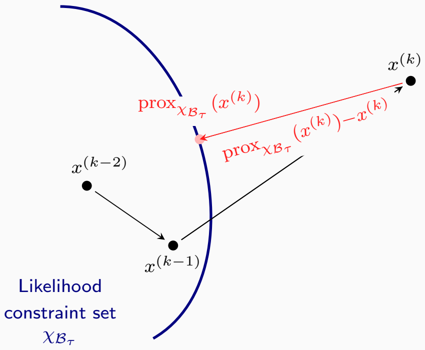


Proximal nested sampling intuition

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

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

1. **$x^{(k)}$ is already in $\mathcal{B}_\tau$**: term $\left[x^{(k)} - \operatorname{prox}^{\lambda}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right]$ disappears and recover usual Langevin MCMC.
2. **$x^{(k)}$ is not in $\mathcal{B}_\tau$**: a step is also taken in the direction $-\left[x^{(k)} - \operatorname{prox}^{\lambda}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right]$, 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

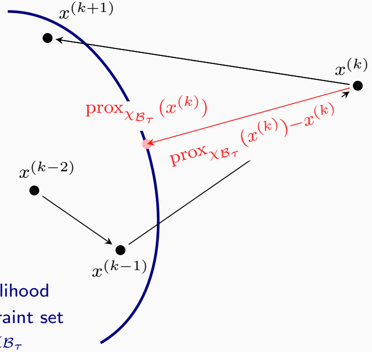
Proximal nested sampling intuition

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. $-\log \pi(x) = \|\Psi^\dagger x\|_1$), can also make Moreau-Yosida approximation $f^\lambda(x)$ and leverage prox to compute gradient ∇f^λ:

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

SciAI UCL

Explicit forms of proximal nested sampling

Must compute the proximity operators.

Explicit forms of proximal nested sampling

Must compute the proximity operators.

Consider common imaging problem as example:

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

Prior

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

Explicit forms of proximal nested sampling

Must compute the proximity operators.

Consider common imaging problem as example:

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

Likelihood

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

Prior

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

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

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

Computing proximal operator for likelihood

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

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

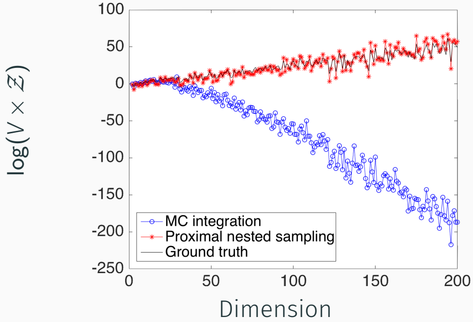
Solve iteratively by primal dual method:

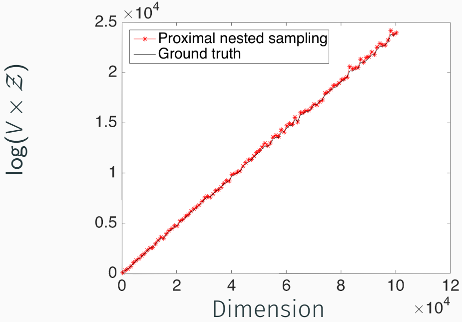
1. $z^{(i+1)} = z^{(i)} + \delta_1 \Phi \bar{x}^{(i)} - \mathrm{prox}_{\chi_{\mathcal{B}'_{\tau'}}}(z^{(i)} + \delta_1 \Phi \bar{x}^{(i)})$,

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

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

Validation on Gaussian problem





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

Also validated in 10^6 **dimensions**.

Truth: 2.3850×10^5 Proximal nested sampling: $(2.3851 \pm 0.0002) \times 10^5$

Denoising wavelet dictionary experiment

Clean image

Noisy image

$\Psi = I$

$\Psi = \text{DB2}$

$\Psi = \text{DB8}$

Denoising wavelet dictionary experiment

Prior	$\log z$	RMSE (Requires ground truth)
$\Psi = I$	-6.54×10^4	41.07
$\Psi = \text{DB2}$	-3.06×10^4	14.29
$\Psi = \text{DB8}$	-3.09×10^4	14.51

Evidence computed by proximal nested sampling correctly compares wavelet dictionaries.

Learned data-driven priors

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 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

Tweedie's formula (Robins 1956)

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

Tweedie's formula gives the posterior expectation of z given x as

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

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

Tweedie's formula

Tweedie's formula (Robins 1956)

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

Tweedie's formula gives the posterior expectation of z given x as

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

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

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

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.

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

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

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.

$\Rightarrow$ 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

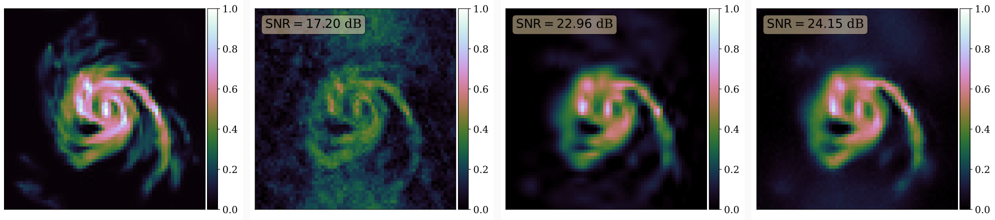
Substituting the denoiser $\nabla \log \pi_\epsilon(x) = \epsilon^{-1}(D_\epsilon(x) - x)$ into the proximal nested sampling Markov chain update:

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

Hand-crafted vs data-driven priors

Consider simple radio interferometric imaging inverse problem with:

- ▷ **hand-crafted prior** based on sparsity-promoting wavelet representation;
- ▷ **data-driven prior** based on a deep convolutional neural network (Ryu et al. 2019).



Ground truth | Backprojected (17.2dB) | Hand-crafted prior (23.0dB) | Data-driven prior (24.2dB)

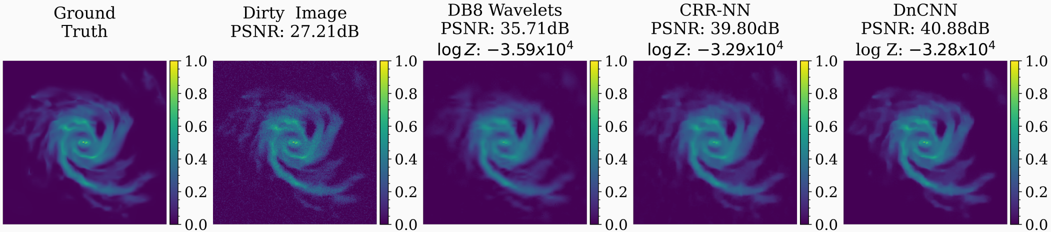
Which model best?

- ▷ SNR ⇒ **data-driven priors best** but require ground-truth;
- ▷ Bayesian evidence ⇒ **data-driven priors best** (no ground-truth knowledge).

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 deep neural networks (Goujon et al. 2023, Ryu et al. 2019).



Which model best?

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

Summary

Summary

- **Proximal nested sampling** (arXiv:2106.03646) framework scales to **high-dimensions**, opening up Bayesian model comparison for, e.g., imaging problems.
- Constrained to **log-convex likelihoods**, which are ubiquitous in imaging sciences.
- Prior not constrained to be log-convex so can be a deep neural network.
- **Learned proximal nested sampling** (arXiv:2307.00056) approach to support **data-driven priors**.
- **Future work**:
  - More extensive experiments to showcase use
  - Remove convexity constraint
  - More expressive data-driven priors (*e.g.* denoising diffusion models)

Extra Slides

Alternatives to marginal likelihood

- **Posterior predictive checks**
  - ✔ Fine for model consistency checks
  - ✖ Not suitable for model comparison
    - Does not guarantee Bayesian consistency
    - Does not penalise model complexity

- **Bayesian leave one out (LOO) cross validation**
  - ✔ Fine for validation
  - ✖ Not suitable for model comparison
    - Does not guarantee Bayesian consistency
    - Does not penalise model complexity

- **Bayesian model complexity and dimensionality**
  - ✔ Only weakly dependent on prior through posterior

- **Bayesian suspicious** for testing for tensions between datasets
  - ✔ Only weakly dependent on prior through posterior

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

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