# Hierarchical Off-Diagonal Low-Rank Approximation for Hessians in Bayesian Inference

Tucker Hartland[1]

joint work with:
Noémi Petra[1], Georg Stadler[2],
Ruanui Nicholson[3],

[1]School of Natural Sciences, University of California, Merced

[2]Courant Institute of Mathematical Sciences, New York University

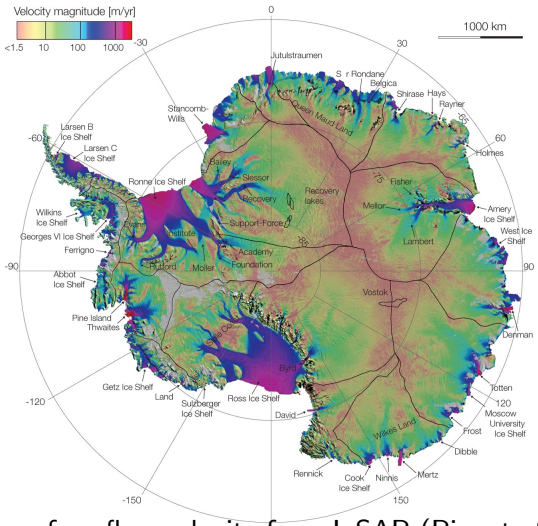
[3]Department of Engineering Science, University of Auckland

SIAM Conference on Computational
Science and Engineering
February 27, 2019

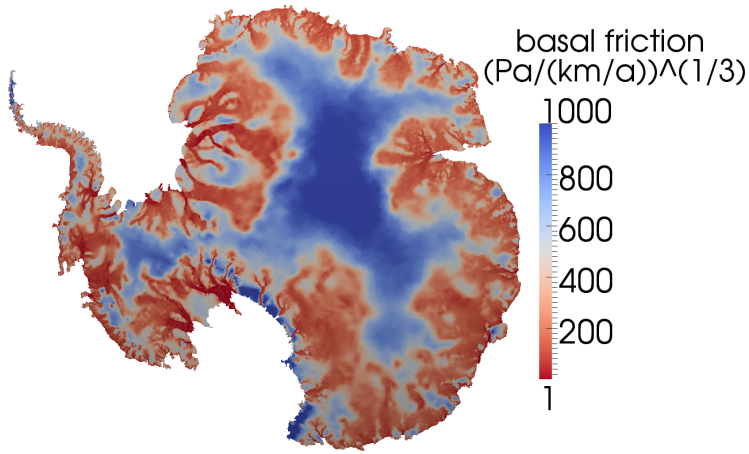
Research Supported By NSF-Grant CAREER-1654311

# Outline

- Motivation: infer basal friction field in an ice sheet model and characterize the associated uncertainties.



Observed surface flow velocity from InSAR (Rignot et. al, 2011)



Antarctic ice sheet inversion for the basal friction parameter field from InSAR surface velocities

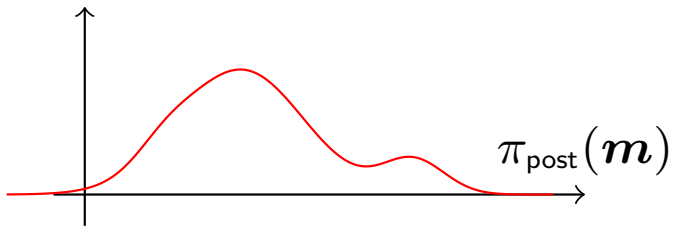
- Bayesian formulation of an inverse problem.
- The role and importance of the Hessian (of the likelihood) in Bayesian inference.
- Hierarchical off-diagonal low-rank (HODLR) approximation of the Hessian.
- Numerical example: inference of membrane log stiffness field in a Poisson model problem.

# Bayesian approach to inverse problems

Inverse problem: given (possibly noisy) data $\boldsymbol{d}$ and (a possibly uncertain) model $\boldsymbol{f}$, infer parameters $\boldsymbol{m}$ that characterize the model, i.e.,

$$\boldsymbol{f}(\boldsymbol{m}) + \boldsymbol{e} = \boldsymbol{d}$$

Interpret $\boldsymbol{m}$, $\boldsymbol{d}$ as random variables; solution of inverse problem is the "posterior" probability density function $\pi_{\text{post}}(\boldsymbol{m})$ for $\boldsymbol{m}$:




## Remarks:

- Bayesian framework quantifies uncertainty in the inverse solution, given uncertainty in the prior, the data, and the model.
- Prior incorporates known information and in infinite dimensions chosen to act as a regularization.
- Bayesian solution is a probability density function in as many dimensions as the number of parameters.

# Exploring the posterior

If the prior is Gaussian with mean $\boldsymbol{m}_{\text{pr}}$ and covariance $\boldsymbol{\Gamma}_{\text{prior}}$, and we assume additive Gaussian noise in the measurements, i.e., $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}_{\text{noise}})$, then using Bayes theorem we obtain the posterior pdf:

$$\pi_{\text{post}}(\boldsymbol{m}) \propto \exp\left(-\tfrac{1}{2} \parallel \boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d} \parallel^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}} - \tfrac{1}{2} \parallel \boldsymbol{m} - \boldsymbol{m}_{\text{pr}} \parallel^2_{\boldsymbol{\Gamma}^{-1}_{\text{prior}}}\right)$$

The "maximum a posteriori" point is

$$\begin{aligned} \boldsymbol{m}_{\text{MAP}} &\overset{\text{def}}{=} \arg \max_{\boldsymbol{m}} \pi_{\text{post}}(\boldsymbol{m}) \\ &= \arg \min_{\boldsymbol{m}} \frac{1}{2} \|\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}} + \frac{1}{2} \|\boldsymbol{m} - \boldsymbol{m}_{\text{pr}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{prior}}} \end{aligned}$$

$\Rightarrow$ deterministic inverse problem with appropriate weighted norms!

Details in: J. Kaipio and E. Somersalo, Statistical and Computational Inverse Problems, 2005

# The Hessian (of the negative log posterior) plays a critical role in inverse problems

- Its spectral properties characterize the degree of ill-posedness.
- The Hessian drives Newton-type optimization algorithms for solving the inverse problem.
- The inverse of the Hessian locally characterizes the uncertainty in the solution of the inverse problem (under the Gaussian assumption, it is precisely the posterior covariance matrix).
- Goal: rapidly perform linear algebraic operations, i.e., manipulation of the Hessian (and its inverse and inverse square root) actions needed by sampling or CG solvers, hence seek approaches to approximate the Hessian(-applies).
- These approximations will then be used as pre-conditioners, and to build MCMC proposals based on local Gaussian approximations.

# Scalability of the inverse solver

## Inexact Newton-CG method applied to an ice sheet inverse problem

| #sdof | #pdof | #N | #CG | avgCG | #Stokes |
|---|---|---|---|---|---|
| 95,796 | 10,371 | 42 | 2718 | 65 | 7031 |
| 233,834 | 25,295 | 39 | 2342 | 60 | 6440 |
| 848,850 | 91,787 | 39 | 2577 | 66 | 6856 |
| 3,372,707 | 364,649 | 39 | 2211 | 57 | 6193 |
| 22,570,303 | 1,456,225 | 40 | 1923 | 48 | 5376 |

- **#sdof**: number of degrees of freedom for the state variables;
- **#pdof**: number of degrees of freedom for the inversion parameter field;
- **#N**: number of Newton iterations;
- **#CG**, **avgCG**: total and average (per Newton iteration) number of CG iterations;
- **#Stokes**: total number of linear(ized) Stokes solves (from forward, adjoint, and incremental forward and adjoint problems)

Details in the SIAG CSE19 Best Paper: T. Isaac, N. Petra, G. Stadler, and O. Ghattas. *Scalable and efficient algorithms for the propagation of uncertainty from data through inference to prediction for large-scale problems, with application to flow of the Antarctic ice sheet*, Journal of Computational Physics, 296, 348-368 (2015).

# MCMC sampling: stochastic Newton

## Performance results / Convergence diagnostics

| | MPSRF | IAT | ESS | MSJ | ARR | #Stokes | time (s) |
|---|---|---|---|---|---|---|---|
| SN | 1.348 | 600 | 875 | 64 | 2 | 8400 | 420 |

- **MPSRF:** multivariate potential scale reduction factor
- **IAT:** integrated autocorrelation time
- **ESS:** effecitive sample size
- **MSJ:** mean squared jump distance
- **ARR:** average rejection rate
- **#Stokes:** # of Stokes solves per independent sample
- **time:** time per independent sample

- **Statistics:** 21 parallel chains (each 25k); # samples: 525k; dof: 139; rank Hessian: 15

Proposal density:

$$q(\boldsymbol{m}_k, \boldsymbol{m}) \propto \exp\left(-\frac{1}{2}\left\langle \boldsymbol{m} - \left(\boldsymbol{m}_k - \boldsymbol{H}^{-1}\boldsymbol{g}_k\right), \boldsymbol{H}\left(\boldsymbol{m} - \left(\boldsymbol{m}_k - \boldsymbol{H}^{-1}\boldsymbol{g}_k\right)\right)\right\rangle_{\boldsymbol{M}}\right)$$

Details in: N. Petra, J. Martin, G. Stadler, O. Ghattas. *A computational framework for infinite-dimensional Bayesian inverse problems: Part II. Stochastic Newton MCMC with application to ice sheet inverse problems*, SIAM Journal on Scientific Computing, 2014

# The role of the Hessian in Stochastic Newton proposal sampling and deterministic inversion

To efficiently sample from the Stochastic Newton proposal density we need computationally effective means to,

- apply $\boldsymbol{H}^{-1}$ to vectors in order to compute the mean, $\boldsymbol{m}_k - \boldsymbol{H}^{-1}\boldsymbol{g}_k$;
- apply $\boldsymbol{H}^{-1/2}$ the square root of the covariance operator $\boldsymbol{H}^{-1}$, to vectors.

To efficiently solve the deterministic inverse problem by means of a Newton-CG method we need an estimate of the Hessian that will serve as a preconditioner.

**Algorithm 1** Inexact Newton-CG

**while** Not sufficiently near a minimizer **do**
  Solve($\boldsymbol{H}\boldsymbol{p} = -\boldsymbol{g}$)
  $\vdots$
  $\boldsymbol{m}^{(k+1)} = \boldsymbol{m}^{(k)} + \alpha\boldsymbol{p}$
**end while**

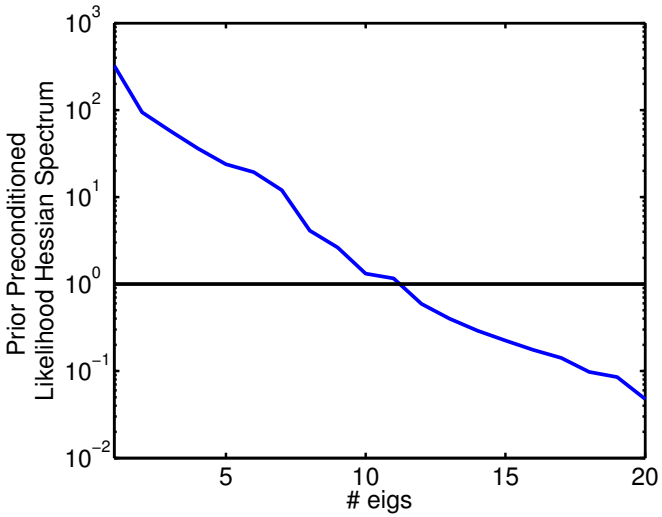
# Low rank approximation of the Hessian of the likelihood

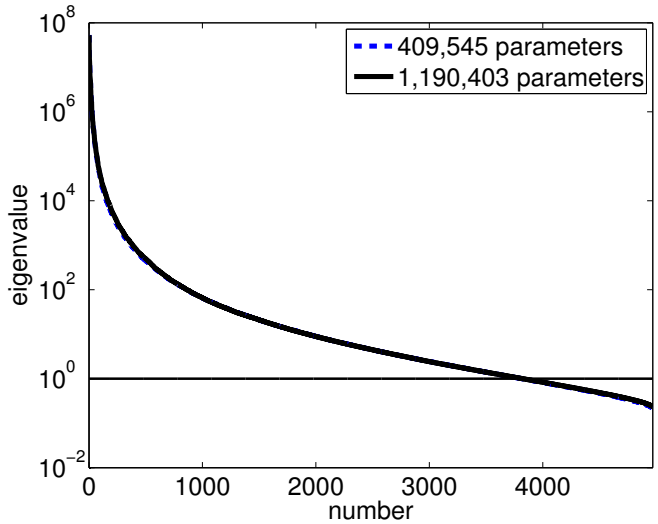
Under the assumption of Gaussian noise and linearized parameter-to-observable map (i.e., Gaussian posterior)

- Invoke low-rank approximation given by the inverse of the Hessian:

$$\boldsymbol{\Gamma}_{\text{post}} = \boldsymbol{H}^{-1} = \left(\boldsymbol{F}^T \boldsymbol{\Gamma}_{\text{noise}}^{-1} \boldsymbol{F} + \boldsymbol{\Gamma}_{\text{prior}}^{-1}\right)^{-1} \approx \boldsymbol{\Gamma}_{\text{prior}}^{1/2} (\boldsymbol{V}_r \boldsymbol{\Lambda}_r \boldsymbol{V}_r^T + \boldsymbol{I})^{-1} \boldsymbol{\Gamma}_{\text{prior}}^{1/2}$$

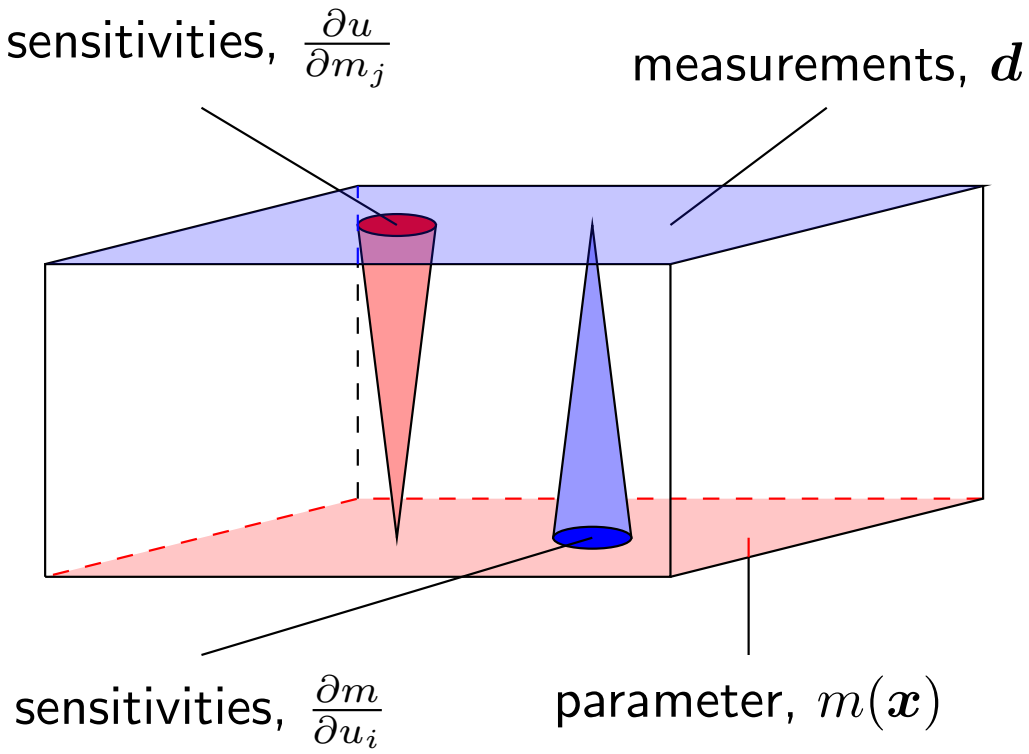
  - $\boldsymbol{V}_r$ and $\boldsymbol{\Lambda}_r$ are the eigenvectors/values of $\boldsymbol{F}^T \boldsymbol{\Gamma}_{\text{noise}}^{-1} \boldsymbol{F} \boldsymbol{v}_i = \lambda_i \boldsymbol{\Gamma}_{\text{prior}}^{-1} \boldsymbol{v}_i$.
  - $\boldsymbol{F}$ is the Jacobian matrix of the parameter to observable map $\boldsymbol{f}(\boldsymbol{m})$.
- Spectrum of the prior-preconditioned likelihood Hessian for the Arolla glacier (139 parameters, left) and Antarctica (1.19M parameters, right).



# Exploiting structure

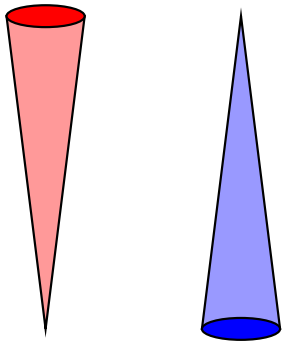
## Taking advantage of the parameter to observable mapping



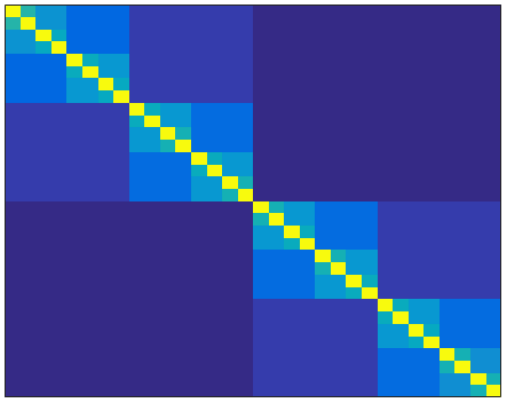

- We expect (hope) for narrow sensitivities.
- $\Rightarrow$ We hope (for a suitable ordering of the discrete degrees of freedom) that the Hessian $\boldsymbol{H}$, will be concentrated on the diagonal.

# Exploiting structure

## Taking advantage of the parameter to observable mapping



- We expect (hope) for narrow sensitivities.
- $\Rightarrow$ We hope that $\boldsymbol{H}$ is concentrated on the diagonal.
- We wish to maintain only *useful* information in $\boldsymbol{H}$.
- $\Rightarrow$ We expect *useful* information concentrated in certain blocks of $\boldsymbol{H}$.
- Apply techniques of hierarchical ($\mathcal{H}$) matrices: Specifically hierarchical off-diagonal low-rank (HODLR) matrices.



# $\mathcal{H}$-matrices

## Hierarchical off-diagonal low-rank (HODLR) matrices

$$\boldsymbol{H}_{\text{HODLR}} = \begin{bmatrix} \boldsymbol{A}_1^{(1)} & \boldsymbol{B}_1^{(1)} \\ \boldsymbol{B}_2^{(1)} & \boldsymbol{A}_2^{(1)} \end{bmatrix}$$

$$= \left[\begin{array}{cc|c} \boldsymbol{A}_1^{(2)} & \boldsymbol{B}_1^{(2)} & \boldsymbol{B}_1^{(1)} \\ \boldsymbol{B}_2^{(2)} & \boldsymbol{A}_2^{(2)} & \\ \hline \boldsymbol{B}_2^{(1)} & & \begin{matrix} \boldsymbol{A}_3^{(2)} & \boldsymbol{B}_3^{(2)} \\ \boldsymbol{B}_4^{(2)} & \boldsymbol{A}_4^{(2)} \end{matrix} \end{array}\right]$$

$$= \left[\begin{array}{cc|c|c} \boldsymbol{A}_1^{(3)} & \boldsymbol{B}_1^{(3)} & \boldsymbol{B}_1^{(2)} & \\ \boldsymbol{B}_2^{(3)} & \boldsymbol{A}_2^{(3)} & & \boldsymbol{B}_1^{(1)} \\ \hline \boldsymbol{B}_2^{(2)} & & \begin{matrix} \boldsymbol{A}_3^{(3)} & \boldsymbol{B}_3^{(3)} \\ \boldsymbol{B}_4^{(3)} & \boldsymbol{A}_4^{(3)} \end{matrix} & \\ \hline \multicolumn{2}{c|}{\boldsymbol{B}_2^{(1)}} & \begin{matrix} \boldsymbol{A}_5^{(3)} & \boldsymbol{B}_5^{(3)} \\ \boldsymbol{B}_6^{(3)} & \boldsymbol{A}_6^{(3)} \end{matrix} & \boldsymbol{B}_3^{(2)} \\ & & \boldsymbol{B}_4^{(2)} & \begin{matrix} \boldsymbol{A}_7^{(3)} & \boldsymbol{B}_7^{(3)} \\ \boldsymbol{B}_8^{(3)} & \boldsymbol{A}_8^{(3)} \end{matrix} \end{array}\right]$$

- Low rank block

Details in: J. Ballani, D. Kressner. *Matrices with Hierarchical Low-Rank Structures*, Exploiting Hidden Structure in Matrix Computations: Algorithms and Applications, 2016

# $\mathcal{H}$-matrices

## Hierarchical compression, direct solve

**Desired characteristics of the preconditioner for the Newton-CG system:**

1. accurately approximates the Hessian,
2. the application of the inverse to vectors is computationally inexpensive.

**Compression:**

- Apply $\boldsymbol{H}$ to random vectors with specified blocks of zeros, in order to sample the column spaces of the off-diagonal blocks.

  The samples are likely aligned with dominant modes, thus allowing the generation of low rank approximations of the off-diagonal blocks.

  Details in: P.G. Martinsson, *Compressing Rank-Structured Matrices Via Randomized Sampling*, SIAM Journal on Scientific Computing, 2016

**Direct Solve:**

- At $\mathcal{O}\left(N \log^2 N\right)$ cost one obtains a factorization of a HODLR matrix.
- Employing the factorization one has a $\mathcal{O}\left(N \log N\right)$ means of applying $\left(\boldsymbol{H}_{\text{HODLR}}\right)^{-1}$ to vectors.

  Details in: S. Ambikasaran, E. Darve, *An $\mathcal{O}(N \log N)$ Fast Direct Solver for Partial Hierarchically Semi-Separable Matrices*, Springer Journal of Scientific Computing, 2013

# Example problem

## Log stiffness coefficient field inversion in an elliptic PDE

**Goal:** determine the field $m(\boldsymbol{x})$ such that the solution $u(\boldsymbol{x})$, of the elliptic PDE is closest to the observation data, $u_{\text{obs}}(\boldsymbol{x})$.

$$\min_m J[m] := \underbrace{\frac{1}{2}\int_\Omega \left(u(\boldsymbol{x}) - u_{\text{obs}}(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x}}_{\text{Misfit}} + \underbrace{\frac{\gamma}{2}\int_\Omega \left(\boldsymbol{\nabla} m \cdot \boldsymbol{\nabla} m\right) \mathrm{d}\boldsymbol{x}}_{\text{Regularization}}$$

Where $u(\boldsymbol{x})$ is uniquely defined by $m(\boldsymbol{x})$, as the solution of the following PDE:

$$\begin{aligned} -\boldsymbol{\nabla} \cdot \left(\exp(m)\, \boldsymbol{\nabla} u\right) &= f(\boldsymbol{x}), \quad \boldsymbol{x} \text{ in } \Omega \\ u(\boldsymbol{x}) &= 0, \qquad \boldsymbol{x} \text{ on } \Gamma \end{aligned}$$

- $u(\boldsymbol{x})$, state, solution of PDE for a given $m(\boldsymbol{x})$,
- $u_{\text{obs}}(\boldsymbol{x})$, observation data,
- $m(\boldsymbol{x})$, log stiffness parameter field,
- $\gamma \geq 0$, regularization parameter,
- $f(\boldsymbol{x})$, source term,
- $\Gamma$, boundary of spatial domain $\Omega$.

# Estimating the minimizer of a deterministic PDE constrained optimization problem

**Method of Solution**

1. Derive 1st and 2nd order derivative information via adjoints.
2. Setup an inexact Newton-CG system.

---

**Algorithm 2** Inexact Newton-CG

---

**while** Not sufficiently near a minimizer **do**

Solve($\boldsymbol{H}\boldsymbol{p} = -\boldsymbol{g}$)

$\vdots$

$\boldsymbol{m}^{(k+1)} = \boldsymbol{m}^{(k)} + \alpha \boldsymbol{p}$

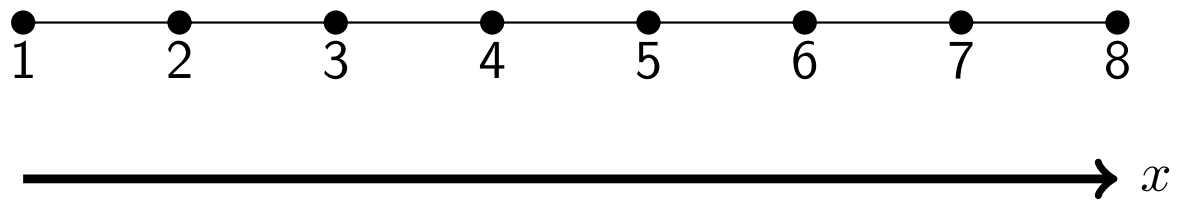
**end while**

---

**Computational Cost:** an application of the Hessian, $\boldsymbol{H}$ to a vector requires two PDE solves.

# Example problem in one dimension

## Ordering the degrees of freedom



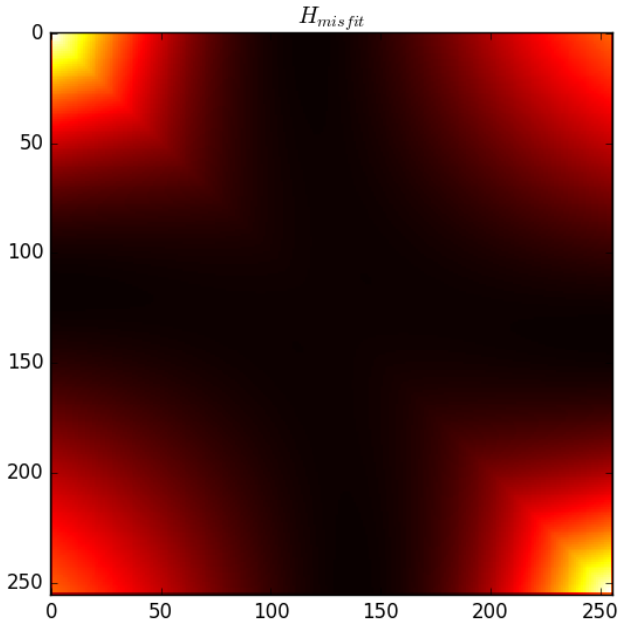

Ordering in 1d naturally respects locality :

- If points $\boldsymbol{x}_i, \boldsymbol{x}_j \in \Omega$ are close to each other then $|i-j|$ is relatively small.

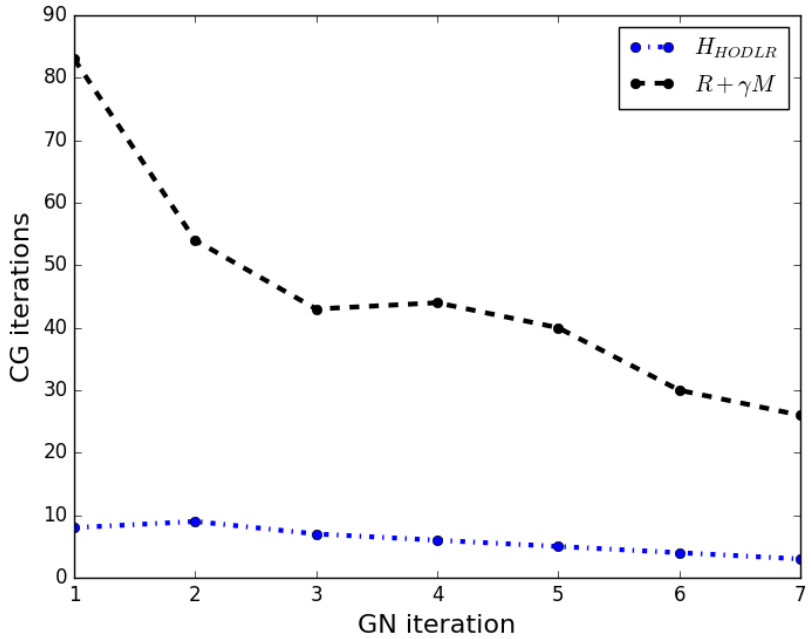
  If this was *not true*, then the Hessian would have large magnitude elements far from the diagonal.
- If $i, j$ are such that $|i-j|$ is relatively small then $||\boldsymbol{x}_i - \boldsymbol{x}_j||$ is small.

  If this was *not true*, then the Hessian would have small magnitude elements near the diagonal.

# HODLR preconditioning in one dimension




Matrix plot of the Hessian



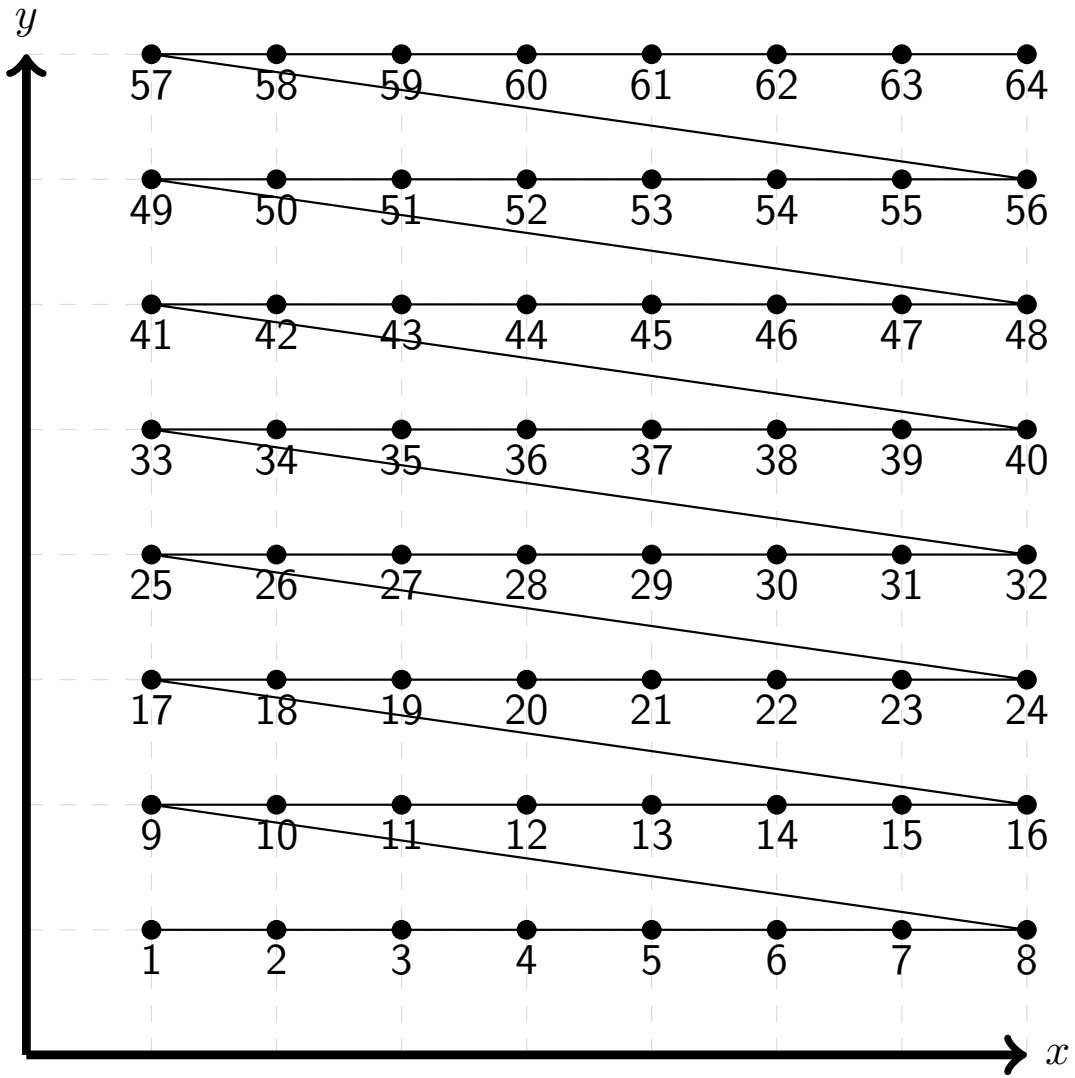

CG iterations vs Newton iteration
$256$ dof, $3$ levels of H-refinement, OD rank $k = 20$

1. The structure of the Hessian indicates that it can be well approximated by a HODLR compression.
2. The HODLR compressed Hessian serves as an effective preconditioner for the

# Example problem in two dimensions

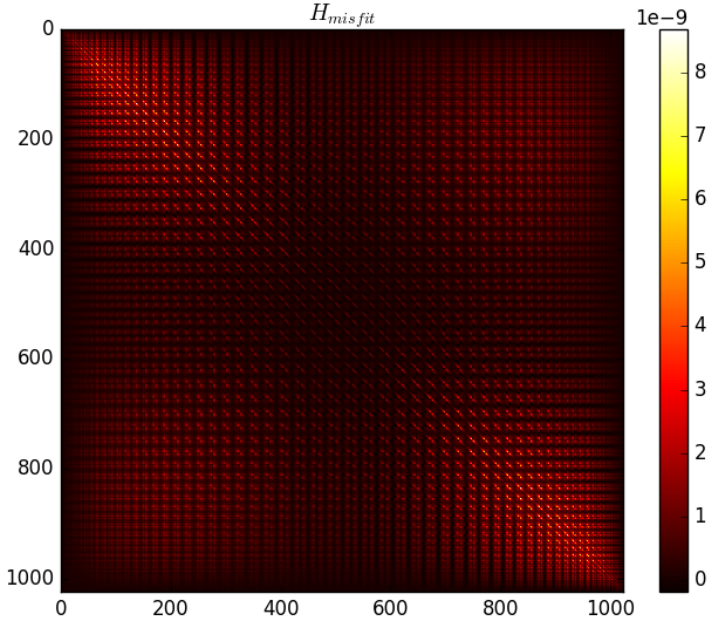
## Ordering the degrees of freedom by row

Ordering in $2$d does *not* naturally respect locality.



# Hessian with a row ordered basis

Naive row ordering basis leads to Hessians unsuitable for hierarchical compression




Matrix plot of the Hessian

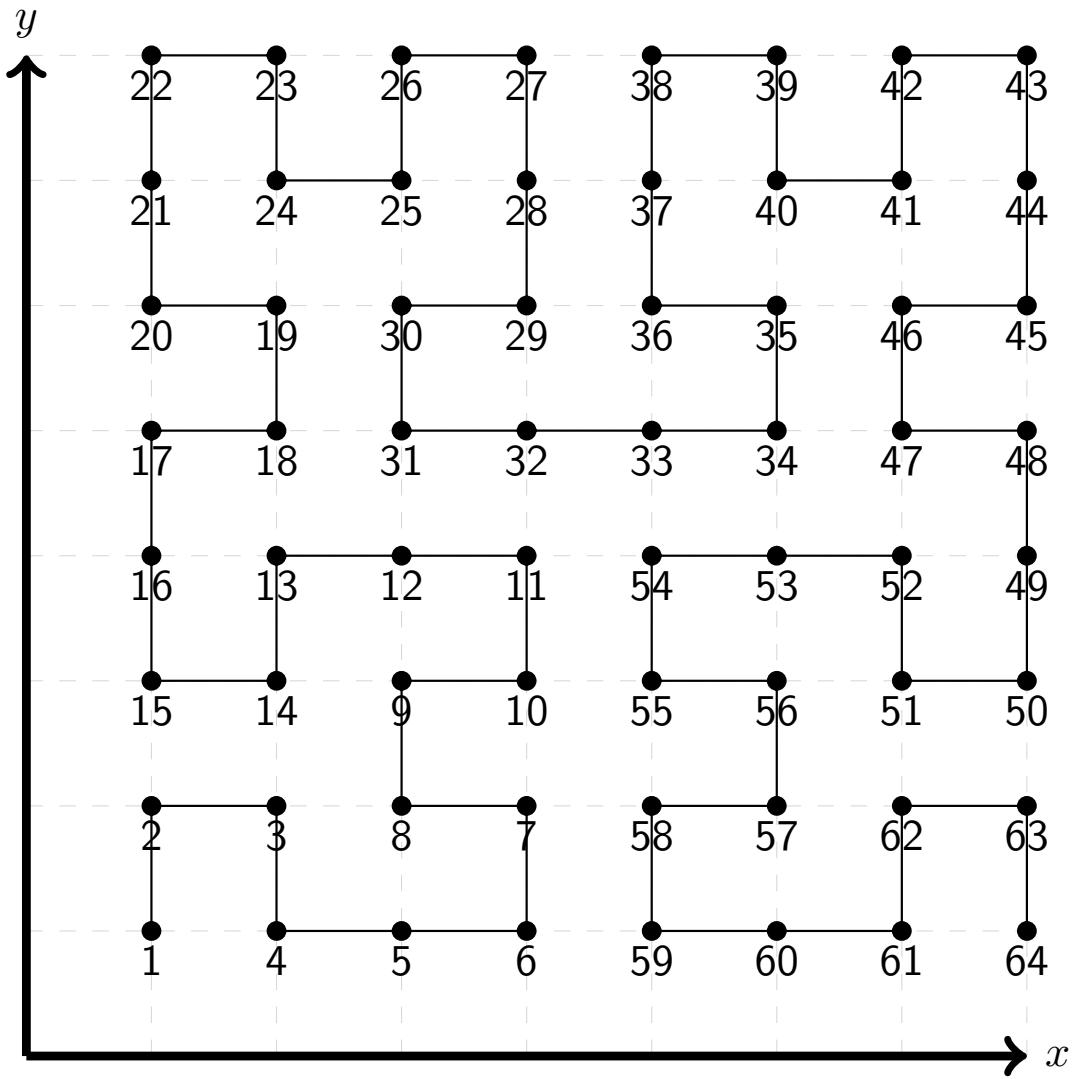
# of Conjugate-Gradient iterations consistently equal to the problem size

When using a default ordering of the degrees of freedom in $2$d:

1. the Hessian does *not* possess HODLR structure,
2. the HODLR compressed Hessian *fails* to effectively precondition the Newton Conjugate-Gradient system.

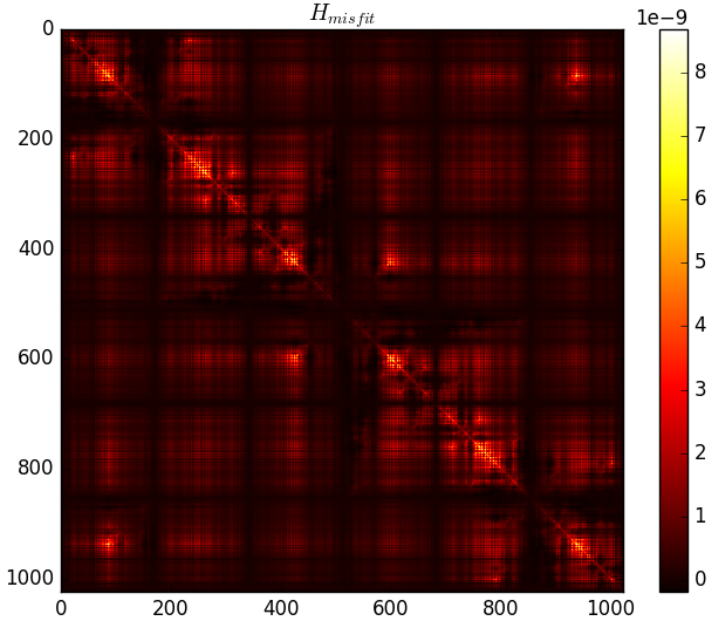
# Example problem in two dimensions

## Ordering the degrees of freedom according to a Hilbert curve

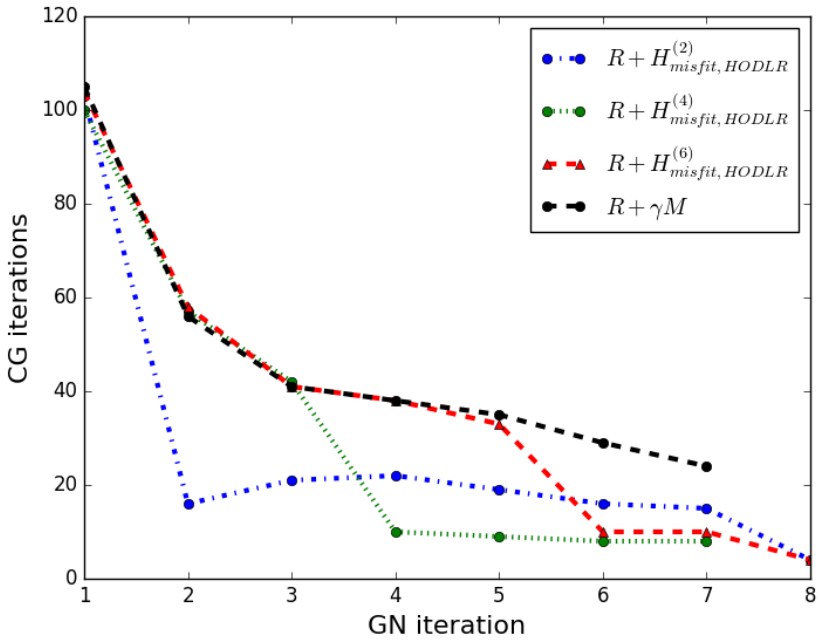


# Hessian with a Hilbert curve ordered basis

Hilbert curve ordering leads to Hessians more suitable for hierarchical compression




Matrix plot of the Hessian




CG iterations vs Newton iteration

When using a Hilbert curve ordering of the degrees of freedom in $2$d:

1. the Hessian possesses HODLR structure,
2. the HODLR compressed Hessian effectively preconditions the Newton Conjugate-Gradient system.

# HODLR compressed Hessian in two dimensions

## Positive definiteness is not preserved!

- For moderate levels of hierarchical refinement and/or moderately low off-diagonal rank, the HODLR compressed Hessian **fails to maintain positive definiteness**, thereby failing to adequately precondition the Newton-CG system.
- We hope to moderately extend recent work on generating symmetric positive definite hierarchical semi-separable (HSS) approximations to HODLR matrices.

  Details in: X. Xing, E. Chow, *Preserving Positive Definiteness in Hierarchically Semiseparable Matrix Approximations*, SIAM Journal on Matrix Analysis and Applications, 2018

# Summary

**Conclusions:**

- Numerical evidence demonstrates that Hessian manipulations can be made tractable by off-diagonal low-rank approximations.
- So that the localized PDE sensitivities manifest as a diagonally concentrated Hessian, it is necessary to reorder the degrees of freedom in a way that respects locality.

**Ongoing and future work:**

- Generate hierarchical approximations that are guaranteed to preserve positive definiteness.
- Apply all of the above to the full Stokes ice sheet problem.
- Apply HODLR approximations of the Hessian to efficiently generate Gaussian proposal distributions for MCMC sampling.
- Look to apply general $\mathcal{H}$-matrices to approximate $\boldsymbol{H}$.

# Diagonal concentration of the Hessian

## Manifestation of localized sensitivities

Let $\boldsymbol{u}(\boldsymbol{m})$ be the solution of the PDE for a given parameter field $\boldsymbol{m}$. $\boldsymbol{M}$, $\boldsymbol{K}$ the mass and stiffness matrices arising from the discretization of the PDE.

$$J\left(\boldsymbol{u}(\boldsymbol{m}), \boldsymbol{m}\right) = \frac{1}{2}\left(\boldsymbol{u} - \boldsymbol{u}_{\text{obs}}\right)^{\top}\left(\boldsymbol{u} - \boldsymbol{u}_{\text{obs}}\right) + \frac{\gamma}{2}\boldsymbol{m}^{\top}\boldsymbol{K}\boldsymbol{m}$$

We take a discretize and then optimize (DTO) approach, to get a sense of the structure of the Hessian in terms of the first and second order sensitivities.

$$\boldsymbol{H}_{i,j} = \frac{\partial^2}{\partial \boldsymbol{m}_i \partial \boldsymbol{m}_j} J\left(\boldsymbol{u}(\boldsymbol{m}), \boldsymbol{m}\right)$$

# Diagonal concentration of the Hessian

## Manifestation of localized sensitivities

$$\boldsymbol{H}_{i,j} = (\boldsymbol{u} - \boldsymbol{u}_{\text{obs}})^{\top} \boldsymbol{M} \frac{\partial^2 \boldsymbol{u}}{\partial \boldsymbol{m}_i \partial \boldsymbol{m}_j} + \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{m}_i}^{\top} \boldsymbol{M} \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{m}_j} + \gamma \boldsymbol{K}_{i,j}$$

- The only components of $\boldsymbol{u}$ which will be sensitive to variations of the parameter $\boldsymbol{m}$ at node $i$ are those components of $\boldsymbol{u}$ that are near node $i$.
- If parameter dof nodes $i$ and $j$ are far from one another, then no component of $\boldsymbol{u}$ will be sensitive to both variations.