Proximal nested sampling

for high-dimensional Bayesian model selection

Jason D. McEwen

www.jasonmcewen.org

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

Frontiers of Nested Sampling, Maximum Entropy Workshop, 2023

Bayesian inference: setting the notation

Bayes' theorem

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

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

Bayesian inference: setting the notation

Bayes' theorem

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

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

For **model selection**, must compute the **Bayesian model evidence** or **marginal likelihood** given by the normalising constant

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

$\rightarrow$ Challenging computational problem.

Nested sampling: reparameterising the likelihood

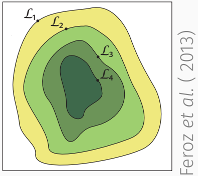
Nested sampling: ingenious approach to efficiently evaluate the evidence (Skilling 2006).

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

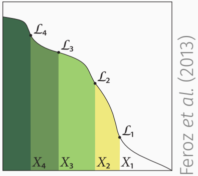
Define the prior volume ξ within Ω_{L^*} by $\xi(L^*) = \int_{\Omega_{L^*}} \pi(x)\mathrm{d}x$.

Evidence can then be rewritten as

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



Nested subspaces



Reparameterised likelihood

Nested sampling: reparameterising the likelihood

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

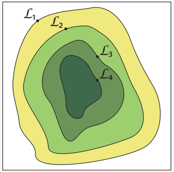
Group the parameter space Ω into a series of **nested subspaces**:
$\Omega_{L^*} = \{x \,|\, \mathcal{L}(x) \geq L^*\}$.

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

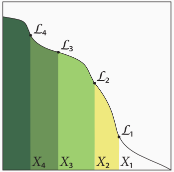
$$z = \int_0^1 \mathcal{L}(\xi) \mathrm{d}\xi.$$

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



Feroz *et al.* (2013)

Nested subspaces



Feroz *et al.* (2013)

Reparameterised likelihood

Nested sampling: constrained sampling

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.

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

$\Rightarrow$ **Exploit structure** of common high-dimensional problems.

Nested sampling: constrained sampling

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.

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

$\Rightarrow$ Exploit structure of common high-dimensional problems.

Nested sampling: constrained sampling

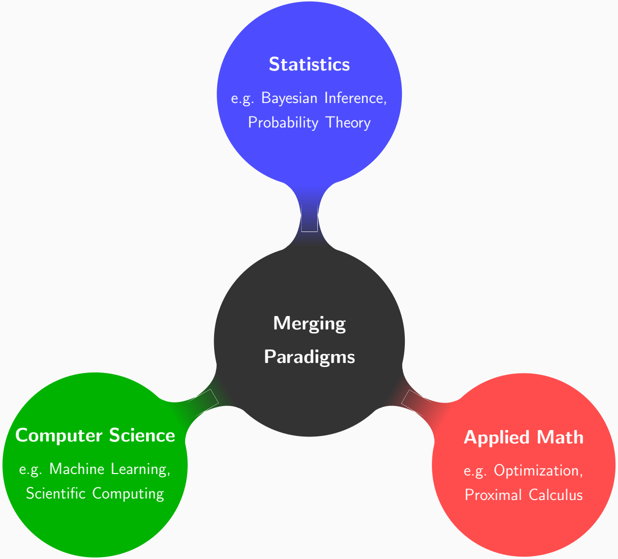
Nested sampling (Skilling 2006)

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.

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

$\Rightarrow$ Exploit structure of common high-dimensional problems.

Merging paradigms



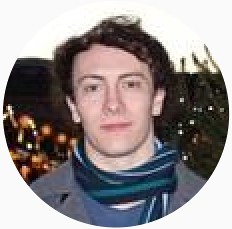
Aside: Learned harmonic mean estimation of the marginal likelihood

- ▷ Learned harmonic mean estimator
  (McEwen *et al.*; arXiv:2111.12720)
- ▷ Bayesian model comparison for simulation-based inference
  (Spurio Mancini *et al.*; arXiv:2207.04037)
- ▷ Learned harmonic mean estimation with normalizing flows [MaxEnt poster!]
  (Polanska *et al.*; arXiv:2307.00048)

Agnostic to sampling strategy ($\rightarrow$ HMC, NUTS).

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

Alicja Polanska



Matt Price

Alessio Spurio Mancini

Outline

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

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):

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

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 $\boldsymbol{\Psi}$):

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

Likelihood

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

Prior

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

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

Data fidelity — Regulariser

$\Rightarrow$ 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):

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

Likelihood Prior

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

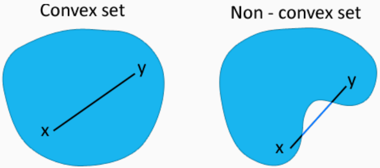
$$\arg\max_x \log p(x \mid y) = \arg\min_x \left[\underbrace{\|y - \mathbf{\Phi} x\|_2^2}_{\text{Data fidelity}} + \underbrace{\lambda \|\Psi^\dagger x\|_1}_{\text{Regulariser}} \right]$$

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

Convex set

$\mathcal{C}$ is a **convex set** if for any $x_1, x_2 \in \mathcal{C}$ and $\alpha \in (0, 1)$ we have

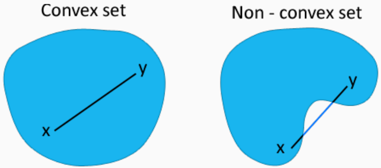
$$\alpha x_1 + (1 - \alpha) x_2 \in \mathcal{C}.$$



Convex set

$\mathcal{C}$ is a **convex set** if for any $x_1, x_2 \in \mathcal{C}$ and $\alpha \in (0,1)$ we have

$$\alpha x_1 + (1-\alpha) x_2 \in \mathcal{C}.$$

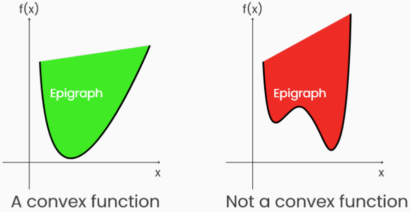


Convex function

The **epigraph** of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is defined by

$$\text{epi}(f) = \{(x, \gamma) \in \mathbb{R}^n \times \mathbb{R} \,|\, f(x) \leq \gamma\}.$$

f is a **convex function** if and only if its **epigraph is convex**.



Sub-differentials

Subdifferential

The **subdifferential** of a convex function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ at $x_0 \in \mathbb{R}^n$ is the set

$$\partial f(x_0) = \{c \,|\, f(x) \geq f(x_0) + c^\mathsf{T}(x - x_0)\}.$$

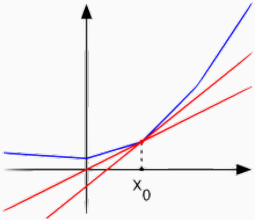


Illustration of sub-gradients

Sub-differentials

Subdifferential

The **subdifferential** of a convex function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ at $x_0 \in \mathbb{R}^n$ is the set

$$\partial f(x_0) = \{c \,|\, f(x) \geq f(x_0) + c^\top (x - x_0)\}.$$

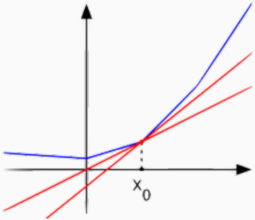


Illustration of sub-gradients

- Each $c \in \partial f(x_0)$ called a **subgradient**.
- If f is differentiable at x_0, then

$$\partial f(x_0) = \{\nabla f(x_0)\}.$$

- Subdifferentials useful for optimising non-differentiable convex functions:

$$0 \in \partial f(x^\star) \Leftrightarrow x^\star \text{ minimises } f.$$

Proximity operator

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

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

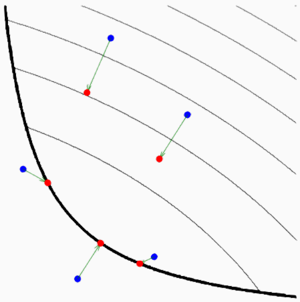


Illustration of prox (Parikh & Boyd 2013)

- 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.

Proximity operator as generalised projection operator

Recall proximity operator:

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

Generalisation of **projection operator**:

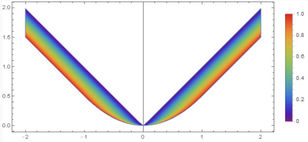
$$\Pi_{\mathcal{C}}(x) = \arg\min_u \Big[\underbrace{\imath_{\mathcal{C}}(u)}_{\text{Indicator}} + \|u - x\|^2 / 2 \Big] ,$$

where $\imath_{\mathcal{C}}(u) = \infty$ if $u \notin \mathcal{C}$ and zero otherwise.

Morea-Yosida approximation

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

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



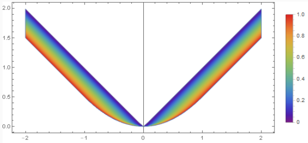
Moreau-Yosida envelope of $|x|$ for varying λ.

Moreau-Yosida approximation

Morea-Yosida approximation

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

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



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

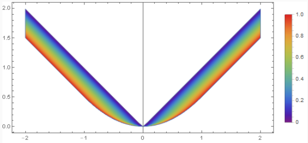
Important **properties** of $f^\lambda(x)$:

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

Morea-Yosida approximation

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

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



Moreau-Yosida 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.

$\Rightarrow$ 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

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

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:

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

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:

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

Then let $\pi_{L^*}(x) = \pi(x)\iota_{L^*}(x)$ represent the 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 $p(x)$ differentiable can adopt **Langevin dynamics**.

MCMC sampling with Langevin dynamics

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

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

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

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

where w is Brownian motion.

MCMC sampling with Langevin dynamics

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

If target distribution $p(x)$ 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.

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 (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):

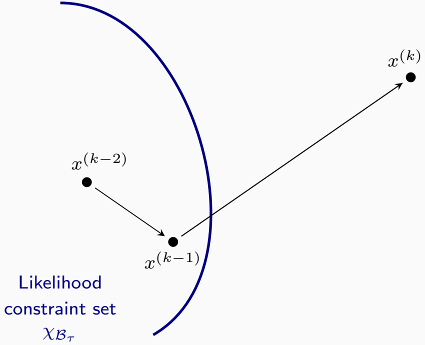
$$x^{(k+1)} = x^{(k)} + \frac{\delta}{2} \nabla \log \pi(x^{(k)}) - \frac{\delta}{2\lambda} \left[x^{(k)} - \operatorname{prox}_{\chi_{\mathcal{B}_\tau}}(x^{(k)}) \right] + \sqrt{\delta} w^{(k+1)}.$$

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.



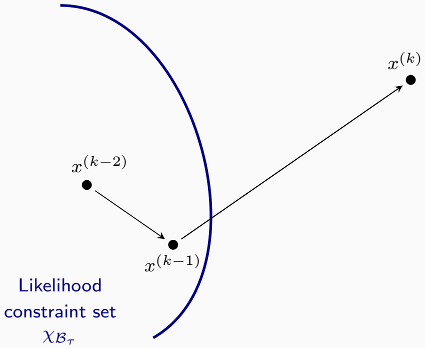
Proximal nested sampling intuition

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.



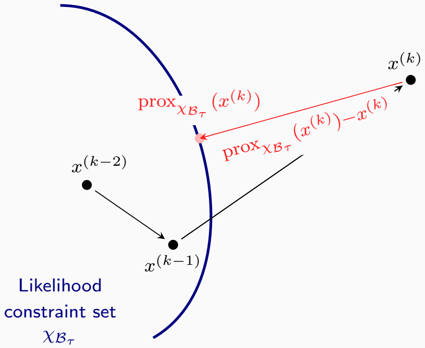
Proximal nested sampling intuition

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.



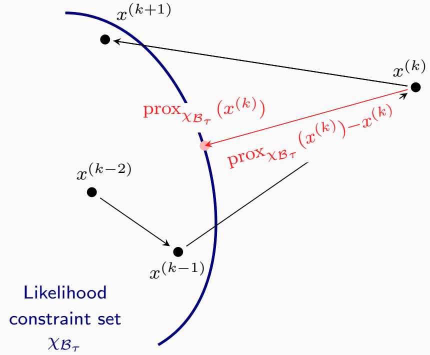
Proximal nested sampling intuition

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.



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

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

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:

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

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.

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

Computing proximal operator for likelihood

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

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

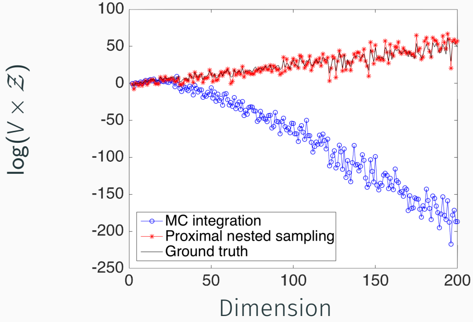
Solve iteratively by primal dual method:

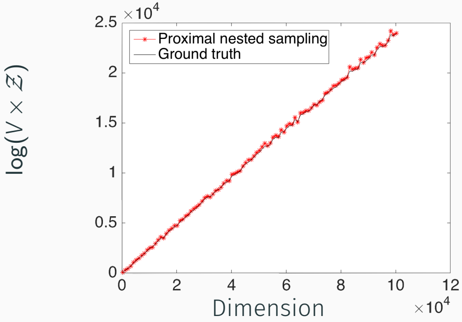
1. $z^{(i+1)} = z^{(i)} + \delta_1 \Phi \bar{x}^{(i)} - \mathrm{prox}_{\chi_{\mathcal{B}'_{\tau'}}}(z^{(i)} + \delta_1 \Phi \bar{x}^{(i)}),$

   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).

Dimension 10^6

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

Denoising wavelet dictionary experiment

Clean image

Noisy image

$\Psi = I$

$\Psi = \mathrm{DB2}$

$\Psi = \mathrm{DB8}$

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

Proximal nested sampling for high-dimensional Bayesian model selection

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

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

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

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.

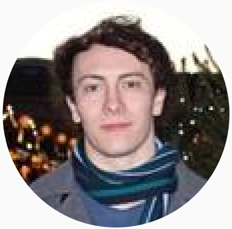
Aim: **integrate learned deep data-driven priors** into proximal nested sampling.

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


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

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

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.

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

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

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

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

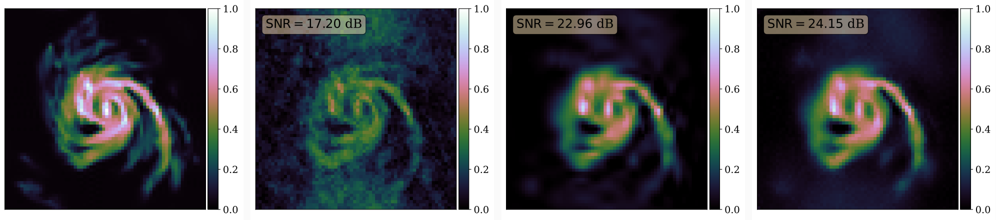
Substituting the denoiser $\nabla \log \pi_\epsilon(x) = \epsilon^{-1}(D_\epsilon(x) - x)$ into the proximal nested sampling Markov chain update:

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

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).

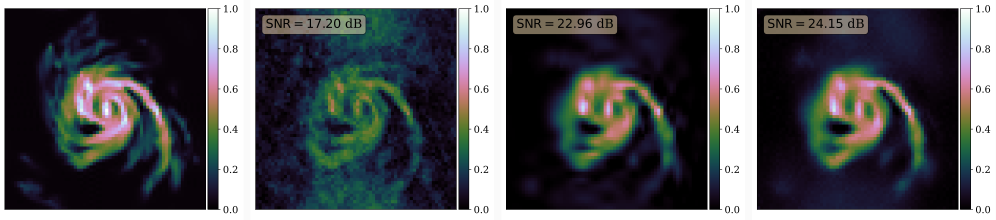


Ground truth | Dirty | Hand-crafted prior | Data-driven prior

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

Which model best?

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

Summary

Summary

- **Proximal nested sampling** 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** approach to support data-driven priors in an empirical Bayes setting.

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

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