Proximal nested sampling

with data-driven priors for inverse imaging

Jason D. McEwen

www.jasonmcewen.org

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

In collaboration with:
Henry Aldridge · Marcelo Pereyra · Matthew Price · Tobias Liaudat · Xiaohao Cai

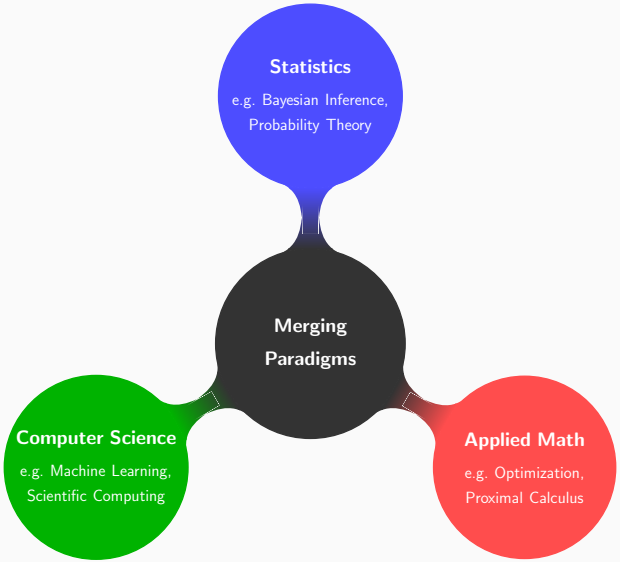
SIAM Conference on Imaging Science, Atlanta, 2024

Goal

Bayesian parameter estimation and model selection

for inverse imaging problems.

Merging paradigms



Outline

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

Nested sampling

Bayesian inference: parameter estimation

Bayes' theorem

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

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

Bayesian inference: parameter estimation

Bayes' theorem

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

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

Bayesian inference: parameter estimation

Bayes' theorem

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

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

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:

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

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

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

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

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

Bayesian inference: model selection

By Bayes' theorem for model M_j:

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

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

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

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

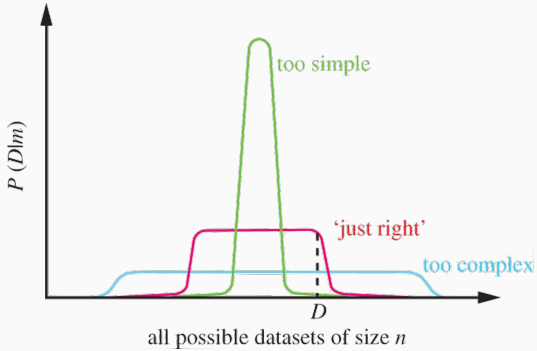
$$z = p(\boldsymbol{y} \,|\, M) = \int \mathrm{d}\theta \, \mathcal{L}(\theta) \, \pi(\theta) \,.$$

→ 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.

- 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 widely.
- 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)

Nested sampling: reparameterising the likelihood

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

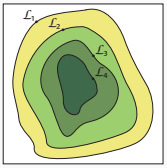
Consider $\Omega_{L^*} = \{x | \mathcal{L}(x) \geq L^*\}$, which groups the parameter space Ω into a series of **nested subspaces**.

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

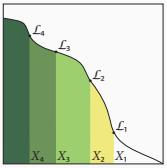
$$\mathcal{Z} = \int_0^1 \mathcal{L}(\xi) \mathrm{d}\xi,$$

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



Feroz *et al.* (2013)

Nested subspaces



Feroz *et al.* (2013)

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:

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

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:

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

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

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

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.

Proximal calculus

Motivating example: high-dimensional inverse imaging problems

Classical high-dimensional imaging problems often consider Gaussian likelihood and sparsity-promoting prior (e.g. in wavelet representation Ψ):

$$p(\boldsymbol{y} \,|\, \boldsymbol{x}) \propto \exp\left(-\|\boldsymbol{y} - \boldsymbol{\Phi x}\|_2^2 / (2\sigma^2)\right)$$

Likelihood

$$p(\boldsymbol{x}) \propto \exp\left(-\|\boldsymbol{\Psi}^\dagger \boldsymbol{x}\|_1\right)$$

Prior

Often compute **MAP estimator** (variational regularisation):

$$\arg\max_{x} \log p(x \,|\, y) = \arg\min_{x} \Big[\underbrace{\|y - \Phi x\|_2^2}_{\text{Data fidelity}} + \underbrace{\lambda \|\Psi^\dagger x\|_1}_{\text{Regulariser}} \Big]$$

$\Rightarrow$ Often solved by **convex optimisation** algorithms (e.g. proximal splitting algorithms).

Motivating example: high-dimensional inverse imaging problems

Classical high-dimensional imaging problems often consider Gaussian likelihood and sparsity-promoting prior (e.g. in wavelet representation Ψ):

$$p(y \mid x) \propto \exp\left(-\|y - \Phi x\|_2^2 / (2\sigma^2)\right) \qquad p(x) \propto \exp\left(-\|\Psi^\dagger x\|_1\right)$$

Likelihood Prior

Often compute **MAP estimator** (variational regularisation):

$$\arg\max_x \log p(x \mid y) = \arg\min_x \left[\underbrace{\|y - \Phi x\|_2^2}_{\text{Data fidelity}} + \underbrace{\lambda \|\Psi^\dagger x\|_1}_{\text{Regulariser}} \right]$$

$\Rightarrow$ Often solved by **convex optimisation** algorithms (e.g. **proximal** splitting algorithms).

Motivating example: high-dimensional inverse imaging problems

Classical high-dimensional imaging problems often consider Gaussian likelihood and sparsity-promoting prior (e.g. in wavelet representation Ψ):

$$p(y \mid x) \propto \exp\left(-\|y - \Phi x\|_2^2/(2\sigma^2)\right) \qquad p(x) \propto \exp\left(-\|\Psi^\dagger x\|_1\right)$$

Likelihood Prior

Often compute **MAP estimator** (variational regularisation):

$$\arg\max_x \log p(x \mid y) = \arg\min_x \Big[\underbrace{\|y - \Phi x\|_2^2}_{\text{Data fidelity}} + \underbrace{\lambda \|\Psi^\dagger x\|_1}_{\text{Regulariser}} \Big]$$

$\Rightarrow$ Often solved by **convex optimisation** algorithms (e.g. proximal splitting algorithms).

Proximity operator

Proximity operator

The **prox** of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ is given by

$$\text{prox}_f^\lambda(x) = \arg\min_u \left[f(u) + \|u - x\|^2 / 2\lambda \right]$$

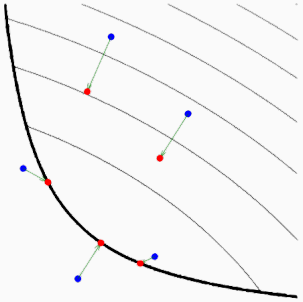


Illustration of prox (Parikh & Boyd 2013)

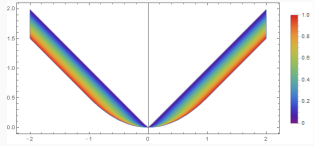
- Thin black lines level curves of convex function.
- Thick black line indicates domain boundary of function.
- Evaluating $prox_f$ at blue points $\mapsto$ red points.

Moreau-Yosida approximation

Morea-Yosida (M-Y) approximation

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

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



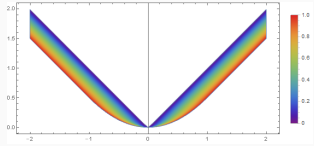
M-Y envelope of $|x|$ for varying λ.

Moreau-Yosida approximation

Morea-Yosida (M-Y) approximation

The Morea-Yosida approximation of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ 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 $f^\lambda(x)$:

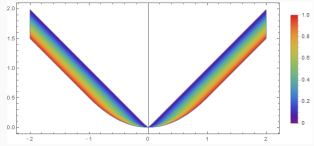
1. As $\lambda \to 0, f^\lambda(x) \to f(x)$
2. $\nabla f^\lambda(x) = (x - \text{prox}_f^\lambda(x))/\lambda$

Moreau-Yosida approximation

Morea-Yosida (M-Y) approximation

The Morea-Yosida approximation of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ 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 $\lambda \to 0, f^\lambda(x) \to f(x)$
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

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

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

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**.

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

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

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

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

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

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:

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

Proximal nested sampling intuition

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

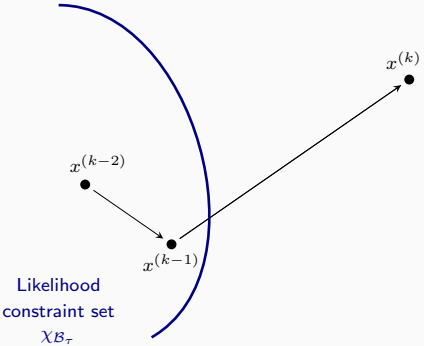
$$x^{(k+1)} = x^{(k)} + \frac{\delta}{2}\nabla \log \pi(x^{(k)}) - \frac{\delta}{2\lambda}\left[x^{(k)} - \text{prox}_{\chi \mathcal{B}_\tau}(x^{(k)})\right] + \sqrt{\delta} w^{(k+1)}.$$

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.



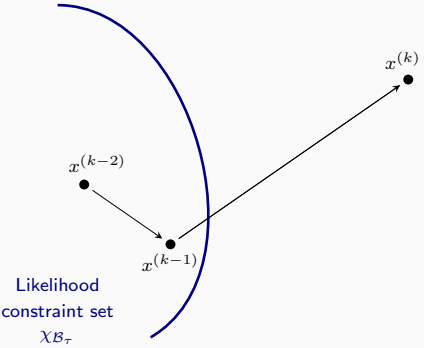
Proximal nested sampling intuition

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.



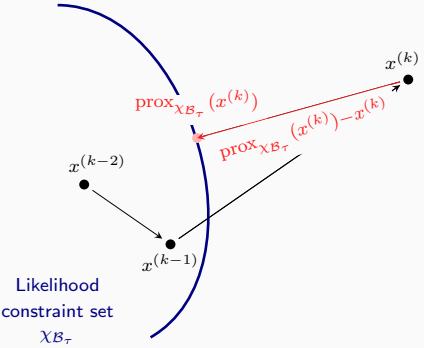
Proximal nested sampling intuition

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.



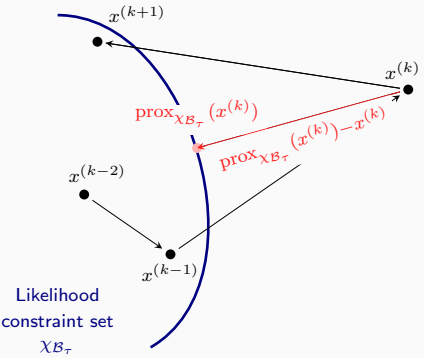
Proximal nested sampling intuition

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

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

Proximal nested sampling

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

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

Explicit forms of proximal nested sampling

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

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

Explicit forms of proximal nested sampling

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

Consider common imaging problem as example:

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

Likelihood

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

Prior

Straightforward when Φ is identity.

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

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

SciAI

Computing proximal operator for likelihood

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

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

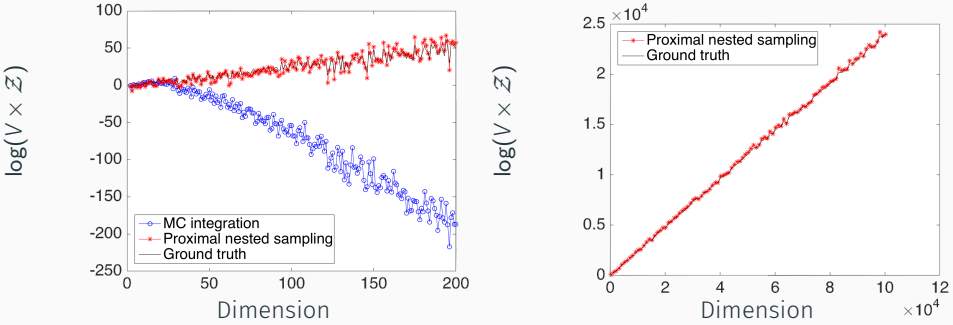
Solve iteratively by primal dual method:

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. $\bar{x}^{(i+1)} = x^{(i+1)} + \delta_3(x^{(i+1)} - x^{(i)})$

Validation on Gaussian problem



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

Also validated in 10^6 dimensions.

Denoising wavelet dictionary experiment

Clean image

Noisy image

$\Psi = I$

$\Psi = \text{DB2}$

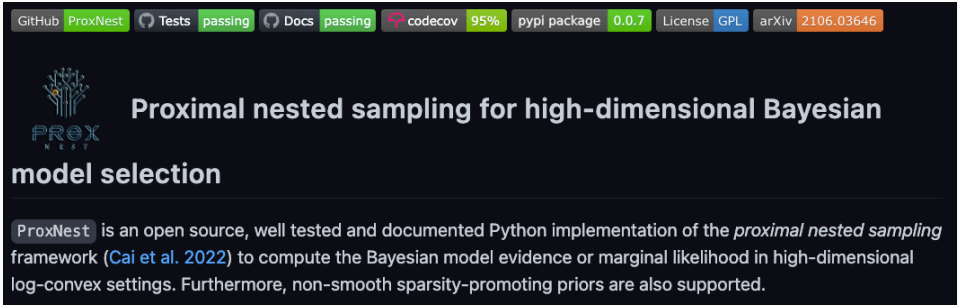
$\Psi = \text{DB8}$

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.

Proxnest code



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

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

Learned deep 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).

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).

Score matching and **denoising diffusion models** achieve state-of-the-art performance in deep generative modelling $\Rightarrow$ denoising closely related to 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

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

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

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

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

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 $\nabla \log p(\boldsymbol{z})$.

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.

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

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

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:

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

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

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

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

Proximal nested sampling with learned data-driven priors

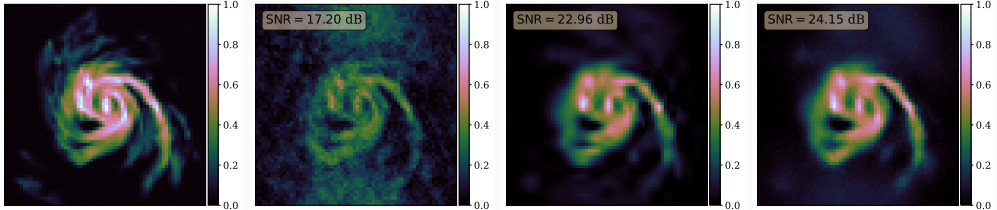
Substituting the denoiser $\nabla \log \pi_\epsilon(x) = \epsilon^{-1}(D_\epsilon(x) - x)$ into the proximal nested sampling Markov chain update:

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

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 | Dirty | Hand-crafted prior | Data-driven prior (DnCNN)

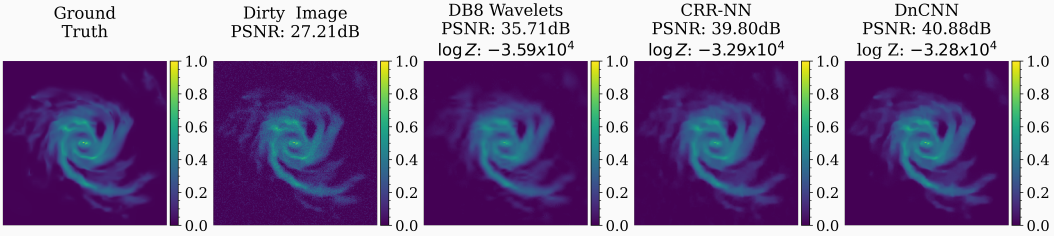
Which model best?

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



Which model best?

- ▷ SNR ⇒ **data-driven priors best** but require ground-truth;
- ▷ Bayesian evidence ⇒ **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.
- Recently developed learned proximal nested sampling (arXiv:2307.00056) approach to support data-driven priors in an empirical Bayes setting.