A Data Scalable Hessian/KKT Preconditioner for Large Scale Inverse Problems*

SIAM CSE17

Nick Alger[1], Umberto Villa[1], Tan Bui-Thanh[2], Omar Ghattas[1]

[1]Center for Computational Geosciences and Optimization (CCGO),
Institute for Computational Engineering and Sciences (ICES),
The University of Texas at Austin
[2]Department of Aerospace Engineering and Engineering Mechanics, and
Institute for Computational Engineering and Sciences (ICES),
The University of Texas at Austin

February 27th, 2017

*This work was funded by DOE grants DE-SC0010518 and DE-SC0009286, and AFOSR grant FA9550-12-1-0484

Overview of the work

KKT system:

$$\underbrace{\begin{bmatrix} \alpha R^* R & & T^* \\ & B^* B & A^* \\ T & A & \end{bmatrix}}_{K} \underbrace{\begin{bmatrix} q \\ u \\ \lambda \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} b_q \\ b_u \\ b_\lambda \end{bmatrix}}_{b}$$

Preconditioner:

$$P := \begin{bmatrix} \alpha R^* R + \rho T^* T & & \\ & B^* B + \rho A^* A & \\ & & \frac{I}{\rho} I \end{bmatrix}$$

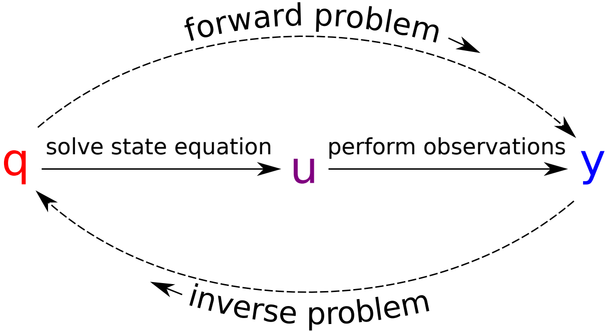
Condition number bound

$$\operatorname{cond}\left(P^{-1/2} K P^{-1/2}\right) \leq \frac{3}{\delta(1-\beta)}$$

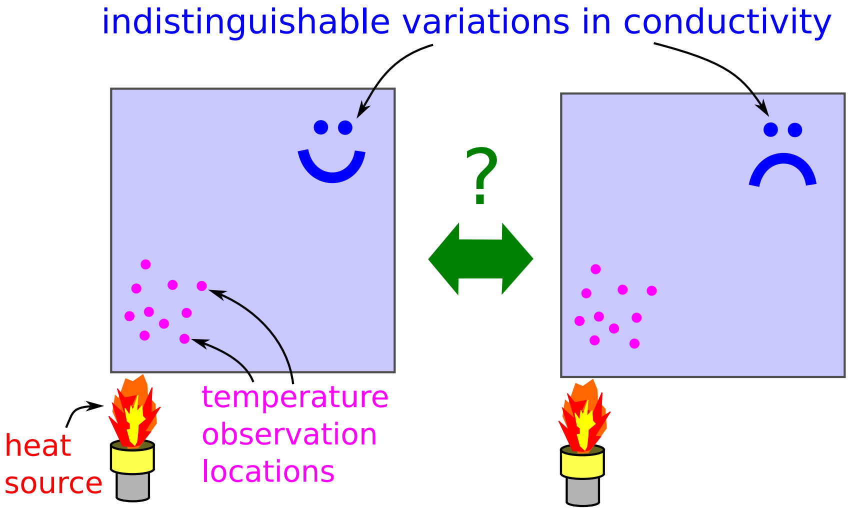
Inverse problem

$$\underbrace{\overset{\text{observations}}{y} = Bu + \overset{\text{noise}}{\zeta},}_{\text{observation process}} \quad \text{where} \quad F\underbrace{(\overset{\text{parameter}}{q},\overset{\text{state variable}}{u}) = 0}_{\text{state equation}}$$

Forward problem: Given q, compute y.
Inverse problem: Given y, estimate q.



Uninformed modes



Optimization problem

$$\underbrace{\overset{\text{observations}}{y} = Bu + \overset{\text{noise}}{\zeta},}_{\text{observation process}} \quad \text{where} \quad \underbrace{F(\overset{\text{parameter}}{q},\overset{\text{state variable}}{u}) = 0}_{\text{state equation}}$$

Regularized least-squares optimization problem:

$$\min_{q,u} \quad \overbrace{\frac{1}{2}||Bu - y||^2}^{\text{try to match observations}} + \overbrace{\frac{\alpha}{2}||Rq||^2}^{\substack{\textbf{regularization:}\\ \text{stabilize reconstruction} \\ \text{of uninformed modes}}}$$

$$\text{such that} \quad F(q, u) = 0.$$

Data misfit

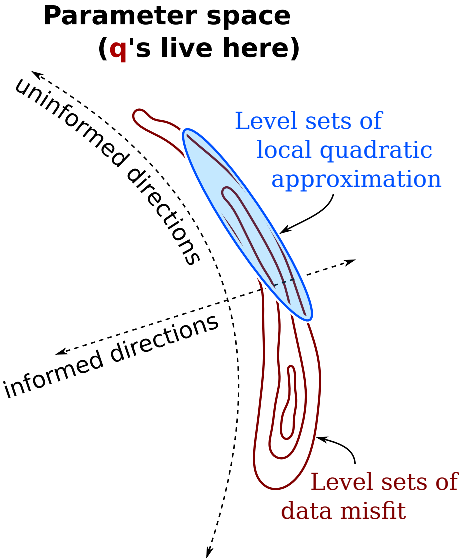
- data misfit:

$$\mathcal{J}_d(q) := \frac{1}{2}||Bu(q) - y||^2$$

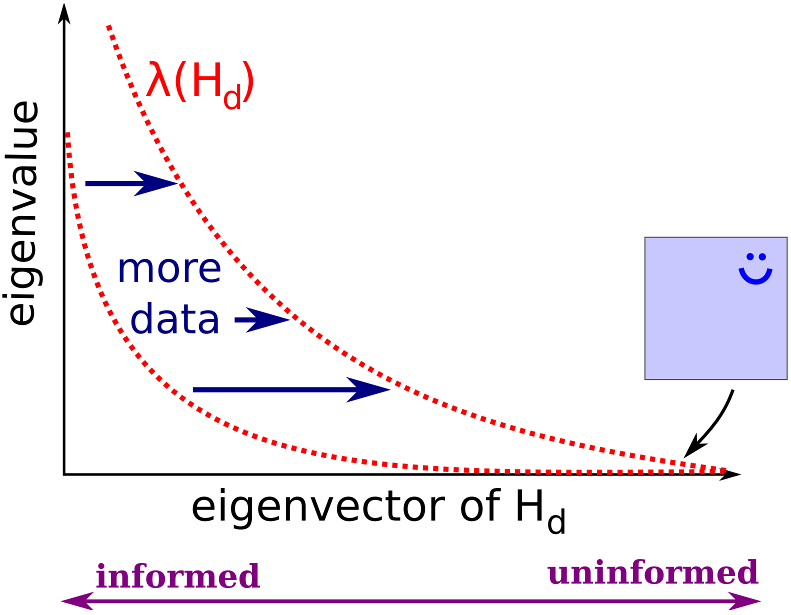
- data misfit Hessian:

$$H_d := \frac{d^2\mathcal{J}_d}{dq^2}$$

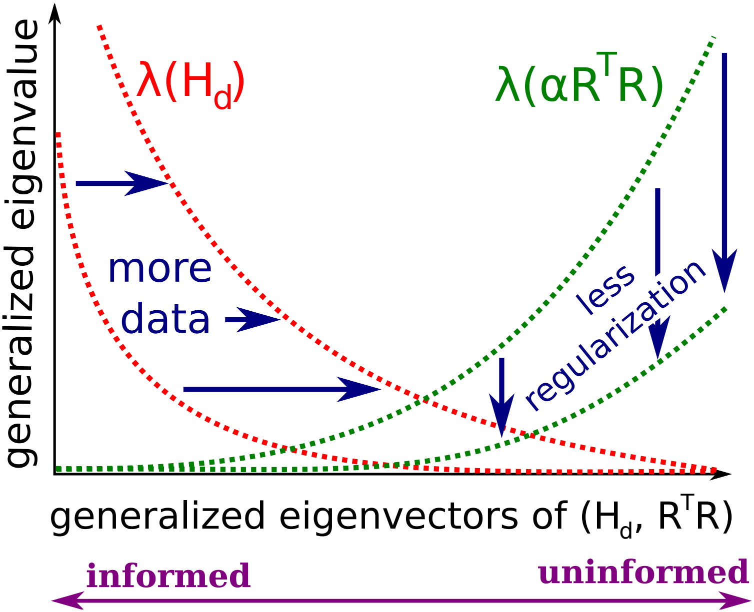
- Eigenstructure of H_d characterizes local sensitivity of observations to parameter perturbations



Spectrum of misfit Hessian



Eigenvalues of data misfit and regularization Hessian



Hessian spectrum and data scalability

- **Spectral structure** of the Hessian controls convergence of optimization schemes.
- **Increasing data** worsens the spectral structure of the Hessian.
- **Data scalable** methods must make progress on *all* informed modes every iteration.

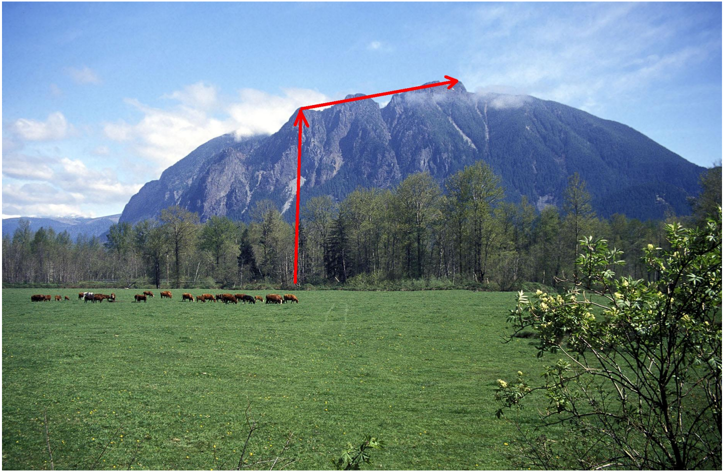
Consequently, the following are *not* data scalable:

- Gradient methods (gradient descent, nonlinear CG, Nesterov, L-BFGS)
- Newton-Krylov methods with regularization preconditioning

Gradient ascent path on a mountain

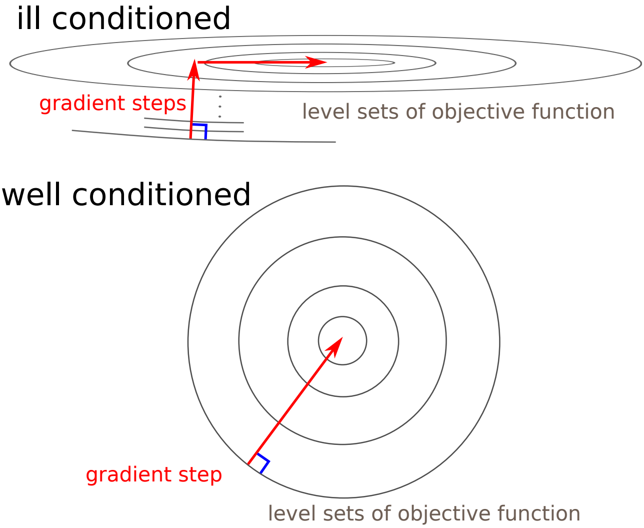
*Original image by Mountains to Sound Greenway Trust, https://commons.wikimedia.org/w/index.php?curid=705297

Gradient ascent path on a mountain

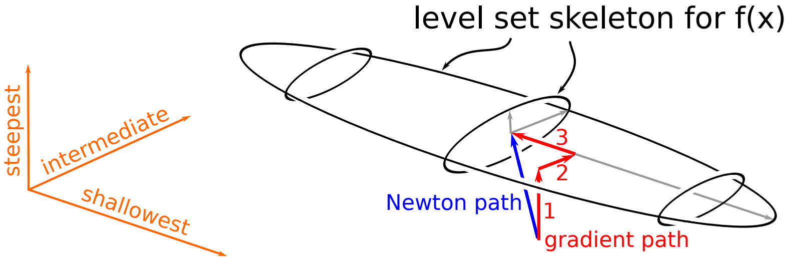


*Original image by Mountains to Sound Greenway Trust, https://commons.wikimedia.org/w/index.php?curid=705297

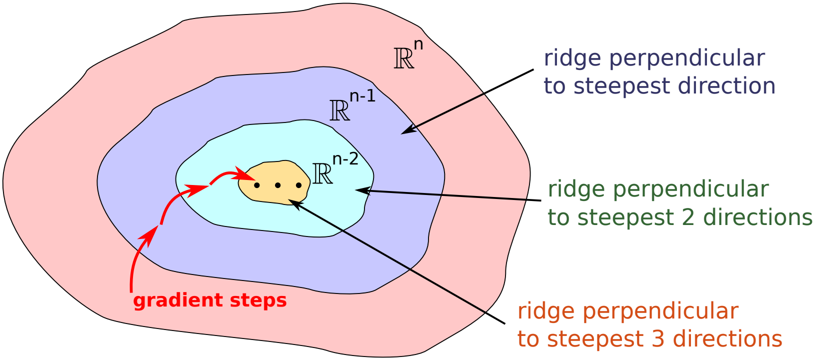
Gradient path in 2D



Gradient path in 3D



Gradient path in nD

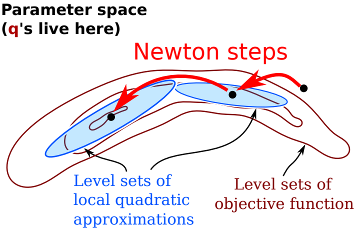


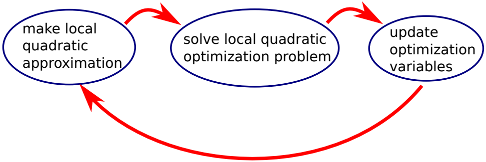
abstract set containment diagram
(not level sets)

Newton/SQP

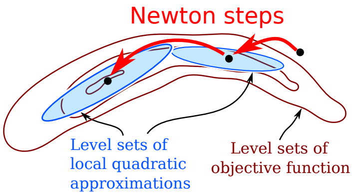
Accounts for scaling in all directions at once?:

- Newton ✓
- Gauss-Newton ✓
- Sequential quadratic programming (SQP) ✓
- L-BFGS ✗





Newton/SQP: address ill-conditioning with linear algebra



Newton-Krylov/SQP-Krylov:

- Solve linear system at each iteration with Krylov method
- Linear systems become harder solve with increasing data
- Decouples nonlinearity from ill-conditioning
- Allows us to directly address ill-conditioning with linear algebra/**preconditioning**

Sequential quadratic programming / Gauss-Newton

- **Linearize constraint equation** at current point:

$$F(q,u) = \underbrace{\frac{\partial F}{\partial q}}_{T} \underbrace{(q - q^{(k)})}_{\delta q} + \underbrace{\frac{\partial F}{\partial u}}_{A} \underbrace{(u - u^{(k)})}_{\delta u} + \text{higher order terms}$$

- Linearized optimization problem:

$$\boxed{\begin{array}{ll} \min\limits_{\delta q, \delta u} & \frac{1}{2}||B\ \delta u - \delta y||^2 + \frac{\alpha}{2}||R\ \delta q - r_k||^2 \\ \text{such that} & T\ \delta q + A\ \delta u = f_0 \end{array}}$$

- SQP/Gauss-Newton:

$$\begin{array}{c}\text{Linearize}\\\text{constraint}\end{array} \rightarrow \begin{array}{c}\text{Solve}\\\text{linearized problem}\end{array} \rightarrow \begin{array}{c}\text{Update}\\\text{optimization variables}\end{array} \rightarrow \text{Repeat}\ldots$$

Linear systems in SQP / Gauss-Newton

- **SQP:**

$$\begin{aligned}\min_{\delta q, \delta u} \quad & \frac{1}{2}\|B\ \delta u - \delta y\|^2 + \frac{\alpha}{2}\|R\ \delta q - r_k\|^2 \\ \text{such that} \quad & T\ \delta q + A\ \delta u = f_0\end{aligned}$$

Must solve system of the form $\mathbf{K}\mathbf{x} = \mathbf{b}$, where

$$\boxed{K = \begin{bmatrix} \alpha R^* R & & T^* \\ & B^* B & A^* \\ T & A & \end{bmatrix}.}$$

- **Gauss-Newton:** (eliminate δu by solving for it)

$$\min_{\delta q} \quad \frac{1}{2}\|\underbrace{BA^{-1}T}_{J}\ \delta q - \widehat{\delta y}\|^2 + \frac{\alpha}{2}\|R\ \delta q - r_k\|^2$$

Must solve system of the form $\mathbf{H}\mathbf{p} = -\mathbf{g}$, where

$$\boxed{H = J^* J + \alpha R^* R.}$$

- Coefficient matrices are equivalent:

$$K \underset{\text{Schur complement}\rightarrow}{\overset{\leftarrow\text{Define auxiliary variables}}{\rightleftharpoons}} H$$

Hessian preconditioning is hard

- Regularization and data misfit terms in Hessian *"fight"* each other by construction
  $\Longrightarrow$ Hard to find preconditioners that work for both terms at once

- Hessian is dense: only accessible via matrix-vector products
  $\Longrightarrow$ Cannot use preconditioners that require entries of the matrix (e.g., algebraic multigrid, algebraic domain decomposition, etc.)

- 15+ years of research by many groups, not much success

- Idea: precondition KKT matrix instead

Regularization preconditioning

Hessian:

$$H = \underbrace{J^* J}_{\substack{\text{compact}\\ \text{operator}}} + \alpha \underbrace{R^* R}_{\substack{\text{differential}\\ \text{operator}}}$$

Regularization preconditioned Hessian:

$$\frac{1}{\alpha} R^{-*} H R^{-1} = \underbrace{\frac{1}{\alpha} R^{-*} H_d R^{-1} + I}_{\substack{\text{compact perturbation}\\ \text{of identity}}}$$

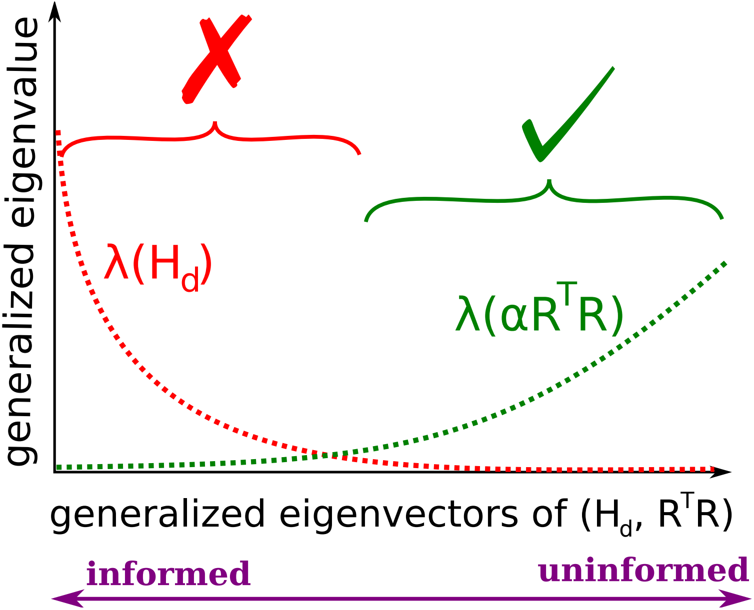
- mesh independent convergence of regularization preconditioned Newton-Krylov.

Problem:

$$\tfrac{1}{\alpha} R^{-T} H_d R^{-1} \text{ becomes "less compact" as:}$$

- the data increases
- the regularization decreases

Regularization preconditioning addresses uninformed modes



Adjoint Schur complement KKT preconditioning

- Generic saddle point optimization problem:

$$\begin{array}{rl} \min\limits_{q,u} & \frac{1}{2}||Bu-y||^2 + \frac{\alpha}{2}||Rq||^2 \\ \text{such that} & Tq + Au = f \end{array} \quad \rightarrow \quad \begin{array}{rl} \min\limits_{x} & \frac{1}{2}x^* Mx + g^* x \\ \text{such that} & Cx = h \end{array}$$

- KKT matrix in generic saddle point form:

$$\begin{bmatrix} \alpha R^* R & & T^* \\ & B^* B & A^* \\ T & A & \end{bmatrix} \rightarrow \begin{bmatrix} M & C^* \\ C & \end{bmatrix}$$

- Murphy, Golub, Wathen:

$$\lambda\left(\begin{bmatrix} M & \\ & C^* M^{-1} C \end{bmatrix}^{-1} \begin{bmatrix} M & C^* \\ C & \end{bmatrix}\right) \quad \text{consists of 3 points.}$$

$\Rightarrow$ Krylov methods converge in 3 iterations!

- **Problem:** M is *not invertible* due to limited observations.

Augmented Lagrangian

- **Problem:** M is not invertible due to limited observations.
- **Solution:** penalize constraint violations *even more*!

$$\begin{array}{rl} \min\limits_x & \frac{1}{2}x^*Mx + g^*x \\ \text{such that} & Cx = h \end{array} \rightarrow \begin{array}{rl} \min\limits_x & \frac{1}{2}x^*Mx + g^*x + \frac{\rho}{2}||Cx-h||^2 \\ \text{such that} & Cx = h \end{array}$$

- Augmented KKT matrix:

$$\begin{bmatrix} M & C^* \\ C & \end{bmatrix} \rightarrow \begin{bmatrix} M + \rho C^*C & C^* \\ C & \end{bmatrix}$$

- $M + \rho C^*C$ is invertible*.

(*provided optimization problem is well-posed)

Augmented adjoint Schur complement preconditioner

- Augmented preconditioner:

$$\begin{bmatrix} M & \\ & C^* M^{-1} C \end{bmatrix} \to \begin{bmatrix} M + \rho C^* C & \\ & C^*(M + \rho C^* C)^{-1} C \end{bmatrix}$$

- Golub, Greif, Varah:

$$\lambda\left(\begin{bmatrix} M + \rho C^* C & \\ & C^*(M + \rho C^* C)^{-1} C \end{bmatrix}^{-1} \begin{bmatrix} M & C^* \\ C & \end{bmatrix} \right)$$

$$\subset \left[-1, \frac{1-\sqrt{5}}{2}\right] \cup \left[1, \frac{1+\sqrt{5}}{2}\right]$$

$\Rightarrow$ Krylov methods converge very fast.

Difficulty: Preconditioner requires solving $M + \rho C^* C$ and $C^*(M + \rho C^* C)^{-1} C$.

Workaround: Replace these with approximations that are easier to solve.

Approximation of Schur complement

We must approximate the following operator:

$$S := C^*(M + \rho C^* C)^{-1} C.$$

S is the (negative) Schur complement for the adjoint variable.

- As $\rho \to \infty$, constraint is enforced in objective function,
  $\implies$ adjoint variable doesn't have to "work as hard"
  $\implies$ better conditioning of the Schur complement
- Use approximation:
  $$S \approx \frac{1}{\rho} I$$
- Approximation exact as $\rho \to \infty$

Approximation of objective block

Next, we must approximate the following operator:

$$M + \rho C^* C = \begin{bmatrix} \alpha R^* R + \rho T^* T & \rho T^* A \\ \rho A^* T & B^* B + \rho A^* A \end{bmatrix}$$

- Off-diagonals scaled by ρ
- ρ small: off-diagonals are less important
- **Approximation:** set off-diagonals to zero

$$M + \rho C^* C \sim \begin{bmatrix} \alpha R^* R + \rho T^* T & \\ & B^* B + \rho A^* A \end{bmatrix}$$

The combined preconditioner

After afforementioned approximations, preconditioner becomes:

$$P := \begin{bmatrix} \alpha R^* R + \rho T^* T & & \\ & B^* B + \rho A^* A & \\ & & \frac{1}{\rho} I \end{bmatrix}$$

- Schur complement block: want ρ large
- Objective 2x2 block: want ρ small

Question: can ρ be chosen just right to make both of these approximations good?

Answer: yes, set $\rho = \sqrt{\alpha}$.

Squared subsystems

Preconditioner subsystems:

$$\alpha R^* R + \rho T^* T$$
$$B^* B + \rho A^* A$$

- Terms do not "fight" each other
- Symmetric positive definite
- Have access to matrix entries*
- Can use algebraic multigrid, algebraic domain decomposition, ..

Condition number bound

- Define the damped projectors:

$$Q_R := T(\frac{\alpha}{\rho} R^* R + T^* T)^{-1} T^*$$
$$Q_J := T(\frac{1}{\rho} J^* J + T^* T)^{-1} T^*.$$

- Let δ, β be AM and GM bounds on Q_R, Q_J:

$$0 < \delta \leq \lambda_{\min}(Q_R + Q_J)$$
$$\lambda_{\max}(Q_R Q_J)^{1/2} \leq \beta < 1.$$

Theorem

$$cond\left(P^{-1/2} K P^{-1/2}\right) \leq \frac{3}{\delta(1-\beta)},$$

Proof.

Use Brezzi theory.

Bounds on δ, β

Theorem

Let R be a spectral filtering regularization operator with eigenvalues of R^*R given by r_i. Denote the eigenvalues of J^*J by d_i^2. Set $\rho = \sqrt{\alpha}$. If the following appropriate regularization assumptions hold:

1. $0 < c_u \leq d_i^2 + \alpha r_i^2$,
2. $d_i r_i \leq c_o < \infty$

then

$$\delta \geq \frac{1}{2}\left(1 + c_o^2\right)^{-1}$$
$$\beta \leq (1 + c_u)^{-1/2}$$

Discussion of appropriate regularization assumptions

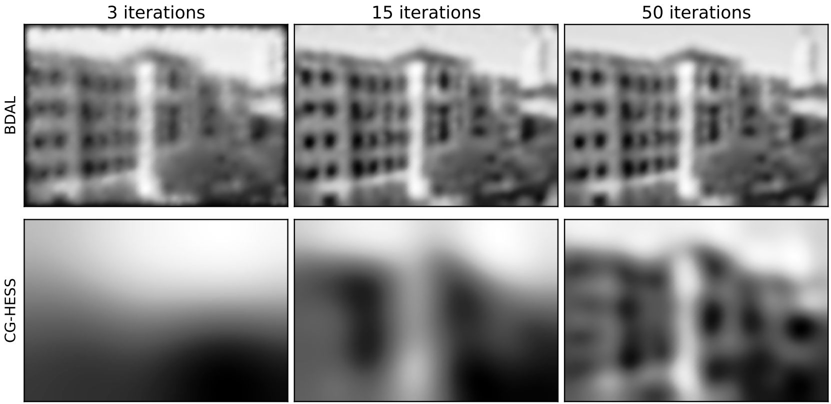
1. $0 < c_u \leq d_i^2 + \alpha r_i^2$,
2. $d_i r_i \leq c_o < \infty$

- Condition 1: Problem not under-regularized (Hessian nonsingular)
- Condition 2: Problem not over-regularized
- Condition 2: Multiplicative nature of condition 2 makes it easily satisfied
- Condition 2: Satisfied with constant $c_o = 1.0$ for Poisson source inversion problem with observations of Fourier modes, and Laplacian or weaker regularisation.

Numerical test problem

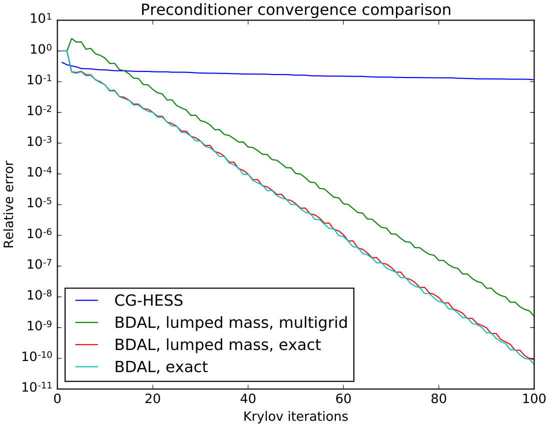
- Poisson source inversion problem
- $\Delta u = q$
- Point measurements of u
- True q: picture of POB building at UT Austin

Iterate comparison



- Top row: Our preconditioner on KKT system
- Bot row: Regularization preconditioning on Hessian

Convergence comparison

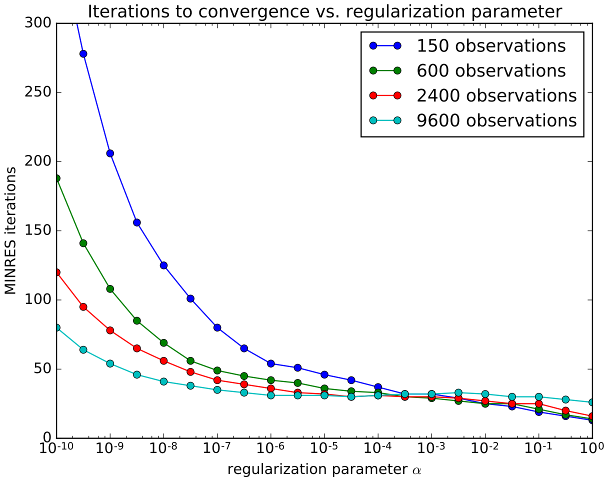


- Our preconditioner converges fast
- Regularization preconditioning stalls
- Replacing subsystem solves with a few multigrid V-cycles results in nearly the same convergence rate

Mesh scalability

h	# triangles	MINRES iterations
5.68e-02	1800	51
2.84e-02	7200	50
1.89e-02	16200	51
1.41e-02	29000	51
1.13e-02	45250	51
9.44e-03	65100	51
8.09e-03	88550	51
7.07e-03	116000	51
6.29e-03	146700	51
5.66e-03	181000	51

Data scalability



- Steady convergence rate over wide range of regularization parameter choices
- Can take regularization parameter very small if there is sufficient data

Conclusion

- Increasing data worsens spectral properties of Hessian
- Existing numerical optimization schemes slow with big data
- We addressed the problem with a data scalable KKT preconditioner
- Performs well when problem is neither over- nor under-regularized

Paper: N. Alger, U. Villa, T. Bui-Thanh, O. Ghattas, *A data scalable augmented Lagrangian KKT preconditioner for large scale inverse problems*. Submitted to SISC (in review). https://arxiv.org/pdf/1607.03556v1.pdf

Thanks to

Co-authors:

- Umberto Villa
- Omar Ghattas
- Tan Bui-Thanh

Colleagues:

- James Martin
- Toby Isaac
- Vishwas Rao
- Aaron Myers

Other:

- Anonymous reviewer: improvement in bound: $2 + 2\sqrt{2} \rightarrow 3$