Fast Mixing MCMC without Gradients

Robert Scheichl

Institute for Mathematics and
Interdisciplinary Center for Scientific Computing (IWR),
Heidelberg University

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

jointly with **R Kutri** (Heidelberg)

as well as T Dodwell (digiLab/Exeter), C Fox (Otago), M Lykkegaard, G Mingas

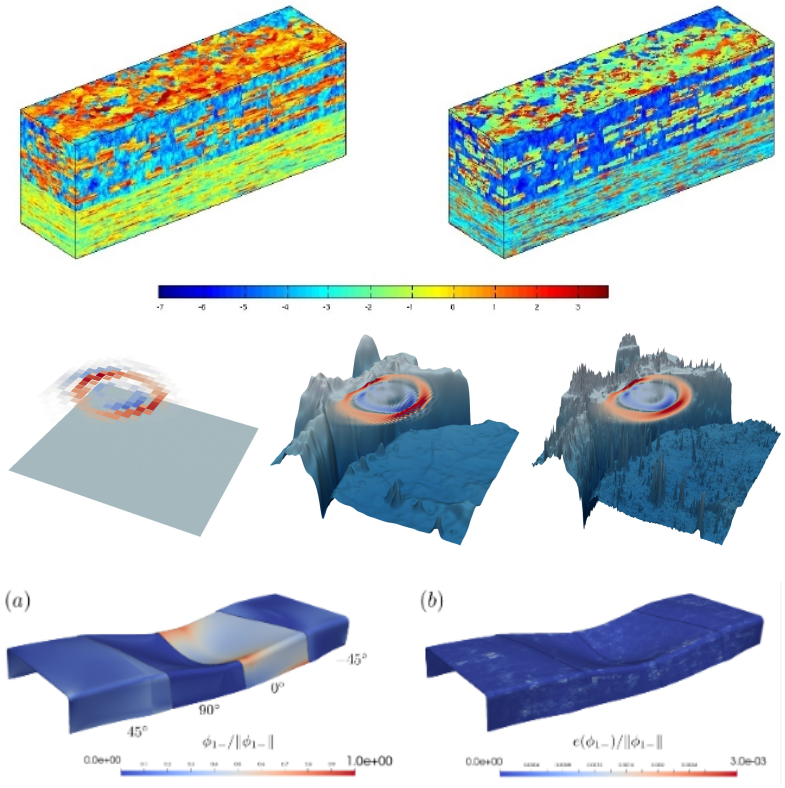
ICERM Workshop on Bayesian Inverse Problems and UQ

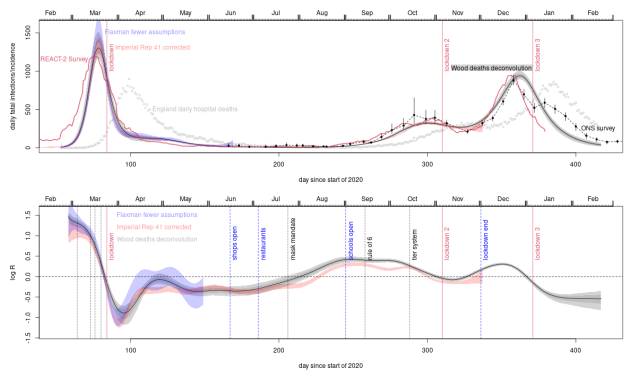
Brown University, Providence, 5th March 2026

Large-Scale Bayesian Inference

From Models to Decisions . . .

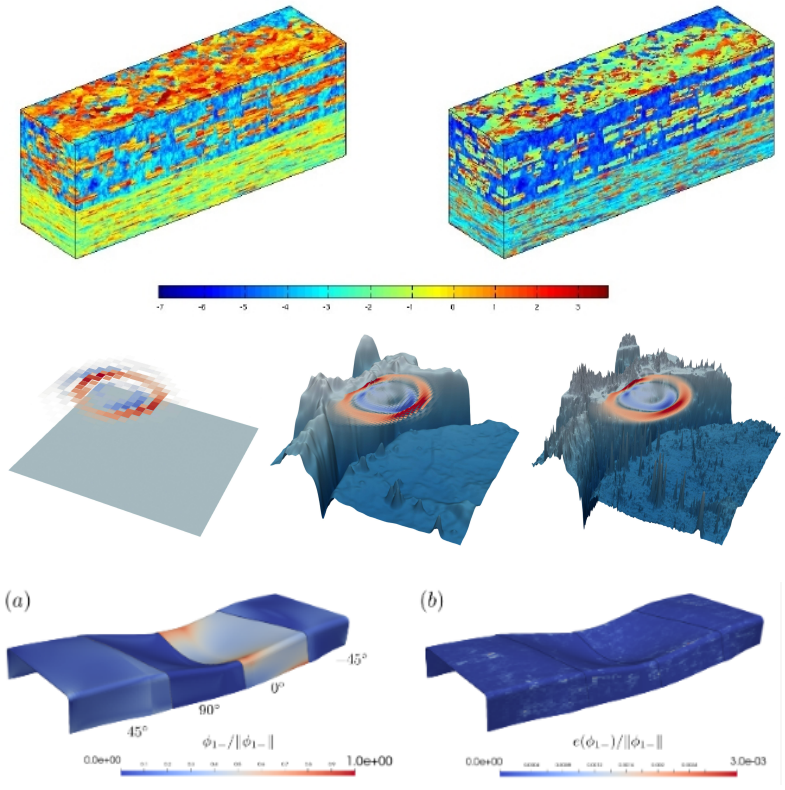
- High level of sophistication in high performance simulations of complex physical problems!
- Huge explosion of ‘data-driven’ methods!
- What do decision makers really care about?
  - The perfect model? → **No!**
  - Full field output → **Often Not!**
  - Understanding of what happens on average → **Often not!**

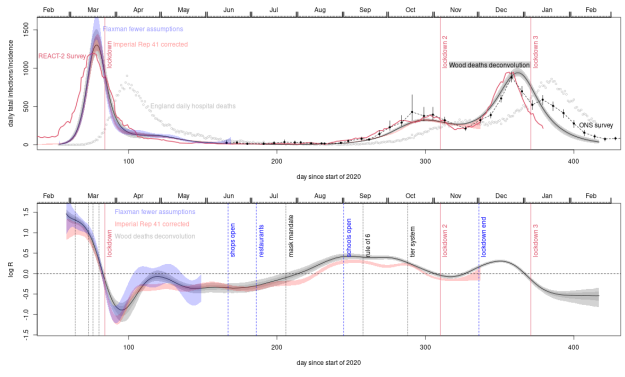




From Models to Decisions . . .

- High level of sophistication in high performance simulations of complex physical problems!
- Huge explosion of ‘data-driven’ methods!
- What do decision makers really care about?
  - The perfect model? → **No!**
  - Full field output → **Often Not!**
  - Understanding of what happens on average → **Often not!**
- **True goal:** Models & data to sing together!
- Predictions to revert to our scientific knowledge of physics in the absence of data - with appropriate measure of uncertainty.





The Challenge

Many excellent **scalable** PDE software packages for complex problems:

- **Commercial:** Ansys, Simulia, Schlumberger, ...
- **Open Source:** OpenFOAM, OPM, FEniCS, FreeFEM++, ...
- **HPC** (open source): deal.II, NGSolve, DUNE, Firedrake, ...

But In real applications

- **Parameters**, BCs, source term, geometry often uncertain / unknown
- **Variability** on many scales $\Rightarrow$ High-dimensional parametrisation
- Need to calibrate model using **Data** (functionals of the solution)

The Challenge

Many excellent **scalable** PDE software packages for complex problems:

- **Commercial:** Ansys, Simulia, Schlumberger, ...
- **Open Source:** OpenFOAM, OPM, FEniCS, FreeFEM++, ...
- **HPC** (open source): deal.II, NGSolve, DUNE, Firedrake, ...

But In real applications

- **Parameters**, BCs, source term, geometry often uncertain / unknown
- **Variability** on many scales $\Rightarrow$ High-dimensional parametrisation
- Need to calibrate model using **Data** (functionals of the solution)

Many excellent **scalable** statistical inference methods (for low dimens.)

(as well as deterministic parameter identification methods)

The **challenge** lies in finding **scalable algorithms** for inference (with UQ) for highly complex (PDE) models when the parameter and/or data space are high-dimensional and especially when the dependence is not smooth.

Bayesian Inverse Problems

- Given (limited) observations of a system

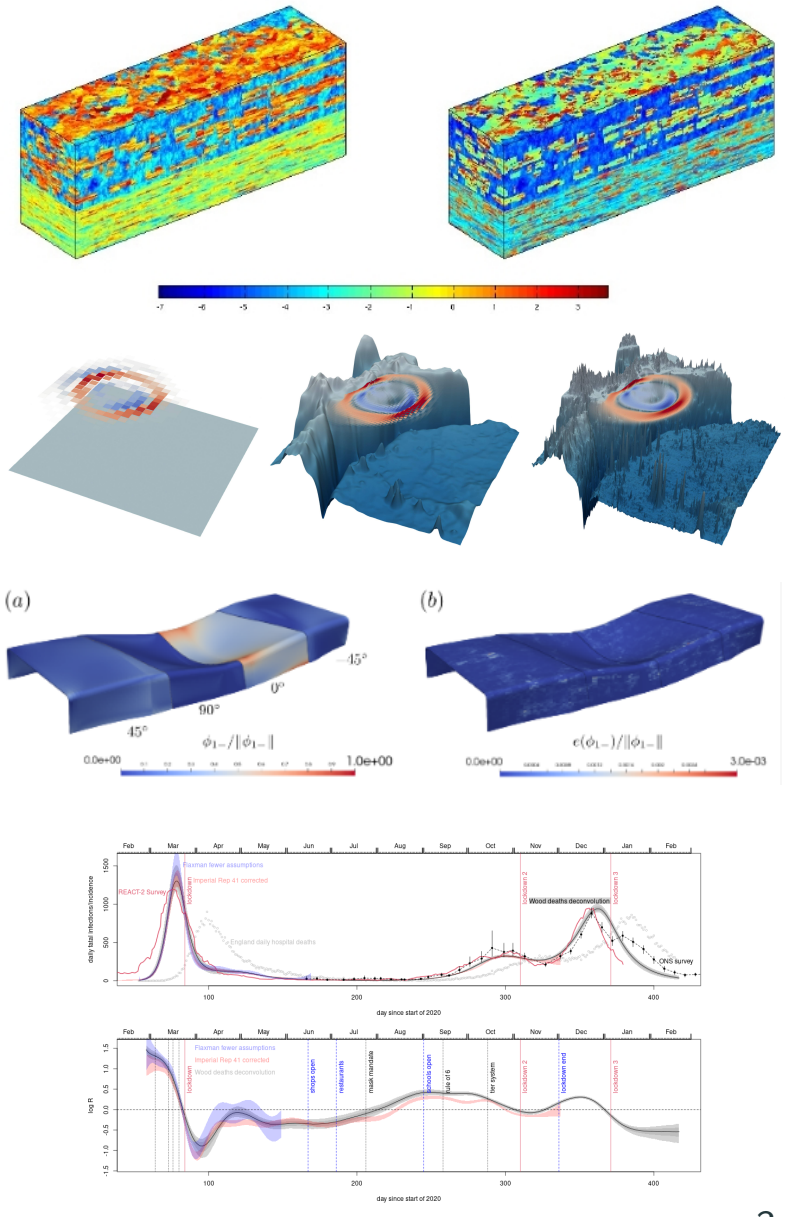
$$\mathbf{d} \in \mathbb{R}^m$$

- A (mathematical) model $\mathcal{F}(\theta) : \mathbb{R}^d \to \mathbb{R}^m$ which predicts our data given parameters θ.
- Connect model and data

$$\boldsymbol{\eta} = \mathbf{d} - \mathcal{F}(\theta) \sim \mathcal{N}(\mathbf{0}, \sigma_\eta^2 \mathbb{I})$$

- Assume some *prior* $\pi(\theta)$ on the parameters
- Wish to find distribution $\pi(\theta|\mathbf{d})$ of parameters given our observations
- Quantity of Interest is expected value of a functional $Q(\theta)$ (statistics), e.g.

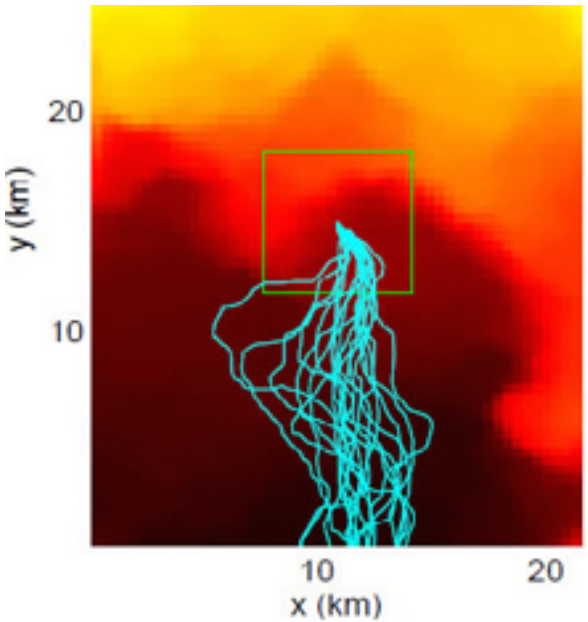
$$\mathbb{E}_{\pi(\theta|\mathbf{d})}[Q(\theta)]$$



Example

Longterm Radioactive Waste Disposal (subsurface flow simulation)

Scenario: Accidental release of radionuclides & transport by groundwater

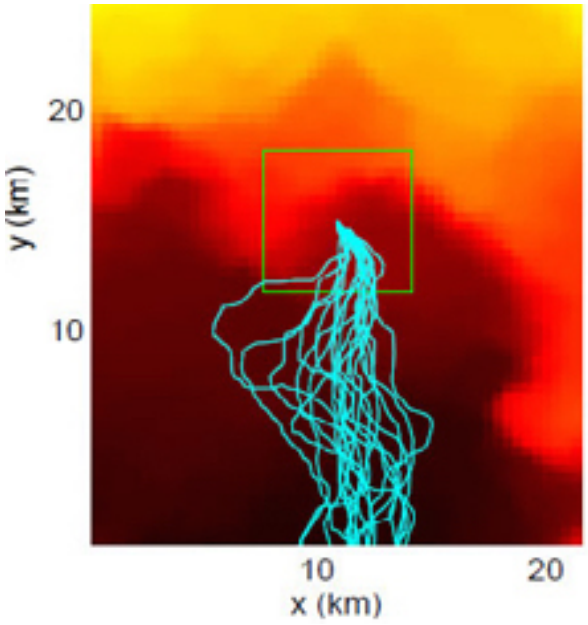


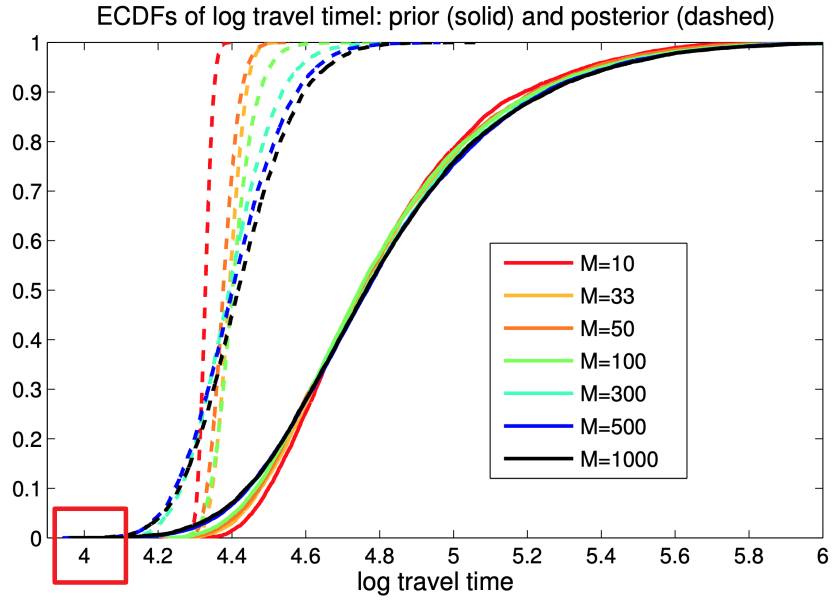
Uncertain particle paths
of leaking radionuclides

Example

Longterm Radioactive Waste Disposal (subsurface flow simulation)

Scenario: Accidental release of radionuclides & transport by groundwater





Source: Ernst et al, 2014

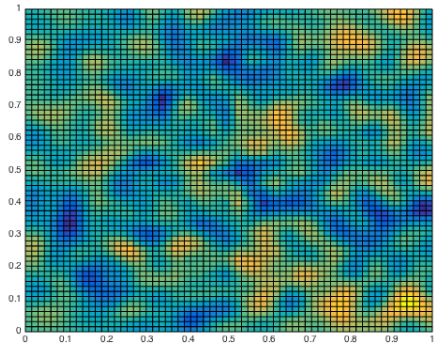
Uncertain particle paths of leaking radionuclides

Quantity of interest: Empirical CDF of travel time, in particular $\mathbb{P}(\tau \leq 10^4 \text{years}) < 10^{-4}$

Simulation unavoidable, **but** material properties are **uncertain!**

Model Inverse Problem (with additive Gaussian noise)

$$y = \underbrace{\mathcal{G} \circ u}_{\mathcal{F}}(\theta_1, \ldots, \theta_d) + \eta$$



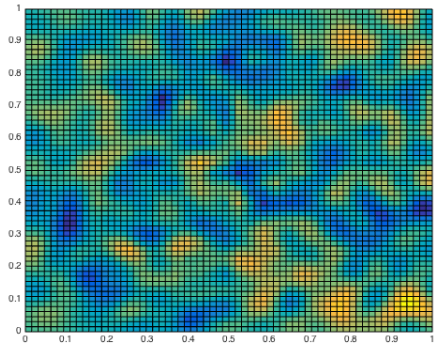
with $u : \mathbb{R}^d \to V$ solution of the (PDE) model

$$-\nabla \cdot (\kappa \nabla u) = f$$

with uncertain $\kappa = \kappa(\theta_1, \ldots, \theta_d)$; observation operator $\mathcal{G} : V \to \mathbb{R}^m$; and Gaussian noise $\eta \sim \mathcal{N}(0, \sigma_\eta^2 I)$

Model Inverse Problem (with additive Gaussian noise)

$$y = \underbrace{\mathcal{G} \circ u}_{\mathcal{F}}(\theta_1, \ldots, \theta_d) + \eta$$



with $u : \mathbb{R}^d \to V$ solution of the (PDE) model

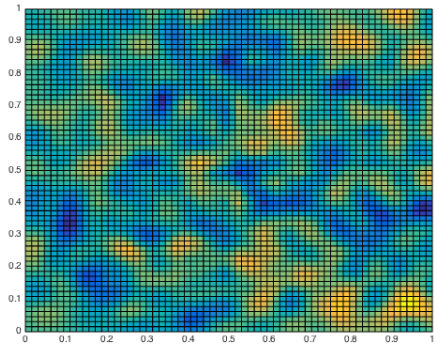
$$-\nabla \cdot (\kappa \nabla u) = f$$

with uncertain $\kappa = \kappa(\theta_1, \ldots, \theta_d)$; observation operator $\mathcal{G} : V \to \mathbb{R}^m$; and Gaussian noise $\eta \sim \mathcal{N}(0, \sigma_\eta^2 I)$

- Prior density $\pi_0(\theta)$
- Data likelihood $L(y|\theta) \propto \exp\left(-\dfrac{\|y - \mathcal{G} \circ u(\theta)\|}{2\sigma_\eta^2}\right)$
- **Bayes' Theorem** $\rightarrow$ Posterior density $\pi(\theta|y) = \dfrac{1}{Z} L(y|\theta)\pi_0(\theta)$
- Quantity of Interest: θ or $Q(\theta)$.

Model Inverse Problem (with additive Gaussian noise)

$$y = \underbrace{\mathcal{G} \circ u}_{\mathcal{F}}(\theta_1, \ldots, \theta_d) + \eta$$



with $u : \mathbb{R}^d \to V$ solution of the (PDE) model

$$-\nabla \cdot (\kappa \nabla u) = f$$

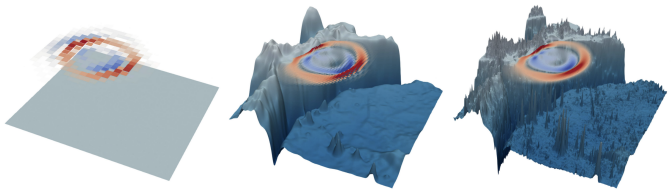
with uncertain $\kappa = \kappa(\theta_1, \ldots, \theta_d)$; observation operator $\mathcal{G} : V \to \mathbb{R}^m$; and Gaussian noise $\eta \sim \mathcal{N}(0, \sigma_\eta^2 I)$

- Prior density $\pi_0(\theta)$
- Data likelihood $L(y|\theta) \propto \exp\left(-\dfrac{\|y - \mathcal{G} \circ u(\theta)\|}{2\sigma_\eta^2}\right)$
- **Bayes' Theorem** $\to$ Posterior density $\pi(\theta|y) = \dfrac{1}{Z} L(y|\theta)\pi_0(\theta)$
- Quantity of Interest: θ or $Q(\theta)$.

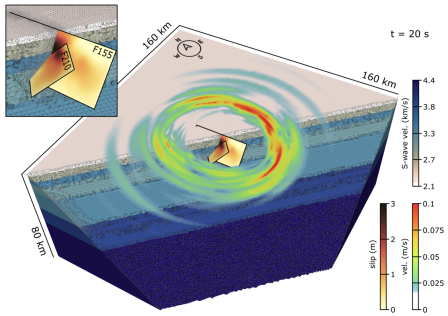
Ultimate task: compute statistics $\equiv$ **high dimensional integrals**

$$\mathbb{E}_\pi Q = \frac{1}{Z} \int Q(\theta) L(y|\theta) \pi_0(\theta) \mathrm{d}\theta, \quad Z = \int L(y|\theta)\pi_0(\theta)\mathrm{d}\theta$$

Other Examples



Tsunamis



Earthquakes

Atmospheric Dispersion

Additive Manufacture

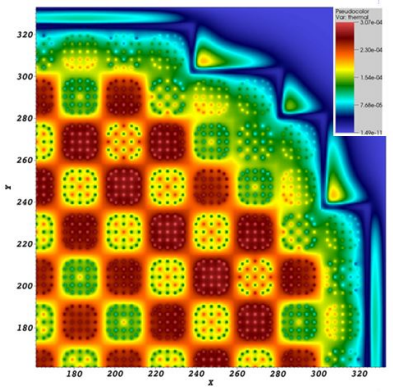


Figure 1 2D Slice of Thermal Flux Distribution near the Core Mid-plane

Nuclear Energy

Geothermal Energy

MCMC

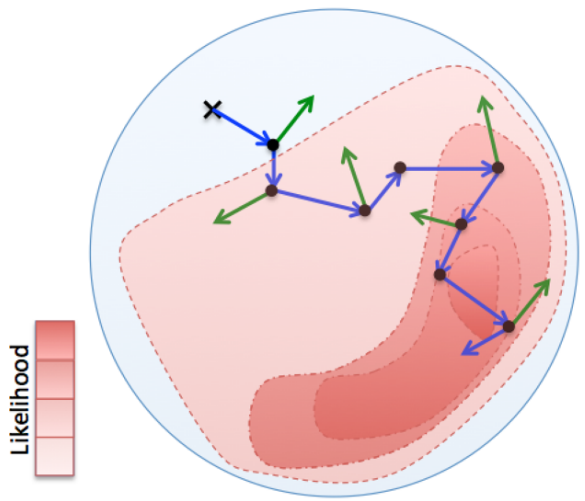
Markov Chain Monte Carlo - Metropolis-Hastings

Algorithm 1. Metropolis-Hastings (MH)

- Given θ^n, generate a proposal x distributed as $q(x|\theta^n)$,
- Accept proposal x as the next state, i.e. set $\theta^{n+1} = x$, with probability

$$\alpha(x|\theta^n) = \min\left\{1, \frac{\pi(x)q(\theta^n|x)}{\pi(\theta^n)q(x|\theta^n)}\right\} \tag{1}$$

otherwise reject x and set $\theta^{n+1} = \theta^n$.



Markov Chain Monte Carlo - Metropolis-Hastings

The **Good Things** about **Metropolis-Hastings**

- **Simple!**
- Repeated iterations generate a (homogeneous) Markov chain.
- **MH** (Alg. 1) kernel is in detailed balance with π, i.e.

$$\pi(\theta) K(x|\theta) = \pi(x) K(\theta|x),$$

- Under mild conditions on $q(\cdot|\cdot)$ and start point

$$\Theta := \{\theta^0, \theta^1, \ldots, \theta^N\} \sim \pi$$

Markov Chain Monte Carlo - Metropolis-Hastings

The **Good Things** about **Metropolis-Hastings**

- **Simple!**
- Repeated iterations generate a (homogeneous) Markov chain.
- **MH** (Alg. 1) kernel is in detailed balance with π, i.e.

$$\pi(\theta) K(x|\theta) = \pi(x) K(\theta|x),$$

- Under mild conditions on $q(\cdot|\cdot)$ and start point

$$\Theta := \{\theta^0, \theta^1, \ldots, \theta^N\} \sim \pi$$

The **Big Challenges** with **Metropolis-Hastings**

1. Evaluating π can be **computationally very expensive**!
2. **Markov Chain** $\Theta := \{\theta^0, \theta^1, \ldots, \theta^N\}$ is strongly correlated.
3. **Difficult to Parallelise** - fundamental challenge since by their nature Markov processes are **sequential**.

Exploiting Hierarchies of Models

Lemma 1. If proposal transition kernel $q(\cdot|\cdot)$ in Alg. 1 is in detailed balance with some distribution π_C, then acceptance probability (1) may be written

$$\alpha(x|\theta^n) = \min\left\{1, \frac{\pi(x)\pi_C(\theta^n)}{\pi(\theta^n)\pi_C(x)}\right\} \tag{2}$$

Proof Substitute detailed balance statement $\pi_C(x)q(\theta^n|x) = \pi_C(\theta^n)q(x|\theta^n)$ into (1) to get (2), almost everywhere.

Exploiting Hierarchies of Models

Lemma 1. If proposal transition kernel $q(\cdot|\cdot)$ in Alg. 1 is in detailed balance with some distribution π_C, then acceptance probability (1) may be written

$$\alpha(x|\theta^n) = \min\left\{1, \frac{\pi(x)\pi_C(\theta^n)}{\pi(\theta^n)\pi_C(x)}\right\} \tag{2}$$

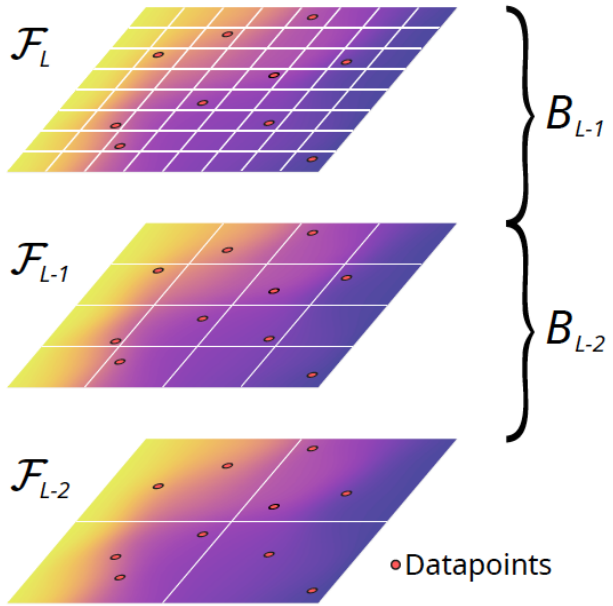
Proof Substitute detailed balance statement $\pi_C(x)q(\theta^n|x) = \pi_C(\theta^n)q(x|\theta^n)$ into (1) to get (2), almost everywhere.

- Idea is to **exploit a surrogate** or a **hierarchy of approximate models** $\mathcal{F}_\ell$:

  FE grid u_ℓ / parameters θ_ℓ / data $\mathbf{d}_\ell$
- Consider just **two levels** and **only** grid resolution (no hierarchy on parameters or data)
- Therefore have

  Fine: Target $\pi \equiv \pi_F$

  Coarse: Approximation π_C



Recall: Multilevel Monte Carlo [Heinrich, '98], [Giles, '07]

Basic Idea: Note that trivially (due to linearity of $\mathbb{E}$)

$$\mathbb{E}[Q_L] = \mathbb{E}[Q_0] + \sum_{\ell=1}^{L} \mathbb{E}[Q_\ell - Q_{\ell-1}]$$

Control Variates!!

Define the following **multilevel MC** estimator for $\mathbb{E}[Q]$:

$$\widehat{Q}_L^{MLMC} := \hat{Q}_0^{\mathrm{MC}} + \sum_{\ell=1}^{L} \hat{Y}_\ell^{\mathrm{MC}} \quad \text{where} \quad Y_\ell := Q_\ell - Q_{\ell-1}$$

Recall: Multilevel Monte Carlo [Heinrich, '98], [Giles, '07]

Basic Idea: Note that trivially (due to linearity of $\mathbb{E}$)

$$\mathbb{E}[Q_L] = \mathbb{E}[Q_0] + \sum_{\ell=1}^{L} \mathbb{E}[Q_\ell - Q_{\ell-1}]$$

Control Variates!!

Define the following **multilevel MC** estimator for $\mathbb{E}[Q]$:

$$\widehat{Q}_L^{MLMC} := \hat{Q}_0^{\mathrm{MC}} + \sum_{\ell=1}^{L} \hat{Y}_\ell^{\mathrm{MC}} \quad \text{where} \quad Y_\ell := Q_\ell - Q_{\ell-1}$$

Key Observation: (Variance Reduction! Corrections cheaper!)

Level L: $\mathbb{V}[Q_L - Q_{L-1}] \to 0$ as $L \to \infty \;\Rightarrow\; N_L = \mathcal{O}(1)$ (best case)

$\vdots$

Level ℓ: N_ℓ optimised to "balance" with cost on levels 0 and L

$\vdots$

Level 0: $N_0 \sim N$ **but** $\mathrm{Cost}_0 = \mathcal{O}(M_0) = \mathcal{O}(1)$ $\to$ Complexity Theorem!

Multilevel Markov Chain Monte Carlo - Bottom Up Approach

[Dodwell, Ketelsen, RS, Teckentrup, SIAM JUQ 3, 2015] & [SIAM Review 61, 2019]

Two key **motivating points**:

1. Exploit multilevel **variance reduction mechanism** (as above)

$$\mathbb{E}_{\pi_F}(Q_F) = \mathbb{E}_{\pi_C}(Q_C) + \Big[\underbrace{\mathbb{E}_{\pi_F}(Q_F) - \mathbb{E}_{\pi_C}(Q_C)}_{\text{Correlate chains!}}\Big]$$

Multilevel Markov Chain Monte Carlo - Bottom Up Approach

[Dodwell, Ketelsen, RS, Teckentrup, SIAM JUQ 3, 2015] & [SIAM Review 61, 2019]

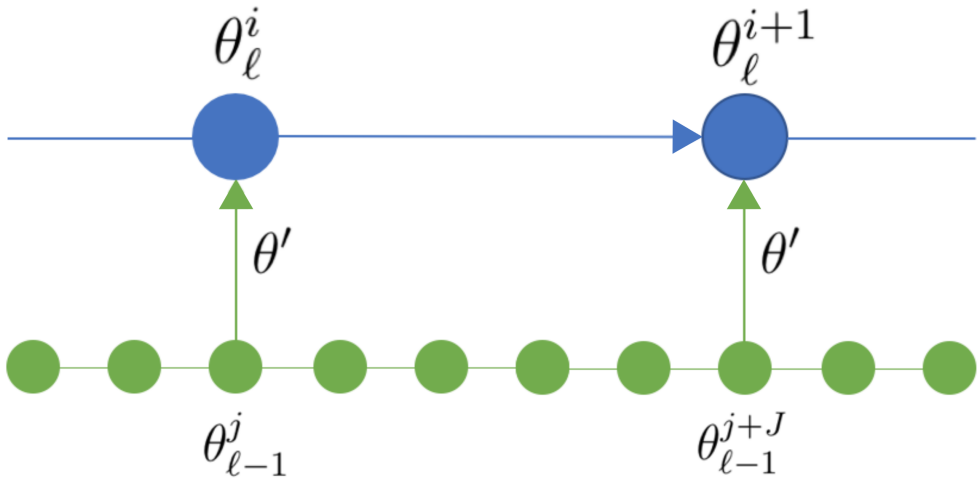
Two key **motivating points**:

1. Exploit multilevel **variance reduction mechanism** (as above)

$$\mathbb{E}_{\pi_F}(Q_F) = \mathbb{E}_{\pi_C}(Q_C) + \Big[\underbrace{\mathbb{E}_{\pi_F}(Q_F) - \mathbb{E}_{\pi_C}(Q_C)}_{\text{Correlate chains!}}\Big]$$

2. **and** use subchains generated π_C to cheaply build **good proposals**.

Algorithm in a picture



Multilevel Markov Chain Monte Carlo - small flies in the ointment!

This MLMCMC Alg. is not (exactly) Markovian!

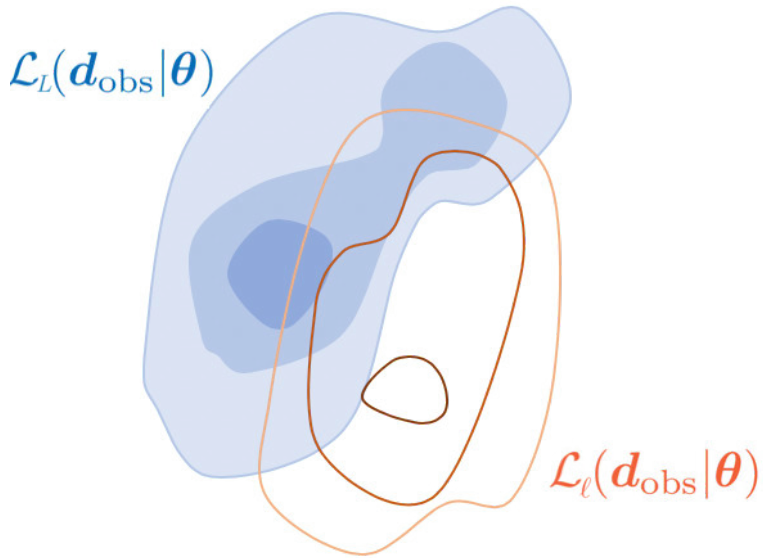
- If we **reject** on the fine, **coarse is not reset**
- Theoretically only works if subsampling rate $J = \infty$
- Works well in practise if $J > \tau$ (autocorrelation length of subchain).

Multilevel Markov Chain Monte Carlo - small flies in the ointment!

This MLMCMC Alg. is not (exactly) Markovian!

- If we **reject** on the fine, **coarse is not reset**
- Theoretically only works if subsampling rate $J = \infty$
- Works well in practise if $J > \tau$ (autocorrelation length of subchain).

Struggles if difference between π_F and π_C is **too big**.

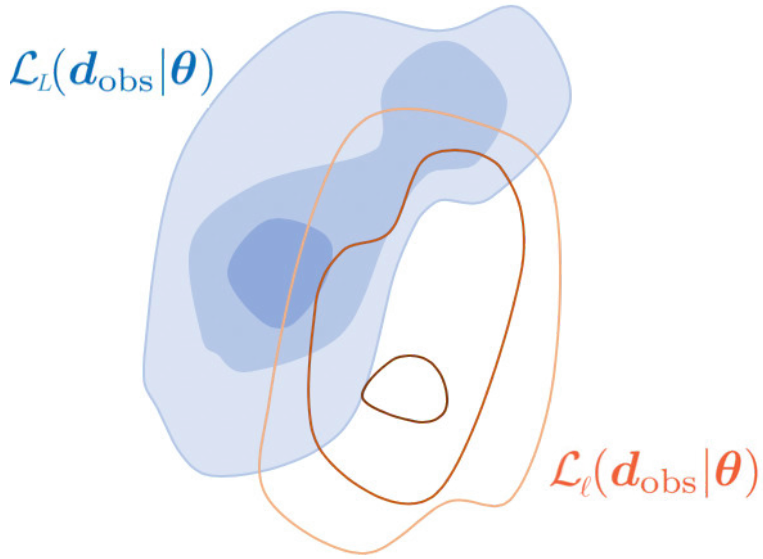


Multilevel Markov Chain Monte Carlo - small flies in the ointment!

This MLMCMC Alg. is not (exactly) Markovian!

- If we **reject** on the fine, **coarse is not reset**
- Theoretically only works if subsampling rate $J = \infty$
- Works well in practise if $J > \tau$ (autocorrelation length of subchain).

Struggles if difference between π_F and π_C is **too big**.



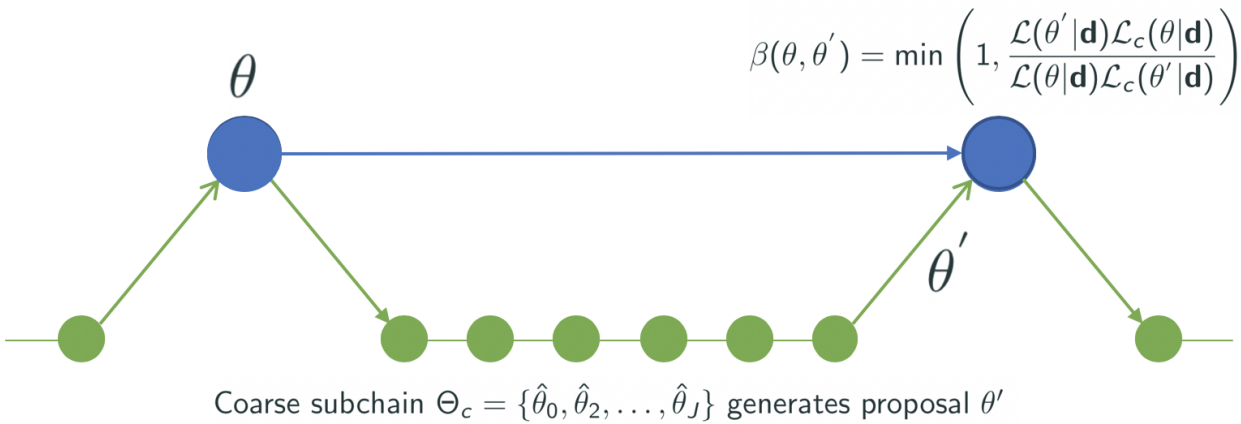
Adaptive Multilevel Delayed Acceptance addresses **both problems!**

Multilevel Delayed Acceptance

Multilevel Delayed Acceptance - Top Down

[Liu, 2001], [Christen, Fox, 2005], [Lykkegaard, Dodwell, Fox, Mingas, RS, 2023]

Run **finite** length subchain of random length $J \sim p(\cdot)$ on coarse level:

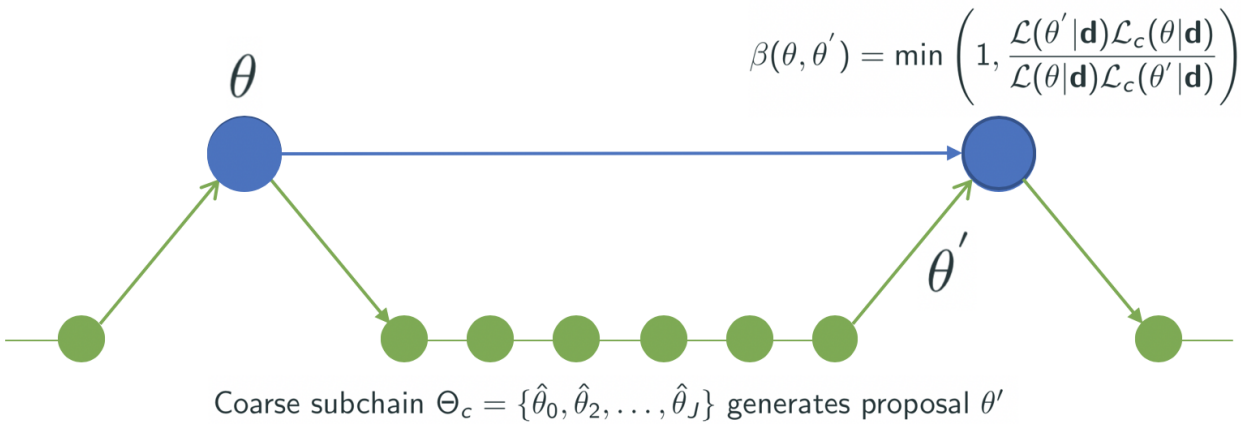


Idea: Use coarse approximation $\pi_C \approx \pi_F$ to generate proposals for π_F.

Multilevel Delayed Acceptance - Top Down

[Liu, 2001], [Christen, Fox, 2005], [Lykkegaard, Dodwell, Fox, Mingas, RS, 2023]

Run **finite** length subchain of random length $J \sim p(\cdot)$ on coarse level:



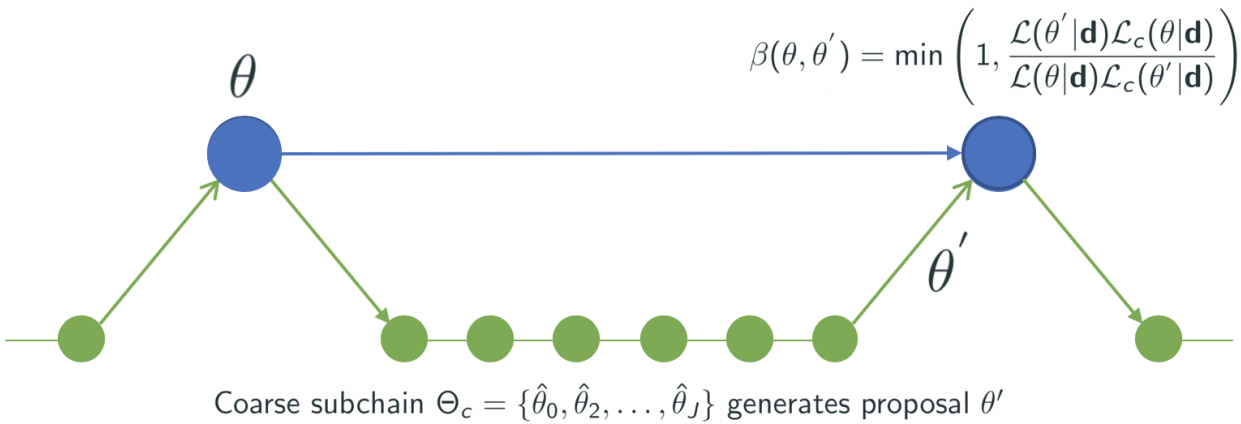
Idea: Use coarse approximation $\pi_C \approx \pi_F$ to generate proposals for π_F.

- Small correlation length and good mixing (e.g. via tempering)
- Cost saving $\approx$ cost ratio between $\mathcal{L}_F$ and $\mathcal{L}_C$ (exponetntial in ℓ) times the **acceptance rate** (typically close to 1)
- Generates a Markov Chain in **detailed balance** with π_F
  (as for surrogate transition method, see below or [Liu, 2001])

Multilevel Delayed Acceptance - Top Down

[Liu, 2001], [Christen, Fox, 2005], [Lykkegaard, Dodwell, Fox, Mingas, RS, 2023]

Run **finite** length subchain of random length $J \sim p(\cdot)$ on coarse level:



Idea: Use coarse approximation $\pi_C \approx \pi_F$ to generate proposals for π_F.

- Small correlation length and good mixing (e.g. via tempering)
- Cost saving $\approx$ cost ratio between $\mathcal{L}_F$ and $\mathcal{L}_C$ (exponetntial in ℓ) times the **acceptance rate** (typically close to 1)
- Generates a Markov Chain in **detailed balance** with π_F
  (as for surrogate transition method, see below or [Liu, 2001])
- **And on top** we can again exploit variance reduction!

Multilevel Delayed Acceptance – Good Proposals

Algorithm 2. Random-Length-Subchain Surrogate Transition

Input: Fine density $\pi_{\mathsf{F}}(\cdot)$, Coarse density $\pi_{\mathsf{C}}(\cdot)$, proposal kernel $q(\cdot|\cdot)$, initial state θ^0

- Draw the subchain length $J \sim p(\cdot)$.
- Starting at θ^n, generate subchain of length J using MH Algorithm 1 for coarse distribution π_{C}, i.e.,

$$x = \mathbf{MH}\left(\pi_{\mathsf{C}}(\cdot), q(\cdot|\cdot), \theta^n, J\right)$$

- Accept proposal x as next sample, i.e. set $\theta^{n+1} = x$, with probability

$$\alpha(x|\theta^n) = \min\left\{1, \frac{\pi_{\mathsf{F}}(x)\pi_{\mathsf{C}}(\theta^n)}{\pi_{\mathsf{F}}(\theta^n)\pi_{\mathsf{C}}(x)}\right\}.$$

otherwise reject and set $\theta^{n+1} = \theta^n$.

Multilevel Delayed Acceptance – Detailed Balance

Lemma 2. Let $K_1(\cdot|\cdot)$ and $K_2(\cdot|\cdot)$ be two commuting transition kernels, each in detailed balance with distribution π, then the composition $(K_1 \circ K_2)$ is in detailed balance with π.

Theorem 3. Alg. 2 simulates a Markov chain in detailed balance with π_{F}.

Multilevel Delayed Acceptance – Detailed Balance

Lemma 2. Let $K_1(\cdot|\cdot)$ and $K_2(\cdot|\cdot)$ be two commuting transition kernels, each in detailed balance with distribution π, then the composition $(K_1 \circ K_2)$ is in detailed balance with π.

Theorem 3. Alg. 2 simulates a Markov chain in detailed balance with π_F.

Proof.

- q_C comutes with itself
- By induction q_C^n (application n times) is in detailed balance with π_C.
- Random subchain gives an effective mixture kernel

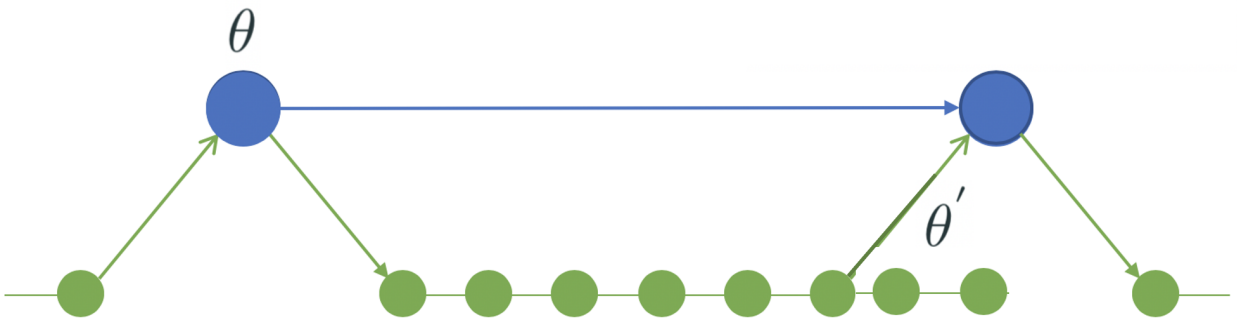
$$\sum_{n \in \mathbb{Z}^+} p(n) q_C^n(\cdot|\cdot)$$

- Apply **Lemma 1** $\Rightarrow$ this kernel is in detailed balance with π_F.

Multilevel Delayed Acceptance – Variance Reduction

Efficient proposal mechanism **but** can also use for variance reduction:

- Coarse subchain **not** $\sim \pi_C$: like **'mini-burn-ins'** from π_F



- For simplicity, fix $J_{\max}$, and choose $p(\cdot)$ uniform on $\{1, \ldots, J_{\max}\}$.
- Samples from 'hybrid' mixture distribution

$$\tilde{\pi}_C = \frac{1}{J_{\max}} \sum_{j=1}^{J_{\max}} K_C^j \, \pi_F \quad \text{with} \quad K_C^j = \underbrace{K_C \circ K_C \circ \ldots \circ K_C}_{j \text{ times}} .$$

- Irrelevant that $\tilde{\pi}_C \neq \pi_C$. In fact better, because closer to π_F!

  (Crucial! Coarse chain resets every time, in particular if the fine chain rejects!)

Multilevel Delayed Acceptance – Variance Reduction

- **Telescoping sum:**

$$\mathbb{E}_{\pi_F}(Q_F) = \mathbb{E}_{\tilde{\pi}_C}(Q_C) + [\mathbb{E}_{\pi_F}(Q_F) - \mathbb{E}_{\tilde{\pi}_C}(Q_C)]$$

- Exploit **variance reduction** in second term:

$$\mathbb{E}_{\pi_F}(Q_F) \approx \underbrace{\frac{1}{NJ_{\max}} \sum_{n=1}^{N} \sum_{j=1}^{J_{\max}} Q_C^{n,j}}_{\text{all samples!}} + \underbrace{\frac{1}{N} \sum_{n=1}^{N} \left(Q_F^n - Q_C^{n,J_n} \right)}_{\text{one sample per subchain!}}$$

where J_n is the randomly chosen index in $\{1, \ldots, J_{\max}\}$ at state n.

Multilevel Delayed Acceptance – Variance Reduction

- **Telescoping sum:**

$$\mathbb{E}_{\pi_F}(Q_F) = \mathbb{E}_{\tilde{\pi}_C}(Q_C) + [\mathbb{E}_{\pi_F}(Q_F) - \mathbb{E}_{\tilde{\pi}_C}(Q_C)]$$

- Exploit **variance reduction** in second term:

$$\mathbb{E}_{\pi_F}(Q_F) \approx \underbrace{\frac{1}{NJ_{\max}} \sum_{n=1}^{N} \sum_{j=1}^{J_{\max}} Q_C^{n,j}}_{\text{all samples!}} + \underbrace{\frac{1}{N} \sum_{n=1}^{N} \left(Q_F^n - Q_C^{n,J_n} \right)}_{\text{one sample per subchain!}}$$

  where J_n is the randomly chosen index in $\{1, \ldots, J_{\max}\}$ at state n.

- Optimal $J_{\max}$ proportional to ratio of variances $\times$ ratio of i.a.c.t.'s.

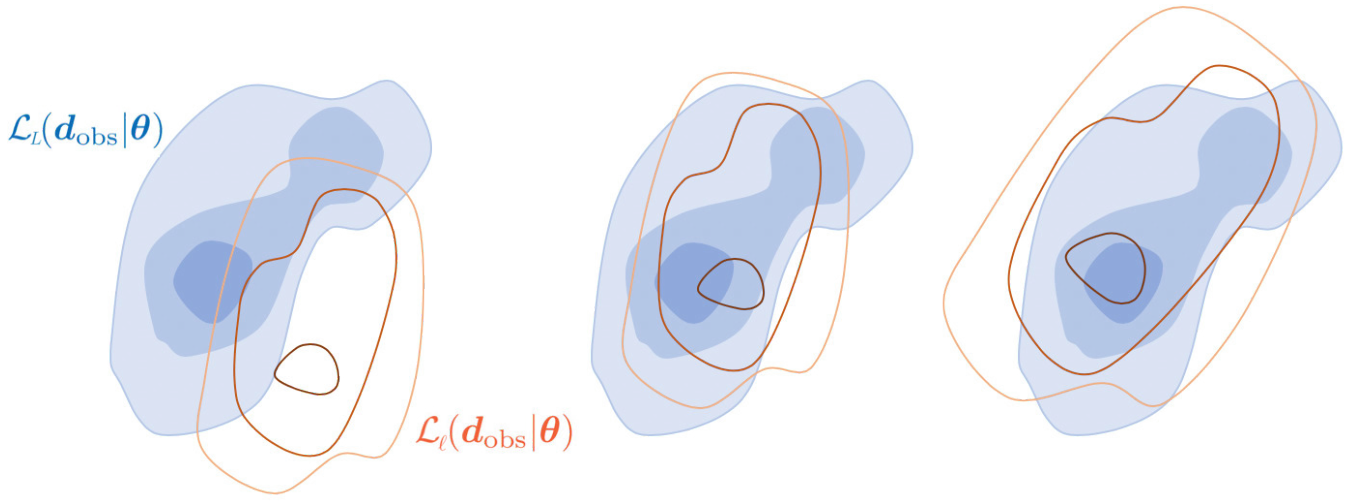
- **Ongoing work (w. Colin Fox):** for elliptic model problem can prove variance reduction and $1 - \alpha_\ell = \mathcal{O}(h^\alpha)$ as for MLMCMC

  (essentially identical proof!)

Adaptive Correction - Wrong models can be made less wrong!

[Cui, Fox, O'Sullivan, 2011], [Cui, Fox, O'Sullivan, 2019], [Lykkegaard et al., 2023]

- **Significant issue** if big difference between fine and coarse posterior!

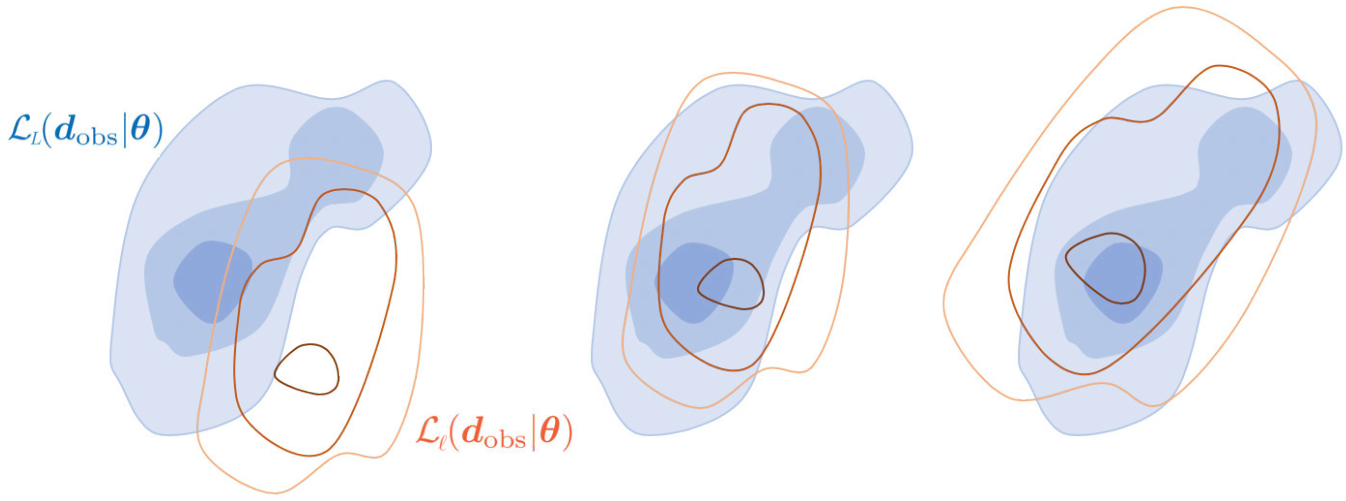


- **However,** every time we accept/reject we get a sample of $\mathcal{F}_F - \mathcal{F}_C$

Adaptive Correction - Wrong models can be made less wrong!

[Cui, Fox, O'Sullivan, 2011], [Cui, Fox, O'Sullivan, 2019], [Lykkegaard et al., 2023]

- **Significant issue** if big difference between fine and coarse posterior!



- **However,** every time we accept/reject we get a sample of $\mathcal{F}_F - \mathcal{F}_C$
- Use multilevel trick on statistical model and estimate correction via a multivariate Gaussian:

$$\mathbf{d} - \mathcal{F}_C = \underbrace{\mathcal{F}_F - \mathcal{F}_C}_{\mathcal{B}_F \sim \mathcal{N}(\mu_{B,F}, \Sigma_{B,F})} + \underbrace{\eta}_{\mathcal{N}(0, \Sigma_\eta)}$$

Adaptive Correction - Learning on-the-fly

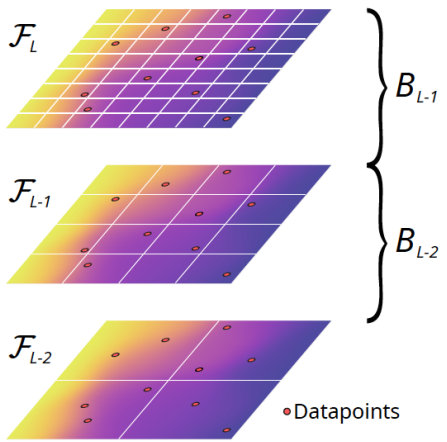
- Likelihood on coarse level now addition of two Gaussians:

$$\mathcal{L}_C = \exp\left(-\frac{1}{2}(\mathcal{F}_C(\theta) + \boldsymbol{\mu}_{\mathcal{B},F} - \mathbf{d})^{\mathsf{T}}(\boldsymbol{\Sigma}_{\mathcal{B},\mathbf{F}} + \boldsymbol{\Sigma}_{\mathbf{e}})^{-1}(\mathcal{F}_{\mathbf{C}}(\theta) + \boldsymbol{\mu}_{\mathcal{B},\mathbf{F}} - \mathbf{d})\right)$$

- Corrections can be built recursively (with very small overhead):

$$\mu_{B,F} \leftarrow \frac{1}{i+1}\left(i\mu_{B,F} + \mathcal{B}(\theta^{i+1})\right) \quad \text{and} \quad \Sigma_{B,F} \leftarrow \frac{i-1}{i}\Sigma_{B,F} + \frac{1}{i}\left(\dots\right)$$

- Repeat over all levels - by summing all biases between levels.
- Use diminishing adaptation to prove convergence.



Implementation in pymc3 - versions $>$ 3.10

https://docs.pymc.io

https://www.pymc.io/projects/examples/en/latest/samplers/MLDA_introduction.html

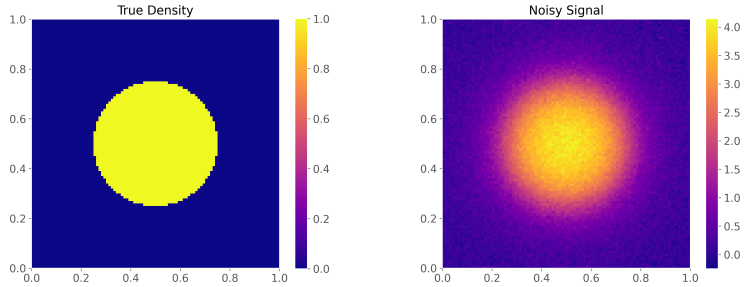
Lightweight code called tinyDA by Mikkel

https://github.com/mikkelbue/tinyDA

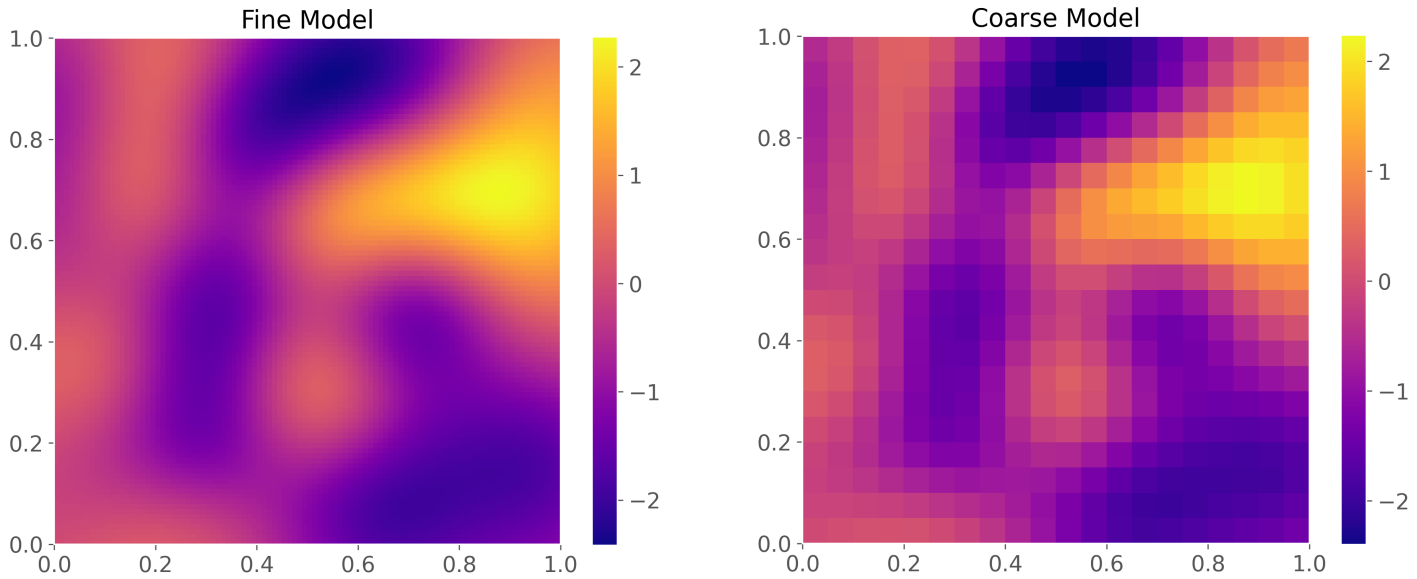
Numerical Results

Numerical Example: Gravitational Survey

- For geological exploration (linear inverse problem; integral equation).
- Toy phantom and noisy signal:
  (with $\sigma_e^2 = 0.01$)

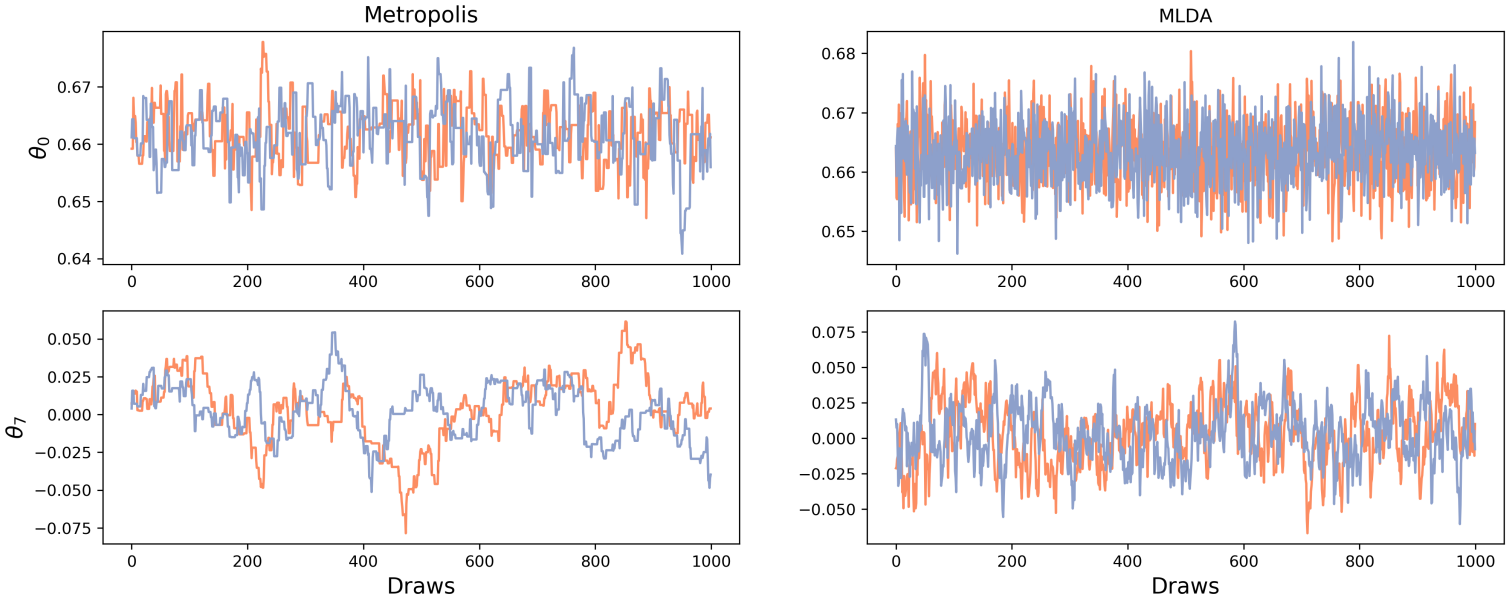


- Gaussian RF prior parametrised via truncated KL expansion with $d = 32$ modes, and 100x100 and 20x20 pixels for π_F and π_C, resp.:

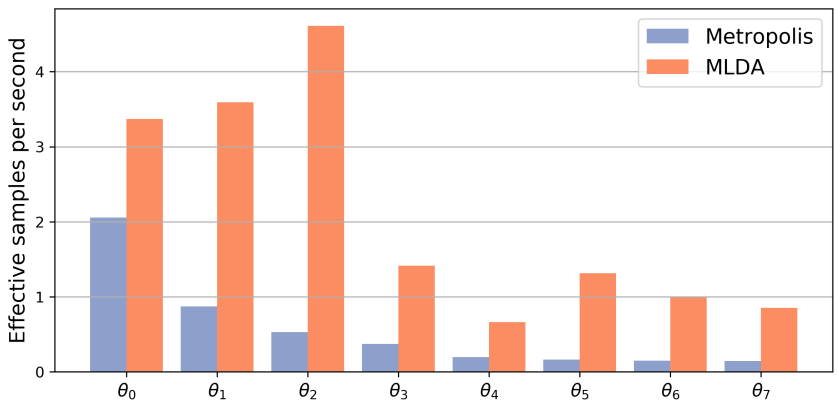


Numerical Example: Gravitational Survey

Traces of θ_1 (top) and θ_9 (bottom) for RWMH (left) and MLDA (right):



Algorithmic performance: giving **effective samples / second** for $\theta_1 \ldots \theta_8$:



Model Inverse Problem – Subsurface Flow

- **Darcy's equation** + BCs.
- $u(\mathbf{x})$ pressure, $k(\mathbf{x}, \boldsymbol{\theta})$ permeability, $f(\mathbf{x})$ source.
- Classical FEM approximation w.r.t. mesh $\mathcal{T}_h$

$$\int_D k(\mathbf{x}, \boldsymbol{\theta}) \nabla u_h \cdot \nabla v_h \, d\mathbf{x} = \int_D f v_h \, d\mathbf{x}$$

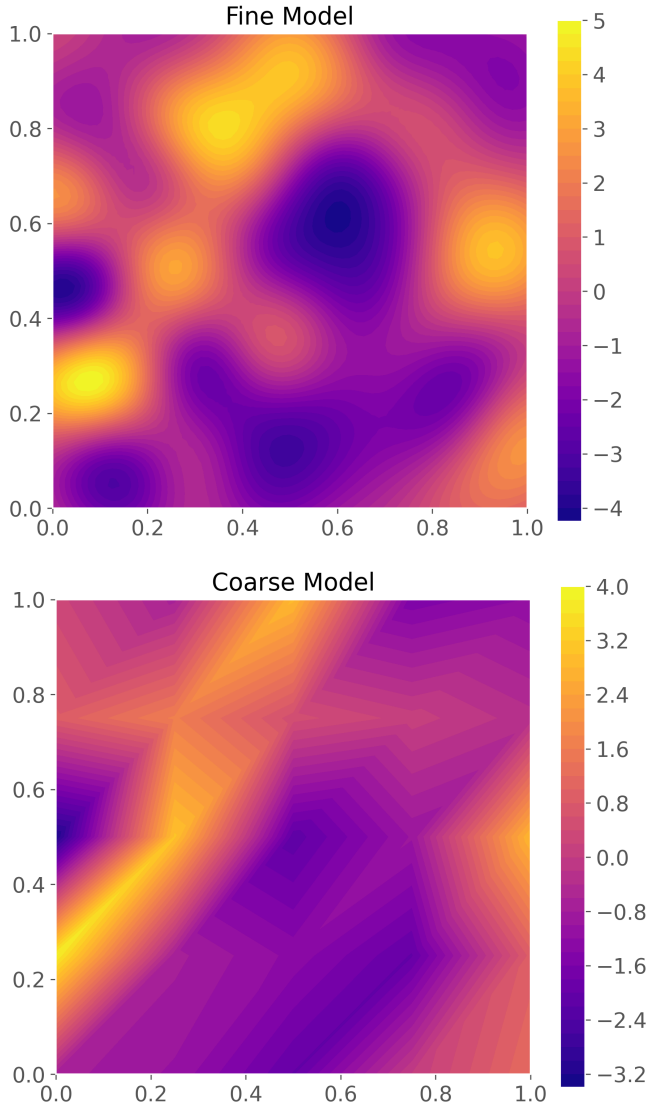
- Level 0 mesh $\mathcal{T}_{h_0}$ **very** coarse: 5×5 !!

Model Inverse Problem – Subsurface Flow

- **Darcy's equation** + BCs.
- $u(\mathbf{x})$ pressure, $k(\mathbf{x}, \boldsymbol{\theta})$ permeability, $f(\mathbf{x})$ source.
- Classical FEM approximation w.r.t. mesh $\mathcal{T}_h$

$$\int_D k(\mathbf{x}, \boldsymbol{\theta}) \nabla u_h \cdot \nabla v_h \, d\mathbf{x} = \int_D f v_h \, d\mathbf{x}$$

- Level 0 mesh $\mathcal{T}_{h_0}$ **very** coarse: 5×5 !!

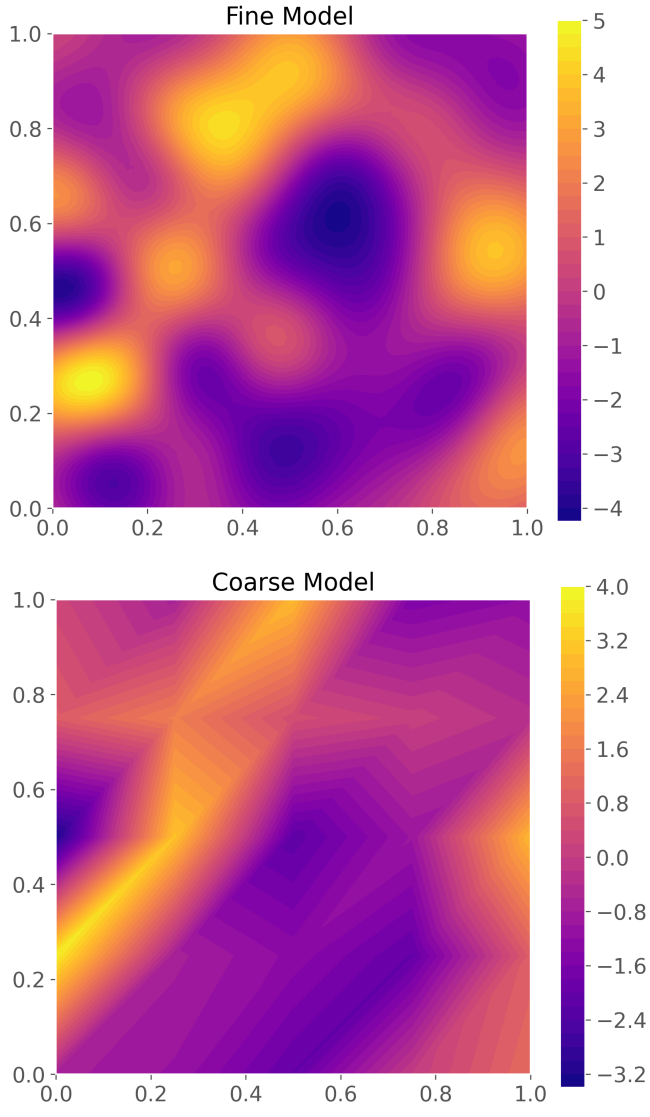


Model Inverse Problem – Subsurface Flow

- **Darcy's equation** + BCs.
- $u(\mathbf{x})$ pressure, $k(\mathbf{x}, \boldsymbol{\theta})$ permeability, $f(\mathbf{x})$ source.
- Classical FEM approximation w.r.t. mesh $\mathcal{T}_h$

$$\int_D k(\mathbf{x}, \boldsymbol{\theta}) \nabla u_h \cdot \nabla v_h \, d\mathbf{x} = \int_D f v_h \, d\mathbf{x}$$

- Level 0 mesh $\mathcal{T}_{h_0}$ **very** coarse: 5×5 !!
- **Data:** Pressure evaluated at $m = 25$ points in D; measurement noise $\sigma_\eta^2 = 10^{-4}$.
- **Prior:** lognormal $k(\mathbf{x}, \boldsymbol{\theta})$ w. Gaussian covariance
  ($\lambda = 0.1$, $\sigma^2 = 2$; parametrised using $d = 64$ KL-modes)



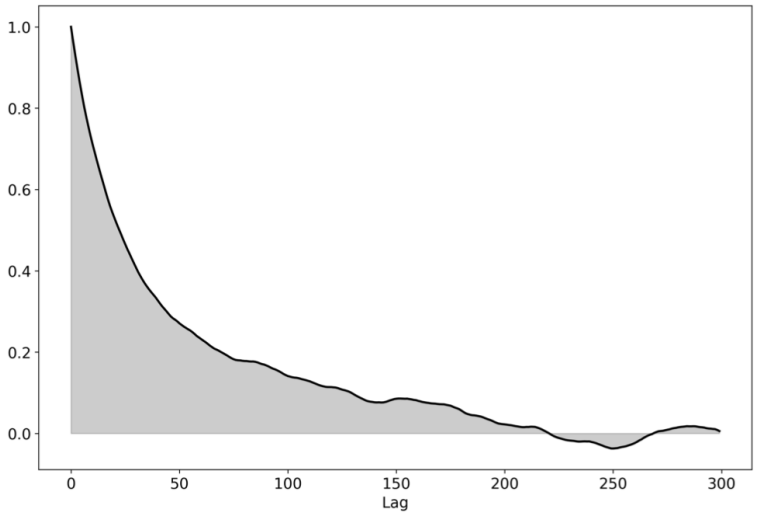
Numerical Experiment: Gains through Adaptive Error Model

Without Adaptive Error Model:

- Effective Sample Size

$$\text{ESS} = 326/40000$$

(iact = 122.7)

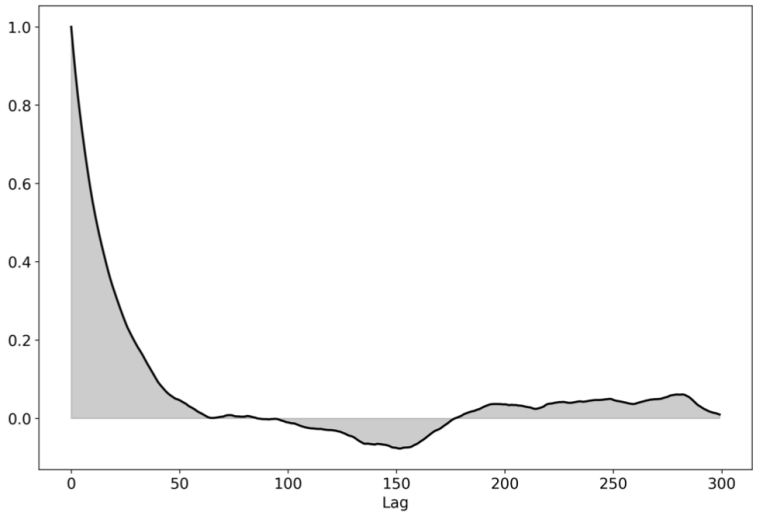


With Adaptive Error Model:

- Effective Sample Size

$$\text{ESS} = 1012/40000$$

(iact = 39.5)



Numerical Comparison (MLMCMC not MLDA!)

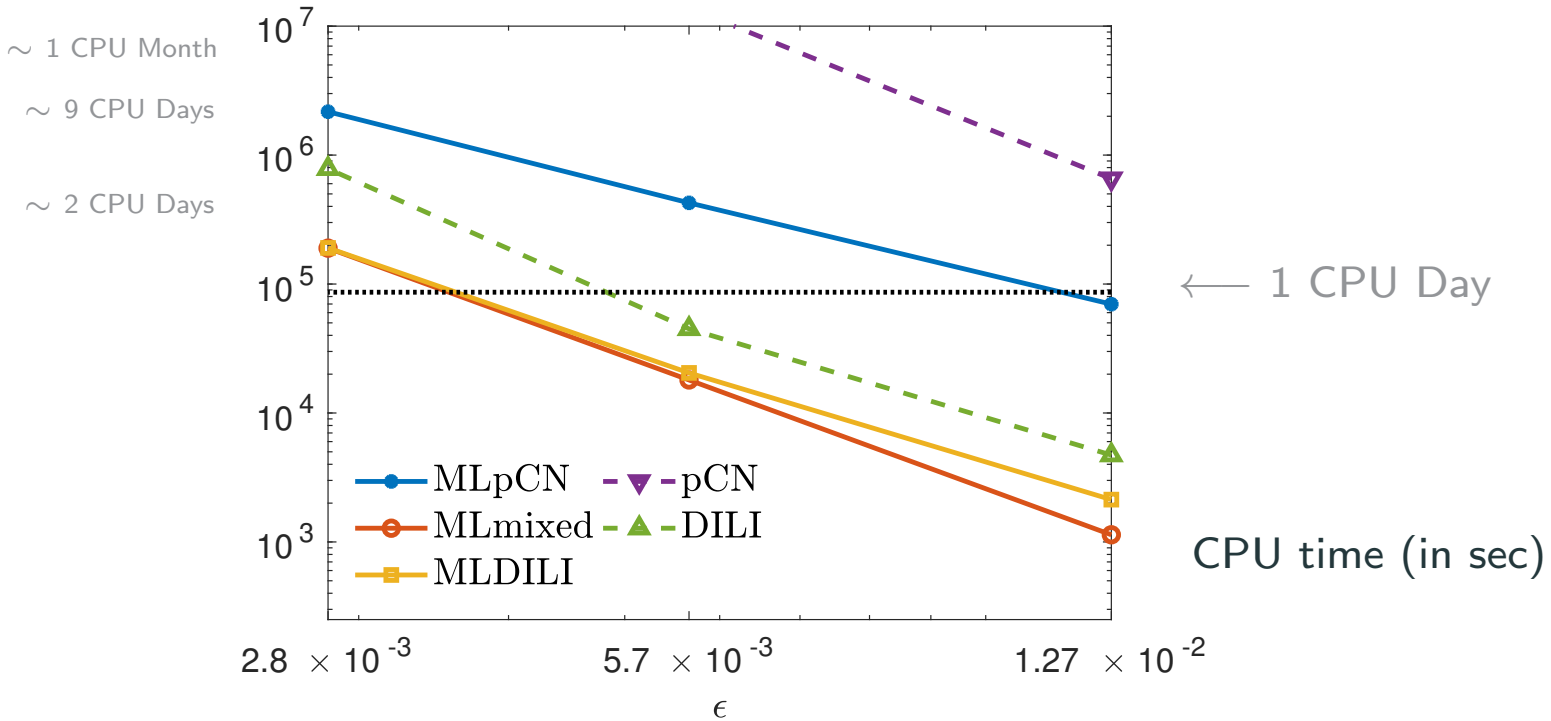
[Cui, Detommaso, RS, 2024]

$$Q_\ell(\theta^n_{\ell,\ell}) - Q_{\ell-1}(\theta^n_{\ell,\ell-1})$$

Level ℓ	0	1	2	3
iact(pCN)	4100	4.9	2.8	1.9
iact(DILI)	9.0	4.6	2.4	1.8

Using model order reduction and **likelihood-informed subspaces** – in particular **DILI MCMC**

[Cui, Law, Marzouk, *JCP* **304**, 2016]



Diving Deeper into the Theory

The Tension

Acceptance probability in MLDA (Target $\pi = \pi_F$; Surrogate $\widetilde{\pi} = \pi_C$)

$$\alpha(x, z) = \min\left\{1, \frac{\pi(z)\,\widetilde{\pi}(x)}{\pi(x)\,\widetilde{\pi}(z)}\right\} \quad \rightarrow \quad \text{When is } \alpha \approx 1?$$

The Tension

Acceptance probability in MLDA (Target $\pi = \pi_F$; Surrogate $\widetilde{\pi} = \pi_C$)

$$\alpha(x, z) = \min\left\{1, \frac{\pi(z)\,\widetilde{\pi}(x)}{\pi(x)\,\widetilde{\pi}(z)}\right\} \quad \rightarrow \quad \text{When is } \alpha \approx 1?$$

(a) **Proximity:**

$$\frac{\pi(z)}{\pi(x)} \to 1 \quad \text{and} \quad \frac{\widetilde{\pi}(x)}{\widetilde{\pi}(z)} \to 1 \quad \text{as } z \to x \quad \text{(Regularity of } \pi, \widetilde{\pi})$$

The Tension

Acceptance probability in MLDA (Target $\pi = \pi_F$; Surrogate $\widetilde{\pi} = \pi_C$)

$$\alpha(x, z) = \min\left\{1,\ \frac{\pi(z)\,\widetilde{\pi}(x)}{\pi(x)\,\widetilde{\pi}(z)}\right\} \quad \rightarrow \quad \text{When is } \alpha \approx 1?$$

(a) **Proximity:**

$$\frac{\pi(z)}{\pi(x)} \to 1 \quad \text{and} \quad \frac{\widetilde{\pi}(x)}{\widetilde{\pi}(z)} \to 1 \quad \text{as } z \to x \quad (\text{Regularity of } \pi, \widetilde{\pi})$$

(b) **Fidelity:**

$$\frac{\pi(z)}{\widetilde{\pi}(z)} \to 1 \quad \text{and} \quad \frac{\widetilde{\pi}(x)}{\pi(x)} \to 1 \quad \text{as } \widetilde{\pi} \to \pi \quad (\textit{pointwise})$$

The Tension

Cheap surrogate

- \+ Can afford many steps
- – Low fidelity

Possible to move far, but small acceptance

→ chain barely moves

Expensive surrogate

- \+ High fidelity
- – Only few steps possible

High acceptance, but only few steps possible

→ little gain over local schemes

Can we break the leap far vs. land safely tradeoff?

Modification 1: Regularisation

Localise the surrogate density: $(\gamma > 0)$

$$\widetilde{\pi} \propto e^{-\tilde{f}} \longrightarrow \widetilde{\pi}_x \propto e^{-\Phi_x}, \qquad \Phi_x(y) := \tilde{f}(y) + \frac{\gamma}{2}\|y - x\|^2$$

$\gamma \to \infty$

- Penalty dominates
- Proposals concentrate near x

$\to$ Local exploration

$\gamma \to 0$

- Surrogate dominates
- Approximately $z \sim \widetilde{\pi}$

$\to$ Global exploration

Modification 1: Regularisation

Localise the surrogate density: $(\gamma > 0)$

$$\widetilde{\pi} \propto e^{-\tilde{f}} \longrightarrow \widetilde{\pi}_x \propto e^{-\Phi_x}, \qquad \Phi_x(y) := \tilde{f}(y) + \frac{\gamma}{2}\|y - x\|^2$$

$\gamma \to \infty$

- Penalty dominates
- Proposals concentrate near x

$\to$ Local exploration

$\gamma \to 0$

- Surrogate dominates
- Approximately $z \sim \widetilde{\pi}$

$\to$ Global exploration

γ controls proximity independently of surrogate fidelity.

Modification 2: Tempering

Interpolate between surrogate $\propto e^{-\tilde{f}}$ and benign reference density $\propto e^{-g}$

$$\tilde{f} \quad \mapsto \quad \theta\,\tilde{f} + (1-\theta)\,g \qquad \theta \in (0,1]$$

Use Gaussian measure as reference (as in regularisation before):

$$\widetilde{\pi}_x \propto e^{-\Phi_x}, \qquad \Phi_x(y) = \theta\tilde{f}(y) + \frac{\gamma}{2}\|y-x\|^2$$

$\theta \to 0$

- isotropic Gaussian (MRW)

$\theta \to 1$

- full surrogate weight

Modification 2: Tempering

Interpolate between surrogate $\propto e^{-\tilde{f}}$ and benign reference density $\propto e^{-g}$

$$\tilde{f} \quad \mapsto \quad \theta\,\tilde{f} + (1-\theta)\,g \qquad \theta \in (0,1]$$

Use Gaussian measure as reference (as in regularisation before):

$$\widetilde{\pi}_x \propto e^{-\Phi_x}, \qquad \Phi_x(y) = \theta\tilde{f}(y) + \frac{\gamma}{2}\|y-x\|^2$$

$\theta \to 0$

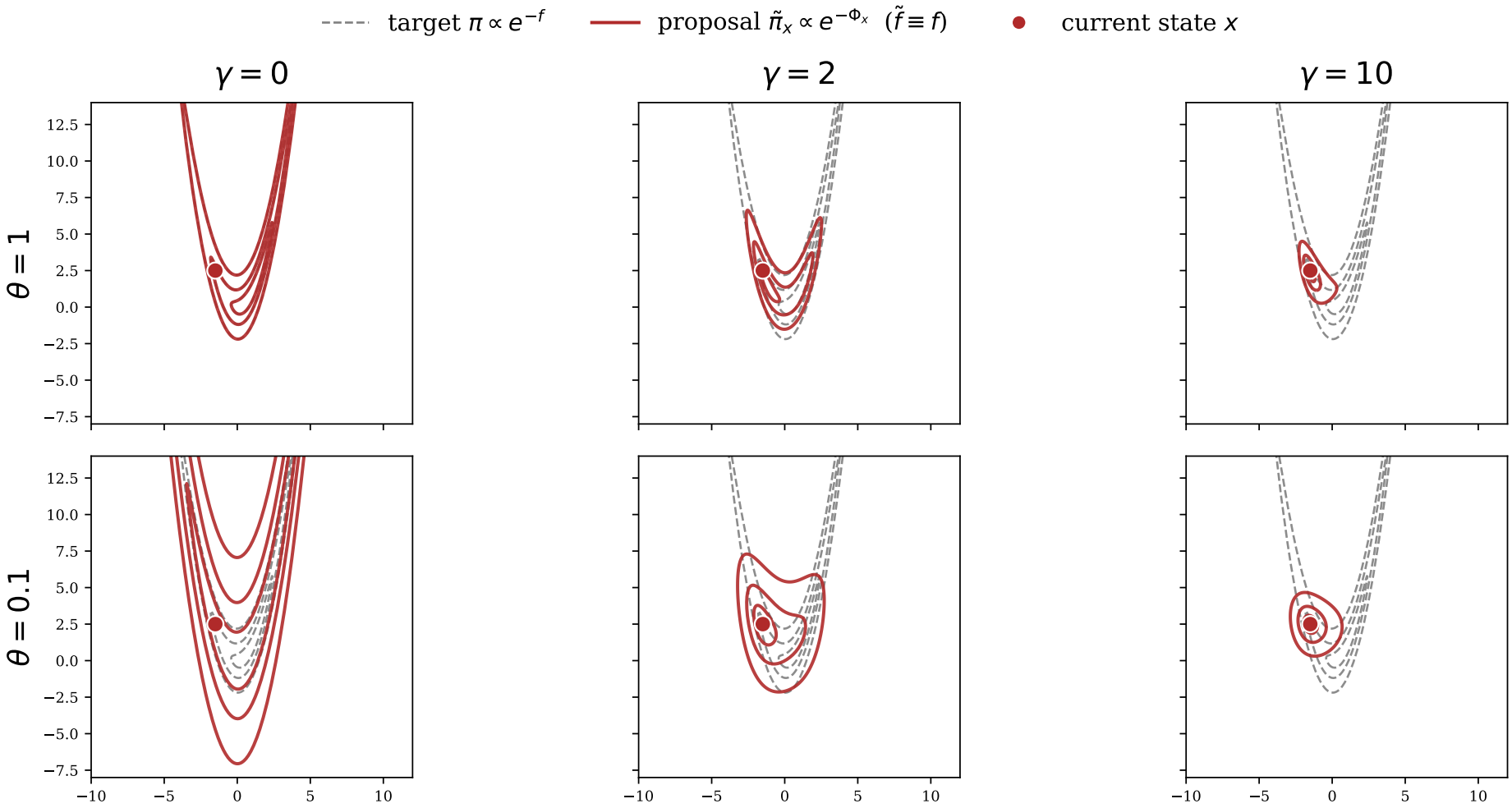
- isotropic Gaussian (MRW)

$\theta \to 1$

- full surrogate weight

θ controls complexity of the geometry

The Proposal Distribution



No Free Lunch

Unfortunately for state-dependent $\widetilde{\pi}_x$ normalisation constants do not cancel:

$$\alpha(x,z) = \min\left\{1,\ \frac{\pi(z)}{\pi(x)}\ \frac{\widetilde{N}_z}{\widetilde{N}_x}\ e^{\theta\,(\tilde{f}(z)-\tilde{f}(x))}\right\}$$

Estimate on the fly using sub-chain states:

$$\frac{\widetilde{N}_z}{\widetilde{N}_x} = \mathbb{E}_{u\sim\Pi_x} e^{\Phi_x(u)-\Phi_z(u)} \approx \frac{1}{J}\sum_{j=1}^{J} e^{\frac{\gamma}{2}(\|u_j-x\|^2-\|u_j-z\|^2)}$$

Stable estimation with low variance, since Exponent is linear in u:

$$\Phi_x(u)-\Phi_z(u) = \frac{\gamma}{2}\Big(\|u-x\|^2-\|u-z\|^2\Big) = \gamma(z-x)^\top u + \frac{\gamma}{2}\Big(\|x\|^2-\|z\|^2\Big)$$

Mixing-time results for log-concave targets

Consider strongly log-concave, log-smooth target Π with density $\pi \propto e^{-f}$:

$$\lambda \mathbb{I} \preceq H_f(x) \preceq L\mathbb{I} \qquad \forall x \in \mathbb{R}^d \qquad H_f : \text{Hessian of } f$$

Condition number: $\kappa := L/\lambda$

Definition. Let $\delta > 0$ and μ some initial measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The *(total variation δ-)mixing time* of a Markov chain w. transition operator T is defined as

$$t_\delta(\mu) := \inf\{n \in \mathbb{N} : \| T^n \mu - \Pi \|_{\mathrm{TV}} < \delta\}.$$

Mixing-time results for log-concave targets

Consider strongly log-concave, log-smooth target Π with density $\pi \propto e^{-f}$:

$$\lambda \mathbb{I} \preceq H_f(x) \preceq L\mathbb{I} \qquad \forall x \in \mathbb{R}^d \qquad H_f : \text{Hessian of } f$$

Condition number: $\kappa := L/\lambda$

Definition. Let $\delta > 0$ and μ some initial measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The *(total variation δ-)mixing time* of a Markov chain w. transition operator T is defined as

$$t_\delta(\mu) := \inf\{n \in \mathbb{N} : \| T^n \mu - \Pi \|_{\mathrm{TV}} < \delta\}.$$

- MALA (uses local gradient): $q(\theta_n|x) = \mathcal{N}(x - h\nabla f(x),\ 2h\,\mathbb{I})$

(Euler–Maruyama on $dX_t = -\nabla f(X_t)\,dt + \sqrt{2}\,dW_t$)

Theorem. [Dwivedi, Chen, Wainwright, Yu, JMLR 20, 2019]

Mixing time $\mathcal{O}(\kappa d)$ from warm start.

Mixing-time results for log-concave targets

Consider strongly log-concave, log-smooth target Π with density $\pi \propto e^{-f}$:

$$\lambda \mathbb{I} \preceq H_f(x) \preceq L\mathbb{I} \qquad \forall x \in \mathbb{R}^d \qquad H_f : \text{Hessian of } f$$

Condition number: $\kappa := L/\lambda$

Definition. Let $\delta > 0$ and μ some initial measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The *(total variation δ-)mixing time* of a Markov chain w. transition operator T is defined as

$$t_\delta(\mu) := \inf\{n \in \mathbb{N} : \| T^n \mu - \Pi \|_{\mathrm{TV}} < \delta\}.$$

- MALA (uses local gradient): $q(\theta_n|x) = \mathcal{N}(x - h\nabla f(x),\ 2h\,\mathbb{I})$

(Euler–Maruyama on $dX_t = -\nabla f(X_t)\,dt + \sqrt{2}\,dW_t$)

Theorem. [Dwivedi, Chen, Wainwright, Yu, JMLR 20, 2019]

Mixing time $\mathcal{O}(\kappa d)$ from warm start.

- Similar results for HMC [Chen, Dvivedi, Wainwright, Yu, JMLR 21, 2020]

Mixing-time results for log-concave targets

Consider strongly log-concave, log-smooth target Π with density $\pi \propto e^{-f}$:

$$\lambda \mathbb{I} \preceq H_f(x) \preceq L\mathbb{I} \qquad \forall x \in \mathbb{R}^d \qquad H_f : \text{Hessian of } f$$

Condition number: $\kappa := L/\lambda$

Definition. Let $\delta > 0$ and μ some initial measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The *(total variation δ-)mixing time* of a Markov chain w. transition operator T is defined as

$$t_\delta(\mu) := \inf\{n \in \mathbb{N} : \| T^n \mu - \Pi \|_{\mathrm{TV}} < \delta\}.$$

- MALA (uses local gradient): $q(\theta_n|x) = \mathcal{N}(x - h\nabla f(x),\ 2h\,\mathbb{I})$

(Euler–Maruyama on $dX_t = -\nabla f(X_t)\,dt + \sqrt{2}\,dW_t$)

Theorem. [Dwivedi, Chen, Wainwright, Yu, JMLR 20, 2019]

Mixing time $\mathcal{O}(\kappa d)$ from warm start.

- Similar results for HMC [Chen, Dvivedi, Wainwright, Yu, JMLR 21, 2020]
- For comparison: MRW has a mixing time of $\mathcal{O}(\kappa^2 d)$.

Main Result

Theorem (Kutri, **RS**, 2026+; Informal).

Let surrogate $\tilde{f}$ be m-strongly convex, M-smooth with $\lambda \leq m \leq M \leq L$. Under the following assumptions

(i) J is large enough to approximate $\tilde{\Pi}_x$ (sub-chain length)

(ii) $\nabla \tilde{f}$ is good inside a high-probability region $\mathcal{K}$: (gradient fidelity)

$$\sup_{x \in \mathcal{K}} \|\nabla \tilde{f}(x) - \nabla f(x)\|^2 \lesssim L \frac{\max\{\kappa, d\}}{d}$$

(iii) parameter choice: $\theta = \frac{1}{2}, \quad \gamma \gtrsim L \max\{\kappa, d\}$ (localisation)

then the chain (started from a β-warm μ) mixes in

$$\mathcal{O}\left(\log(\tfrac{2\beta}{\delta})\, \kappa \max\{\kappa, d\}\right) \text{ steps.}$$

Main Result (Discussion)

- Surrogate density $\widetilde{\pi}_x$ has condition number $\tilde{\kappa}_x \leq \frac{\gamma+L/2}{\gamma+\lambda/2} \ll \kappa$

  $\rightarrow$ significantly faster mixing
- For $d \geq \kappa$: recovers MALA rate $\mathcal{O}(\kappa d)$ — without gradients!
- For $d < \kappa$: rate $\mathcal{O}(\kappa^2)$ — independent of d!
- β-warmness through burn-in on surrogate chain

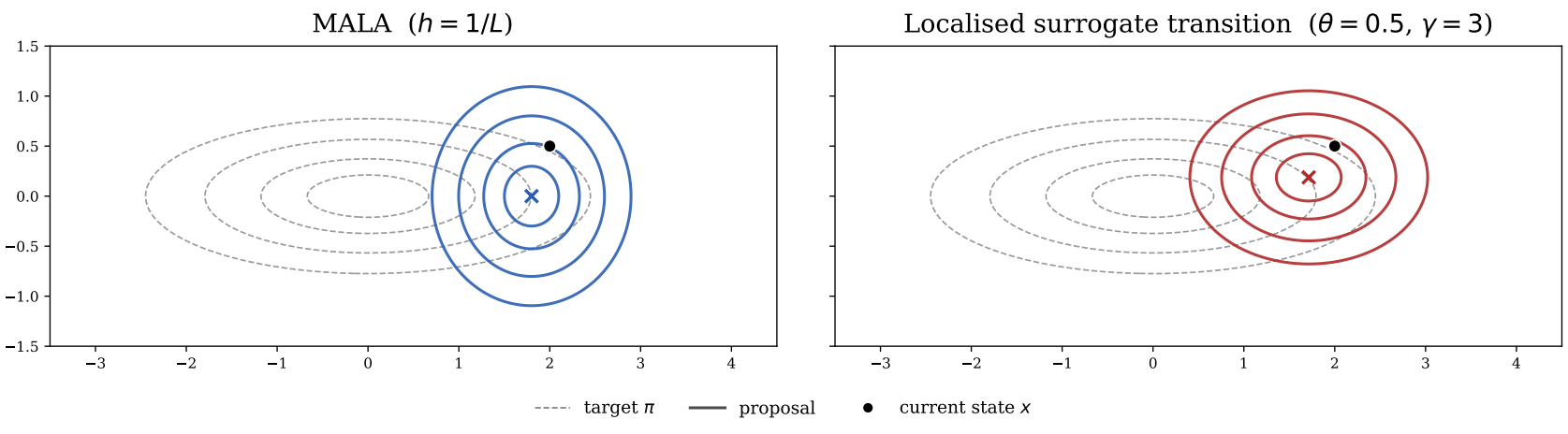
Main Result (Discussion)

- Surrogate density $\widetilde{\pi}_x$ has condition number $\tilde{\kappa}_x \leq \frac{\gamma+L/2}{\gamma+\lambda/2} \ll \kappa$

  $\rightarrow$ significantly faster mixing
- For $d \geq \kappa$: recovers MALA rate $\mathcal{O}(\kappa d)$ — without gradients!
- For $d < \kappa$: rate $\mathcal{O}(\kappa^2)$ — independent of d!
- β-warmness through burn-in on surrogate chain

Intuition why rates are comparable to MALA and why IAT is even better:

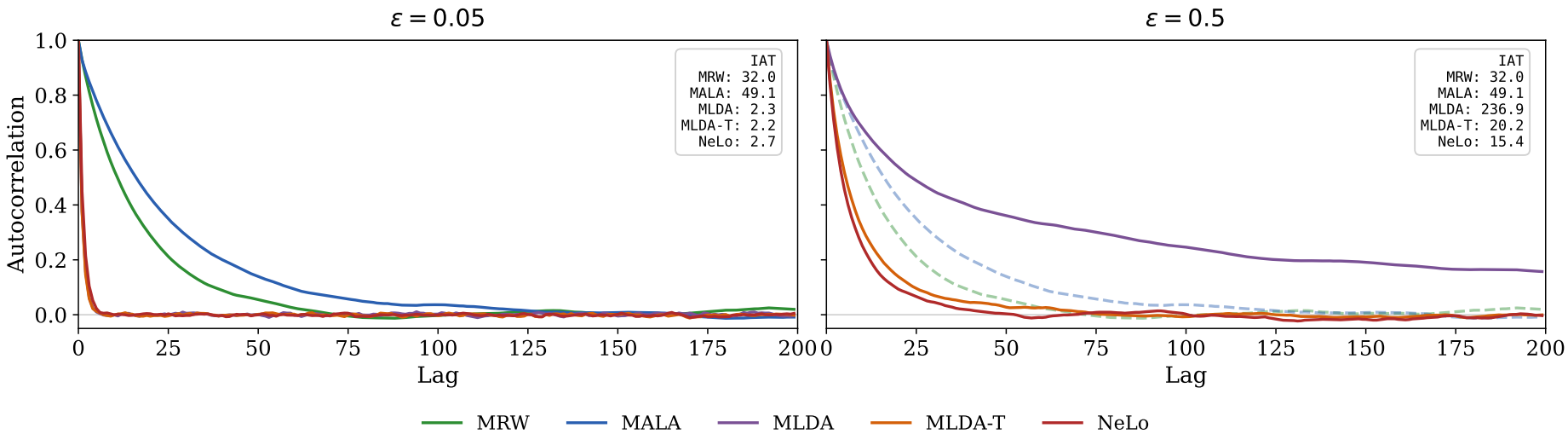
(see below)



Autocorrelation

Example. π and $\widetilde{\pi}$ 2D-Gaussians with $\kappa = 50$

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 1/\kappa & 0 \\ 0 & 1 \end{pmatrix}, \quad \text{and} \quad \tilde{\mu} = \begin{pmatrix} \varepsilon \\ \varepsilon \end{pmatrix}, \widetilde{\Sigma} = \Sigma + \frac{\varepsilon}{2\sqrt{\kappa}} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$



Acceptance tuned to optima: $(J = 20, \theta = 0.5)$

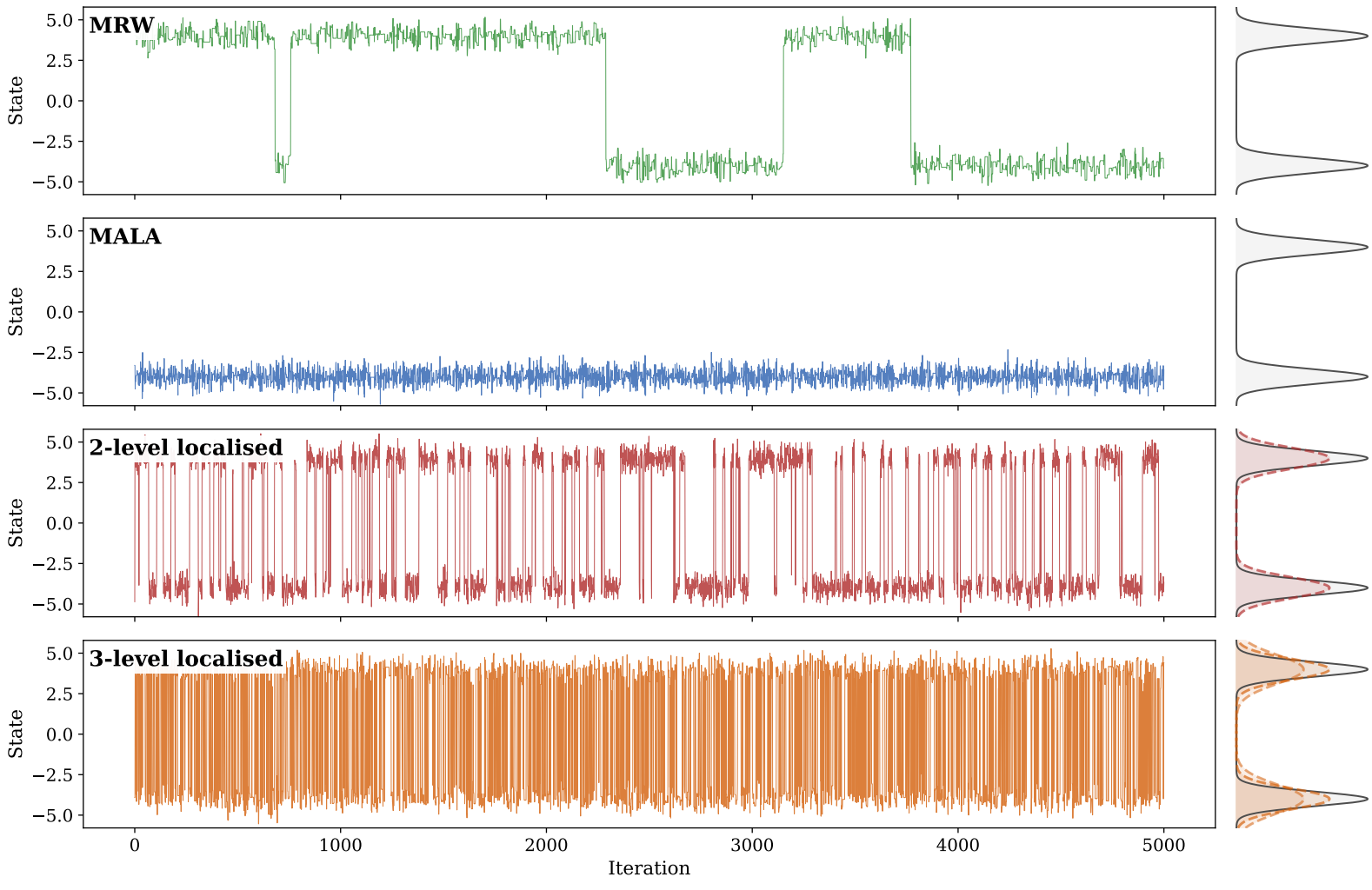
MRW: ~ 0.25 [Gelman, Gilks, Roberts, 1997]

MALA: ~ 0.5 [Roberts, Rosenthal, 1998]

MLDA, NeLo: ~ 0.5 (empirically)

Tempering and Multimodality

Example. Mixture of $\mathcal{N}(-4, 0.2\mathbb{I})$ and $\mathcal{N}(4, 0.2\mathbb{I})$ (ratio peak/valley $\approx 10^{18}$)



IAT=572.2

IAT=1.4(??)

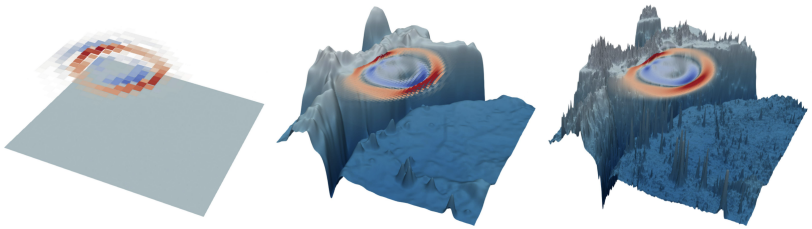
IAT=25.3

IAT=2.7

(2-level: $\theta = 0.5, J = 20$; 3-level: $\theta_0 = 0.25, J_0 = 20$; $\theta_1 = 0.5, J_1 = 10$)

Concluding Remarks

- **MLDA – Scalable Bayesian inference for complex problems!**
- Large additional gains through a simple **adaptive error model**
- Can use general model hierarchy: 'multi-fidelity' or spatial adaptivity!

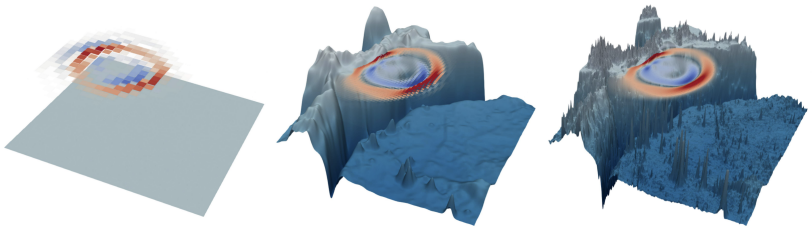


e.g. bathymetry in tsunami simulation

- Robustness & **theoretical guartantees** via localisation & tempering

Concluding Remarks

- **MLDA – Scalable Bayesian inference for complex problems!**
- Large additional gains through a simple **adaptive error model**
- Can use general model hierarchy: 'multi-fidelity' or spatial adaptivity!



e.g. bathymetry in tsunami simulation

- Robustness & **theoretical guartantees** via localisation & tempering

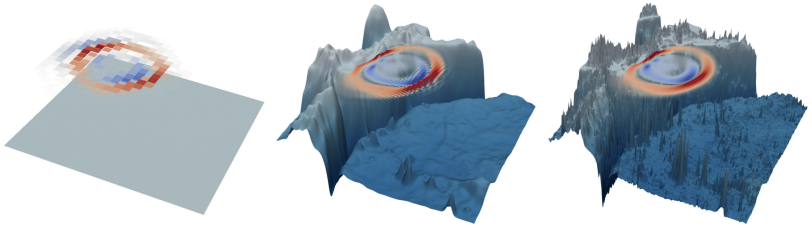
- Open source easy-to-use code: `github.com/mikkelbue/tinyDA`
- Easy integration of all codes with sophisticated back-ends in

  `UM-Bridge` `github.com/um-bridge`

  [Seelinger et al., *Democratizing Uncertainty Quantification*, JCP **521**, 2025]

Concluding Remarks

- **MLDA – Scalable Bayesian inference for complex problems!**
- Large additional gains through a simple **adaptive error model**
- Can use general model hierarchy: 'multi-fidelity' or spatial adaptivity!



e.g. bathymetry in tsunami simulation

- Robustness & **theoretical guartantees** via localisation & tempering

- Open source easy-to-use code: `github.com/mikkelbue/tinyDA`
- Easy integration of all codes with sophisticated back-ends in

  `UM-Bridge` `github.com/um-bridge`

  [Seelinger et al., *Democratizing Uncertainty Quantification*, JCP **521**, 2025]

- Complexity analysis & a posteriori error control (with Fox)
- Multi-index delayed acceptance (with Haji-Ali, Seelinger, Teckentrup,. . .)

References

- Liu, Monte Carlo Strategies in Scientific Computing, Springer, New York, 2004
- Christen, Fox, *MCMC using an approximation*, J. Comp. Graph. Stat., **14**, 2005
- Dodwell, Ketelsen, **RS**, Teckentrup, *A hierarchical multilevel MCMC algorithm with applications to UQ in subsurface flow*, SIAM/ASA J. Uncertain. Q., **3**, 2015
- Lykkegaard, Mingas, **RS**, Fox, Dodwell, *Multilevel delayed acceptance MCMC with an adaptive error model in PyMC3*, NeurIPS, 2020
- Seelinger, Reinarz, Rannabauer, Bader, Bastian, **RS**, *High performance UQ with parallelized multilevel Markov chain Monte Carlo*, Proceed. SC'21, No. 75, 2021
- **Lykkegaard, Dodwell, Fox, Mingas, RS**, **Multilevel delayed acceptance MCMC**, SIAM/ASA J. Uncertain. Q., **10**, 2023
- Cui, Detommaso, **RS**, *Multilevel dimension-independent likelihood-informed MCMC for large-scale inverse problems*, Inverse Prob., **40**, 2024
- Seelinger, Reinarz, Lykkegaard, . . . , Dodwell, **RS**, *Democratizing Uncertainty Quantification*, J. Comput. Phys., **521**, 2025
- **Kutri, RS**, **Localized surrogate transition: Fast mixing MCMC without gradients**, in preparation, 2026+