Randomized Quasi-Newton Updates are Linearly Convergent Matrix Inversion Algorithms

Robert Mansel Gower
Joint work with Peter Richtárik

ICCOPT Tokyo, August 2016

Inverting a Matrix

The Problem

$\in \mathbb{R}^{n \times n}$

Identity matrix

$$\underbrace{A}_{n \times n} X = I$$

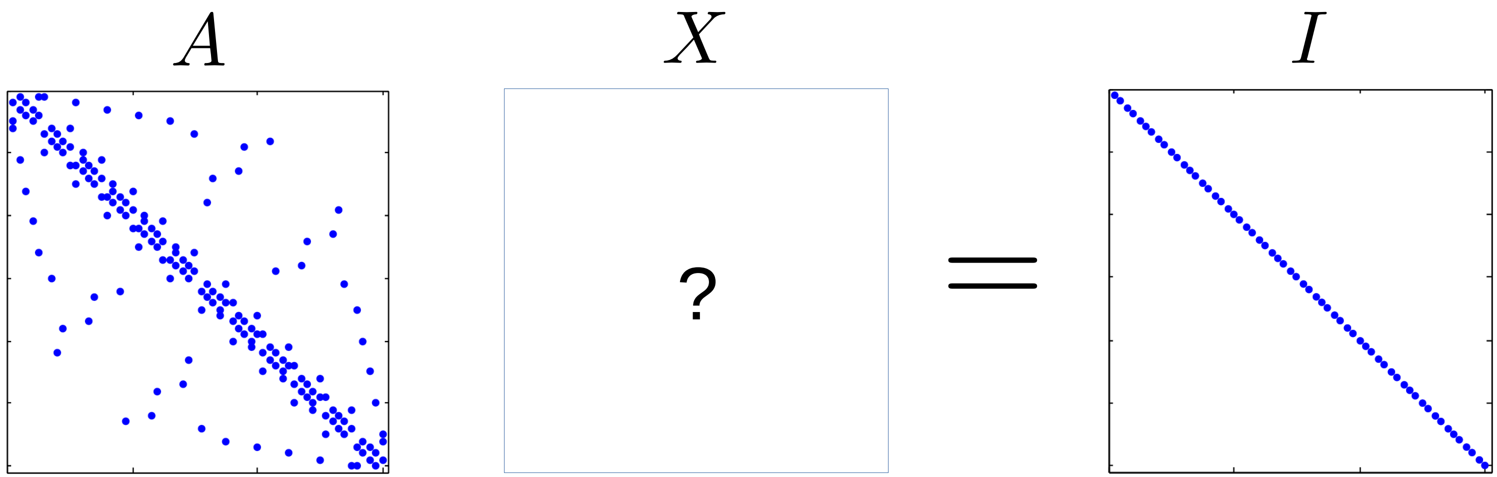
Assumption: The matrix A is nonsingular

Why iteratively invert a matrix?

- Matrix inverse standard tool (needed to calculate a Schur complement or a projection operator)
- Starting point for **randomized variable metric**
- Starting point for **randomized preconditioning**

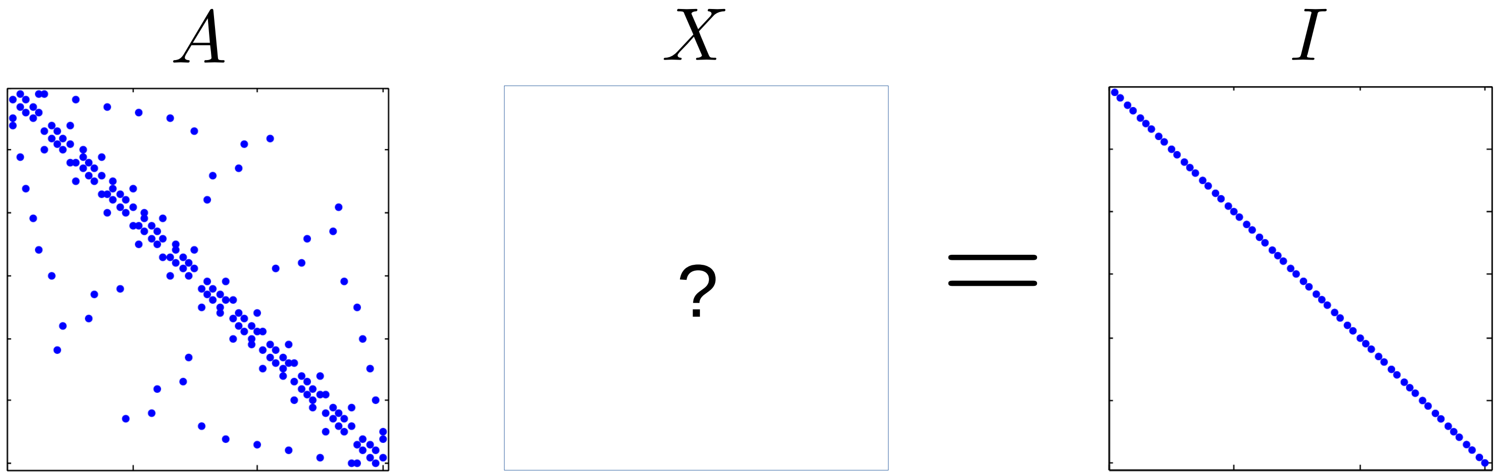
Randomized Methods for Nonsymmetric Matrices

The Sketching Idea



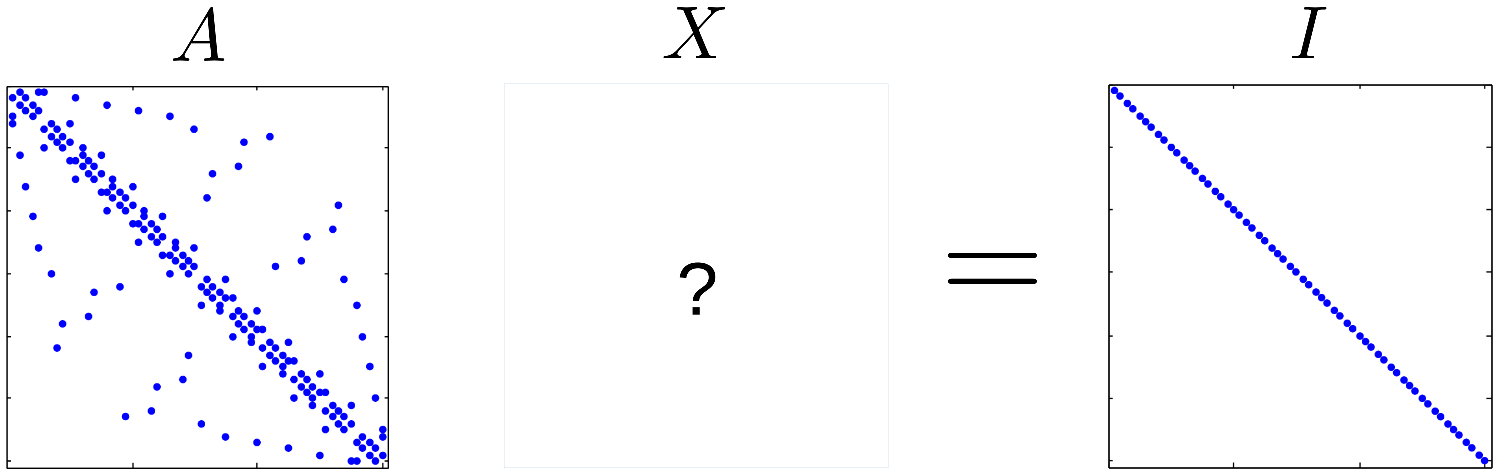
Compress system with random thin **random** matrix $S \in \mathbf{R}^{n \times \tau}, \tau \ll n.$

The Sketching Idea



Compress system with random thin **random** matrix $S \in \mathbf{R}^{n \times \tau}, \tau \ll n.$

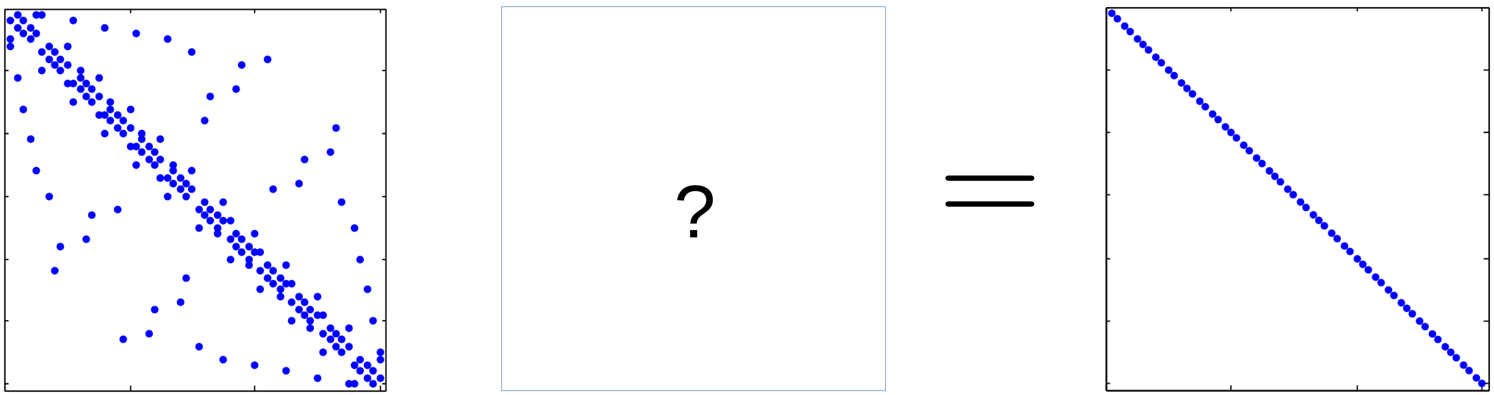
The Sketching Idea



Compress system with random thin **random** matrix $S \in \mathbf{R}^{n\times\tau}, \tau \ll n.$

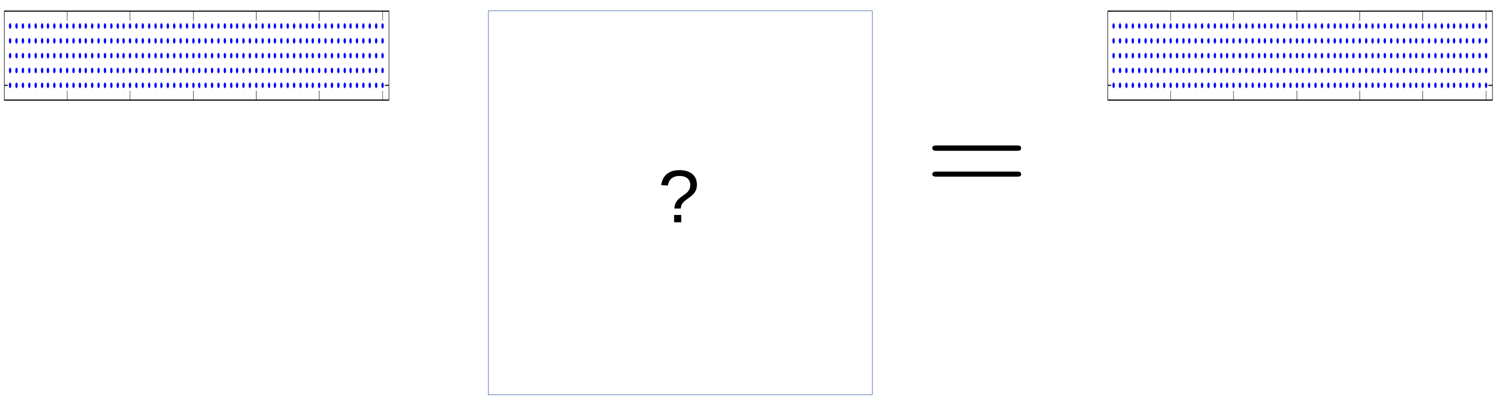
The Sketching Idea

$A \qquad X \qquad = \qquad I$



Compress system with random thin **random** matrix $S \in \mathbf{R}^{n \times \tau}, \tau \ll n.$

$S^T A \qquad X \qquad = \qquad S^T$



The quasi-Newton Viewpoint: “Sketch and Project”

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - X_k||^2_{F(W^{-1})}$$

$$\text{subject to} \quad S^T A X = S^T$$

The quasi-Newton Viewpoint: “Sketch and Project”

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - X_k||^2_{F(W^{-1})}$$

$$\text{subject to} \quad S^T A X = S^T$$

The quasi-Newton Viewpoint: "Sketch and Project"

$$\langle X, Y \rangle_{F(W^{-1})} = \mathbf{Tr}(X^T W^{-1} Y W^{-1}), \quad ||X||_{F(W^{-1})} = \sqrt{\langle X, X \rangle_{F(W^{-1})}}$$

W : Symmetric and positive definite

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - X_k||^2_{F(W^{-1})}$$

$$\text{subject to} \quad S^T A X = S^T$$

The quasi-Newton Viewpoint: "Sketch and Project"

$$\langle X, Y\rangle_{F(W^{-1})} = \mathbf{Tr}(X^T W^{-1} Y W^{-1}), \quad ||X||_{F(W^{-1})} = \sqrt{\langle X, X\rangle_{F(W^{-1})}}$$

W : Symmetric and positive definite

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n\times n}} ||X - X_k||^2_{F(W^{-1})}$$

$$\text{subject to} \quad S^T A X = S^T$$

S: random $n \times \tau$ matrix

The quasi-Newton Viewpoint: "Sketch and Project"

$$\langle X, Y\rangle_{F(W^{-1})} = \mathbf{Tr}(X^T W^{-1} Y W^{-1}), \quad ||X||_{F(W^{-1})} = \sqrt{\langle X, X\rangle_{F(W^{-1})}}$$

W : Symmetric and positive definite

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n\times n}} ||X - X_k||^2_{F(W^{-1})}$$

$$\text{subject to} \quad S^T A X = S^T$$

S: random $n \times \tau$ matrix

$$X_{k+1} = X_k - W A^T S (S^T A W A^T S)^{\dagger} S^T (A X_k - I)$$

The quasi-Newton Viewpoint: "Sketch and Project"

$$\langle X, Y \rangle_{F(W^{-1})} = \mathbf{Tr}(X^T W^{-1} Y W^{-1}), \quad ||X||_{F(W^{-1})} = \sqrt{\langle X, X \rangle_{F(W^{-1})}}$$

W : Symmetric and positive definite

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - X_k||^2_{F(W^{-1})}$$

$$\text{subject to} \quad S^T A X = S^T$$

S: random $n \times \tau$ matrix

$$X_{k+1} = X_k - W A^T S (S^T A W A^T S)^\dagger S^T (A X_k - I)$$

Includes methods good/bad Broyden, simultaneous Kaczmarz ...etc

Randomized Methods for Symmetric Matrices

$$A = A^T$$

1. The quasi-Newton Viewpoint: "Sketch and Project"

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - X_k||^2_{F(W^{-1})}$$

$$\text{subject to} \quad S^T A X = S^T, \quad X = X^T$$

1. The quasi-Newton Viewpoint: “Sketch and Project”

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - X_k||^2_{F(W^{-1})}$$

$$\text{subject to} \quad S^T A X = S^T, \quad X = X^T$$

1. The quasi-Newton Viewpoint: “Sketch and Project”

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - X_k||^2_{F(W^{-1})}$$

$$\text{subject to} \quad S^T A X = S^T, \quad X = X^T$$

Connection to quasi-Newton Methods: randomized block extension of the quasi-Newton updates.

$$S = \delta \in \mathbb{R}^n \quad \text{and} \quad \gamma := A\delta$$

and *A* is a “Hessian” we would like to invert. To be cheap, we can only sample the action *Aδ*

secant equation: $XAS = S$ ⟹ $X\gamma = \delta$

Goldfarb, D. (1970). **A Family of Variable-Metric Methods Derived by Variational Means**. Mathematics of Computation, 24(109), 23.

1. The quasi-Newton Viewpoint: "Sketch and Project"

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - X_k||^2_{F(W^{-1})}$$

$$\text{subject to} \quad S^T A X = S^T, \quad X = X^T$$

$$\int_0^1 \nabla^2 f(x_k + t\delta)\delta dt$$

si-Newton Methods: randomized

e quasi-Newton updates.

$$S = \delta \in \mathbb{R}^n \quad \text{and} \quad \gamma := A\delta$$

and A is a "Hessian" we would like to invert. To be cheap, we can only sample the action $A\delta$

secant equation: $XAS = S \quad \Rightarrow \quad X\gamma = \delta$

Goldfarb, D. (1970). **A Family of Variable-Metric Methods Derived by Variational Means**. Mathematics of Computation, 24(109), 23.

2. The Approx. Preconditioning viewpoint: “Constrain and Approximate”

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - A^{-1}||^2_{F(W^{-1})}$$

$$\text{subject to} \quad X = X_k + YS^TAW + WA^TSY^T$$

$$Y \in \mathbb{R}^{n \times \tau} \text{ is free}$$

2. The Approx. Preconditioning viewpoint: “Constrain and Approximate”

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - A^{-1}||^2_{F(W^{-1})}$$

$$\text{subject to} \quad X = X_k + YS^T AW + WA^T SY^T$$

$$Y \in \mathbb{R}^{n \times \tau} \text{ is free}$$

Duality: This is a dual problem of the sketch and project viewpoint, new insight into quasi-Newton methods.

New viewpoint for BFGS

Sketch and project

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - X_k||^2_{F(A)}$$

$$\text{subject to} \quad X\gamma = \delta, \quad X = X^T$$

New viewpoint for BFGS

Sketch and project

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - X_k||^2_{F(A)}$$

$$\text{subject to} \quad X\gamma = \delta, \quad X = X^T$$

New viewpoint for BFGS

Sketch and project

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - X_k||^2_{F(A)}$$

$$\text{subject to} \quad X\gamma = \delta, \quad X = X^T$$

Constrain and approximate

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - A^{-1}||^2_{F(A)}$$

$$\text{subject to} \quad X = X_k + y\delta^T + \delta y^T$$

$$y \in \mathbb{R}^n \text{ is free}$$

New viewpoint for BFGS

Sketch and project

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - X_k||^2_{F(A)}$$

$$\text{subject to} \quad X\gamma = \delta, \quad X = X^T$$

Constrain and approximate

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - A^{-1}||^2_{F(A)}$$

$$\text{subject to} \quad X = X_k + y\delta^T + \delta y^T$$

$$y \in \mathbb{R}^n \text{ is free}$$

Duality: The BFGS projects the inverse onto a 2-dimensional space of symmetric matrices

3. Algebraic Viewpoint
"Random Update"

$$H := S(S^T A W A^T S)^\dagger S^T$$

$$X_{k+1} = X_k - (X_k A - I) H A W$$
$$+ W A H (A X_k - I)(A H A W - I)$$

3. Algebraic Viewpoint
"Random Update"

Moore-Penrose pseudo inverse

$$H := S(S^T AWA^T S)^\dagger S^T$$

$$X_{k+1} = X_k - (X_k A - I) HAW + WAH(AX_k - I)(AHAW - I)$$

Fact: Every (not necessarily square) real matrix M has a real pseudo-inverse $M^\dagger$.

3. Algebraic Viewpoint
"Random Update"

Moore-Penrose pseudo inverse

$$H := S(S^T AWA^T S)^\dagger S^T$$

$$X_{k+1} = X_k - (X_k A - I)HAW + WAH(AX_k - I)(AHAW - I)$$

Low rank 3 X τ update

Fact: Every (not necessarily square) real matrix M has a real pseudo-inverse $M^\dagger$.

4. Analytic Viewpoint
“Random Fixed Point”

$$R_k := X_k - A^{-1}$$

$$R_{k+1} = (I - WA^T HA) R_k (I - AHA^T W)$$

4. Analytic Viewpoint
“Random Fixed Point”

$$R_k := X_k - A^{-1}$$

$$R_{k+1} = (I - WA^T HA)R_k(I - AHA^T W)$$

A positive Linear operator applied to old residual defines the new residual

Complexity / Convergence

Theorem [GR'16]

If S has full column rank with probability one then

1. $$||\mathbf{E}[X_k - A^{-1}]||_{W^{-1}} \leq \rho^k \, ||X_0 - A^{-1}||_{W^{-1}}$$

Complexity / Convergence

Theorem [GR'16]

$$||A||_{W^{-1}} = ||W^{-1/2} A W^{-1/2}||_2$$

If S has full column rank with probability one then

1. $$||\mathbf{E}[X_k - A^{-1}]||_{W^{-1}} \leq \rho^k \, ||X_0 - A^{-1}||_{W^{-1}}$$

Complexity / Convergence

Theorem [GR'16]

$$\|A\|_{W^{-1}} = \|W^{-1/2} A W^{-1/2}\|_2$$

If S has full column rank with probability one then

1. $$\|\mathbf{E}[X_k - A^{-1}]\|_{W^{-1}} \leq \rho^k \|X_0 - A^{-1}\|_{W^{-1}}$$

$$\rho := 1 - \lambda_{\min}(W^{1/2} A^T \mathbf{E}[H] A W^{1/2})$$

$\lambda_{\min}(\cdot) =$ smallest eigenvalue

Complexity / Convergence

Theorem [GR'16]

$$||A||_{W^{-1}} = ||W^{-1/2} A W^{-1/2}||_2$$

If S has full column rank with probability one then

1. $$||\mathbf{E}[X_k - A^{-1}]||_{W^{-1}} \leq \rho^k \, ||X_0 - A^{-1}||_{W^{-1}}$$

$$\rho := 1 - \lambda_{\min}(W^{1/2} A^T \mathbf{E}[H] A W^{1/2})$$

$\lambda_{\min}(\cdot) =$ smallest eigenvalue

and if $\mathbf{E}[H] \succ 0$ then

2. $$\mathbf{E}[||X_k - A^{-1}||^2_{F(W^{-1})}] \leq \rho^k \, ||X_0 - A^{-1}||^2_{F(W^{-1})}$$

Case study of $\mathbf{E}[H]$

$$H := S(S^T A W A^T S)^\dagger S^T$$

$$W = I$$

$$\mathbf{P}(S = e^i) = \frac{1}{m}$$

Case study of $\mathbf{E}[H]$

$$H := S(S^T A W A^T S)^\dagger S^T$$

Special Choice of Parameters

$$W = I$$

$$\mathbf{P}(S = e^i) = \frac{1}{m} \Rightarrow S = e^i$$

Case study of $\mathbf{E}[H]$

$$H := S(S^T A W A^T S)^{\dagger} S^T$$

Special Choice of Parameters

$$\mathbf{P}(S = e^i) = \frac{1}{m}$$

$$W = I$$

$$S = e^i$$

$$\mathbf{E}[H] = \frac{1}{m} \sum_{i=1}^{m} \frac{e_i e_i^T}{||A_{i:}||_2^2} = \mathrm{diag}(||A_{i:}||_2^2)$$

Case study of $\mathbf{E}[H]$

$$H := S(S^T AWA^T S)^{\dagger} S^T$$

Special Choice of Parameters

$$\mathbf{P}(S = e^i) = \frac{1}{m}$$

$$W = I$$

$$S = e^i$$

$$\mathbf{E}[H] = \frac{1}{m}\sum_{i=1}^{m} \frac{e_i e_i^T}{||A_{i:}||_2^2} = \mathrm{diag}(||A_{i:}||_2^2)$$

Case study of $\mathbf{E}[H]$

$$H := S(S^T AWA^T S)^{\dagger} S^T$$

Special Choice of Parameters

$$\mathbf{P}(S = e^i) = \frac{1}{m}$$

$$W = I$$

$$S = e^i$$

$$\mathbf{E}[H] = \frac{1}{m}\sum_{i=1}^{m} \frac{e_i e_i^T}{||A_{i:}||_2^2}$$

$$= \operatorname{diag}(||A_{i:}||_2^2)$$

Case study of $\mathbf{E}[H]$

$$H := S(S^T AWA^T S)^\dagger S^T$$

Special Choice of Parameters

$$\mathbf{P}(S = e^i) = \frac{1}{m}$$

$$W = I$$

$$S = e^i$$

$$\mathbf{E}[H] = \frac{1}{m}\sum_{i=1}^{m} \frac{e_i e_i^T}{||A_{i:}||_2^2} = \mathrm{diag}(||A_{i:}||_2^2)$$

Positive definite when *A* has no zero rows

The rate: lower and upper bounds

Theorem [RG'15]

$$0 \leq 1 - \frac{\tau}{n} \leq \rho \leq 1$$

The rate: lower and upper bounds

Theorem [RG'15]

$$0 \leq 1 - \frac{\tau}{n} \leq \rho \leq 1$$

Insight: The method is a *contraction* (with only full rank assumption on *S*). That is, things can not get worse.

The rate: lower and upper bounds

Theorem [RG'15]

$$0 \leq 1 - \frac{\tau}{n} \leq \rho \leq 1$$

Insight: The method is a *contraction* (with only full rank assumption on *S*). That is, things can not get worse.

The rate: lower and upper bounds

Theorem [RG'15]

$$0 \leq 1 - \frac{\tau}{n} \leq \rho \leq 1$$

Insight: The method is a *contraction* (with only full rank assumption on *S*). That is, things can not get worse.

Insight: The lower bound on the rate is better for *S* high rank, that is, when the dimension of the search space in the "constrain and approximate" viewpoint grows.

Special Case:
Randomized Block BFGS

Randomized BFGS

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - X_k||^2_{F(A)}$$

$$\text{subject to} \quad S^T A X = S^T, \quad X = X^T$$

Randomized BFGS

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - X_k||^2_{F(A)}$$

$$\text{subject to} \quad S^T A X = S^T, \quad X = X^T$$

Randomized BFGS

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - X_k||^2_{F(A)}$$

$$\text{subject to} \quad S^T A X = S^T, \quad X = X^T$$

Special Choice of Parameters

positve definite → $W = A^{-1}$

$\mathbf{P}(S = e_i) = p_i$ → $S = e_i$

Randomized BFGS

$$X_{k+1} = \arg\min_{X \in \mathbb{R}^{n \times n}} ||X - X_k||^2_{F(A)}$$

$$\text{subject to} \quad S^T A X = S^T, \quad X = X^T$$

Special Choice of Parameters

positve definite → $W = A^{-1}$

$\mathbf{P}(S = e_i) = p_i$ → $S = e_i$

⇒

$$X_{k+1} = H + (I - HA)X_k(I - AH)$$

Randomized BFGS

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - X_k||^2_{F(A)}$$
$$\text{subject to} \quad S^T A X = S^T, \quad X = X^T$$

Special Choice of Parameters

positve definite → $W = A^{-1}$

$\mathbf{P}(S = e_i) = p_i$ → $S = e_i$

$$X_{k+1} = H + (I - HA) X_k (I - AH)$$

Complexity Rate. A is positive definite → $\mathbf{E}[H] \succ 0$

$$p_i = \frac{A_{ii}}{\mathbf{Tr}(A)}$$

$$\mathbf{E}[||AX_k - I||^2_F] \leq \left(1 - \frac{\lambda_{\min}(A)}{\mathbf{Tr}(A)}\right)^k ||AX_0 - I||^2_F$$

Convenient probability

Theorem [GR‘15]

$$\bar{S} := [S_1, \ldots, S_r] \quad \text{is nonsingular}$$

$$\mathbf{P}(S = S_i) = p_i = \frac{\mathbf{Tr}(S_i^T AWA^T S_i)}{\mathbf{Tr}(\bar{S}^T AWA^T \bar{S})}$$

Convenient probability

Theorem [GR‘15]

$$\bar{S} := [S_1, \ldots, S_r] \quad \text{is nonsingular}$$

$$\mathbf{P}(S = S_i) = p_i = \frac{\mathbf{Tr}(S_i^T A W A^T S_i)}{\mathbf{Tr}(\bar{S}^T A W A^T \bar{S})}$$

$$\Downarrow$$

$$\rho = 1 - \frac{1}{\kappa^2(W^{1/2} A^T \bar{S})}$$

$$\kappa(W^{1/2} A^T \bar{S}) := ||(W^{1/2} A^T \bar{S})^{-1}||_2 ||W^{1/2} A^T \bar{S}||_F \geq \sqrt{n}$$

Randomized Block BFGS

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - X_k||^2_{F(A)}$$

$$\text{subject to} \quad S^T A X = S^T, \quad X = X^T$$

Special Choice of Parameters

positve definite $\Rightarrow W = A^{-1}$

$\mathbf{P}(S = S_i) = p_i \Rightarrow S = S_i$

$$\Rightarrow X_{k+1} = H + (I - HA)X_k(I - AH)$$

Complexity Rate. If A is positive definite $\Rightarrow \mathbf{E}[H]$ is nonsingular

$$p_i = \frac{\mathbf{Tr}(S_i^T A S_i)}{\mathbf{Tr}(\bar{S}^T A \bar{S})} \Rightarrow \mathbf{E}[||AX_k - I||^2_F] \leq \left(1 - \frac{1}{\kappa^2(A^{1/2}\bar{S})}\right)^k ||AX_0 - I||^2_F$$

Randomized Block BFGS

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - X_k||^2_{F(A)}$$

$$\text{subject to} \quad S^T A X = S^T, \quad X = X^T$$

Special Choice of Parameters

positve definite $\Rightarrow W = A^{-1}$

$\mathbf{P}(S = S_i) = p_i \Rightarrow S = S_i$

$$X_{k+1} = H + (I - HA)X_k(I - AH)$$

Idea: To minimize condition number, choose *S* so that $\bar{S}$ is an approximate inverse of $A^{1/2}$

$$p_i = \frac{\mathbf{Tr}(S_i^T A S_i)}{\mathbf{Tr}(\bar{S}^T A \bar{S})}$$

$$\mathbf{E}[||AX_k - I||^2_F] \le \left(1 - \frac{1}{\kappa^2(A^{1/2}\bar{S})}\right)^k ||AX_0 - I||^2_F$$

Adaptive Randomized Block BFGS (adaRBFGS)

$$\mathbf{E}[||AX_k - I||_F] \leq \left(1 - \frac{1}{\kappa^2(A^{1/2}\bar{S})}\right)^k ||AX_0 - I||_F$$

To minimize condition number:

$$\text{If } \bar{S} = A^{-1/2} \text{ then } \kappa(A^{1/2}\bar{S}) = \kappa(I) = \sqrt{n}$$

$$X_k \to A^{-1} \quad \Rightarrow \quad X_k^{1/2} \to A^{-1/2}$$

$$¿\bar{S} = X_k^{1/2}?$$

Adaptive Randomized Block BFGS (adaRBFGS)

$$\text{Maintain and update } L_k = X_k^{1/2}$$

`adaRBFGS_cols:`

$$S = L_k I_{:C}, \quad C \subset \{1, \ldots, n\} \text{ random set}$$

$$\downarrow$$

$$\bar{S} = L_k = X_k^{1/2}$$

`adaRBFGS_guass:` $S \sim \mathcal{N}(0, X_k)$

Adaptive Randomized Block BFGS (adaRBFGS)

Maintain and update $L_k = X_k^{1/2}$

`adaRBFGS_cols:`

$$S = L_k I_{:C}, \quad C \subset \{1, \ldots, n\} \text{ random set}$$

$$\bar{S} = L_k = X_k^{1/2}$$

`adaRBFGS_guass:` $\quad S \sim \mathcal{N}(0, X_k)$

Gratton, S., Sartenaer, A., & Ilunga, J. T. (2011). **On a Class of Limited Memory Preconditioners for Large-Scale Nonlinear Least-Squares Problems**. SIAM Journal on Optimization, 21(3), 912–935.

Experiments

Current state of the art

Symmetric Newton-Schulz

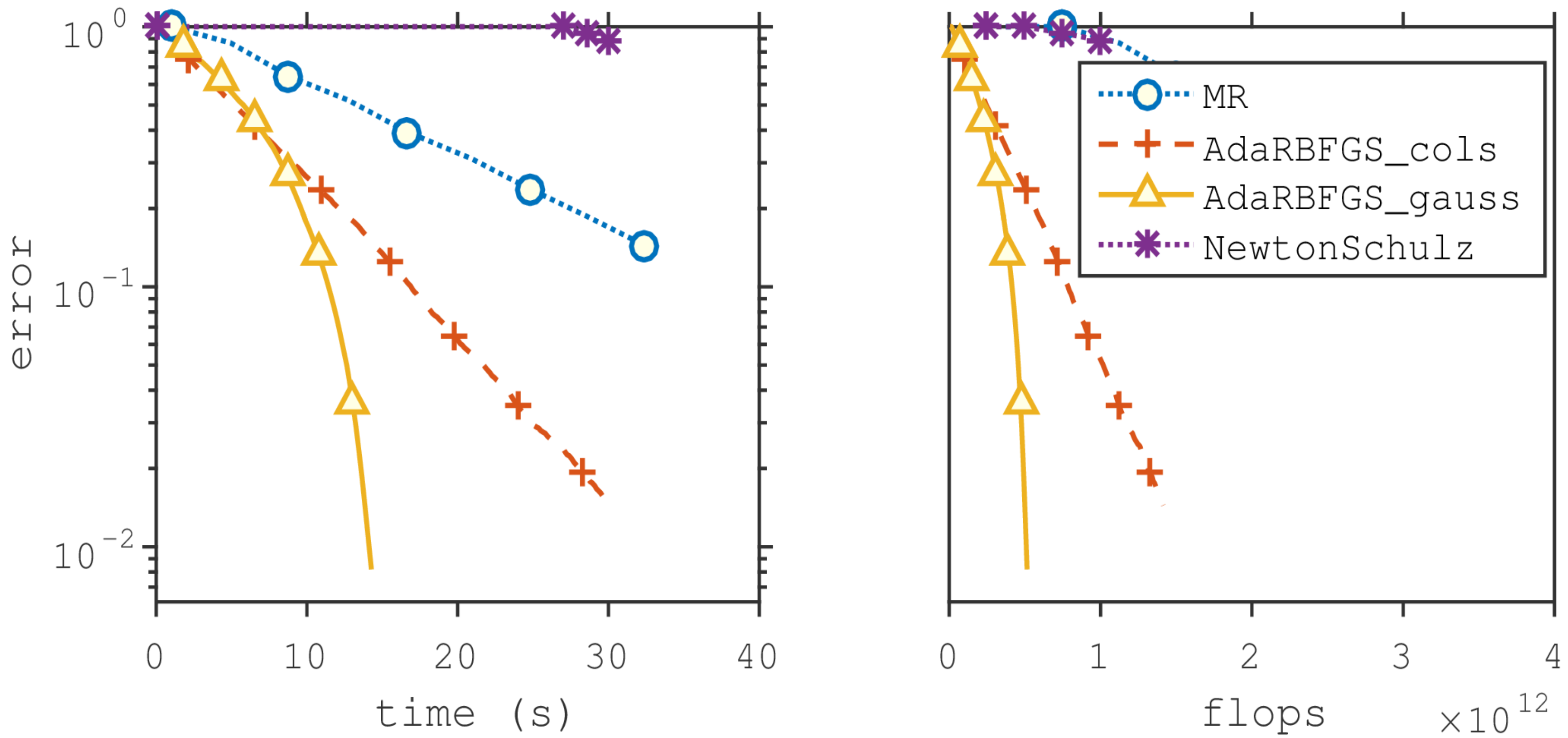
$$X_{k+1} = 2X_k - X_k A X_k$$

Self-conditioning Minimal Residual (MR)

$$\begin{aligned} X_{k+1} = \arg_X \quad & \min ||AX - I||_F^2 \\ & \text{subject to} \quad X = X_k + \alpha X_k (A X_k - I) \end{aligned}$$

Synthetic Problem

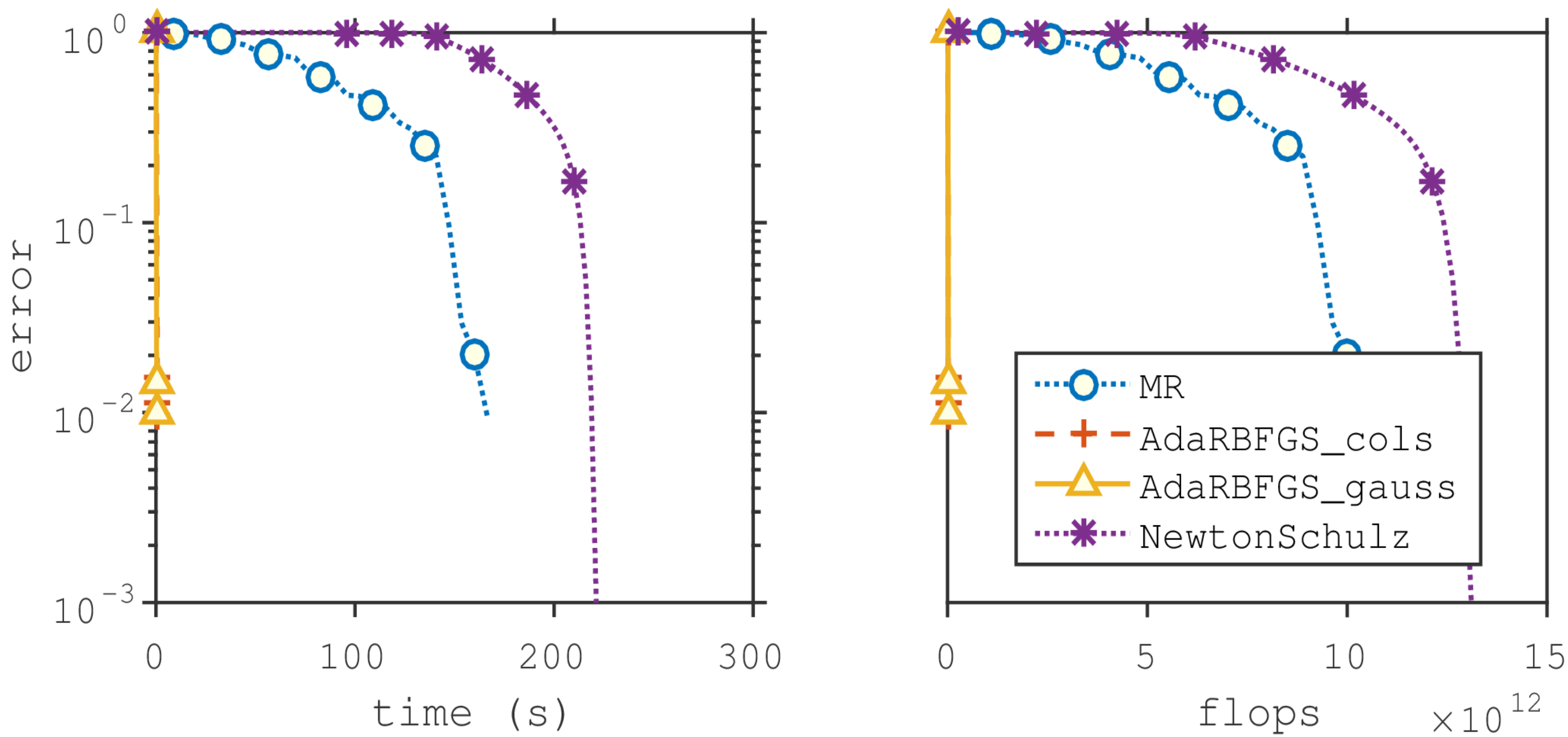
Synthetic data



(*randn*, *n* = 5000)

Ridge Regression Hessian

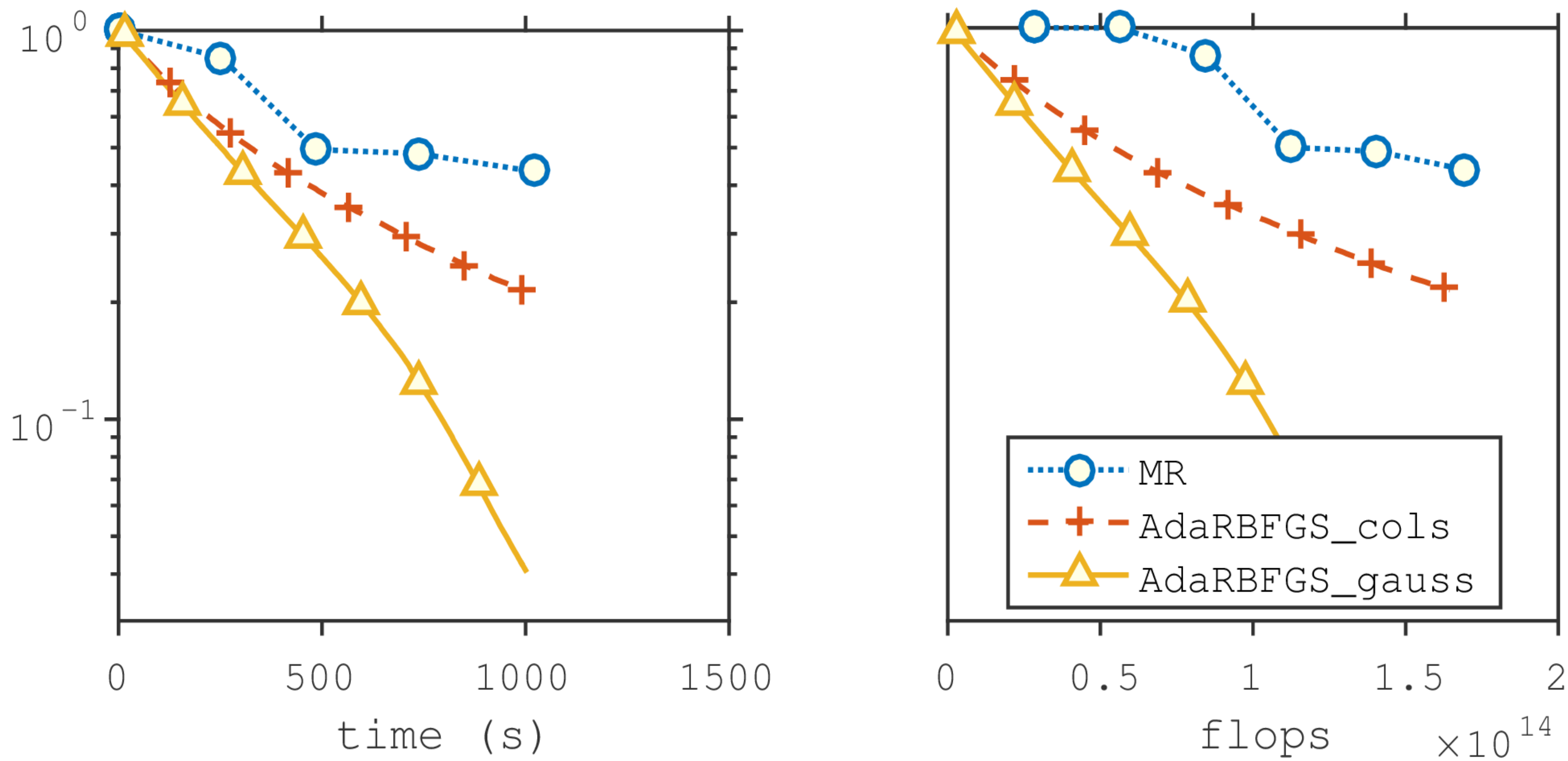
LIBSVM data



(gisette, n = 5,000)

Sparse Matrices from Engineering

UF collection



(GHS psdef/wathen100, n = 30,401)

Conclusion

- **New randomized methods** for calculating of approximate inverses of large-scale matrices

- **Convergence rates** which can form the basis of convergence of preconditioning or variable metric methods.

- **Dual viewpoints** of classic quasi-Newton methods, connection to Approximate Inverse Preconditioning methods

- **Can be extended** to calculating pseudo-inverse

Thank you,
Questions?

RMG and Peter Richtárik
Randomized Iterative Methods for Linear Systems
SIAM. J. Matrix Anal. & Appl., 36(4), 1660–1690, 2015

RMG, D. Goldfarb and Peter Richtárik
Stochastic Block BFGS: Squeezing More Curvature out of Data
ICML, 2016

RMG and Peter Richtárik
Randomized Quasi-Newton Updates are Linearly Convergent Matrix Inversion Algorithms
Preprint arXiv:1602.01768, 2016