Fast and Robust Solution Techniques for Large Scale Linear Least Squares Problems

M.S. Thesis Presentation

Department of Electrical and Electronics Engineering, Bilkent University

İbrahim Kurban Özaslan

<u>Committee</u>
Orhan Arıkan (Advisor)
Sinan Gezici
Elif Vural

13 July 2020

Outline

1. Review of Linear Least Squares Problems
   - Problem Formulation
   - Traditional Approaches
     - Reconstruction for a given regularization parameter
     - Methods for estimation of the unknown regularization parameter
   - Random Projection Based Approaches for the LS Problems
2. Proposed Momentum Iterative Hessian Sketch (M-IHS) Techniques
   - Techniques for a Given Regularization Parameter
   - Hybrid Techniques to Estimate Unknown Regularization Parameter
3. Conclusions and Future Work

Linear Least Squares Problems

- Linear systems of equations:

$$\mathbf{A}\mathbf{x}_0 + \mathbf{w} = \mathbf{b}, \qquad \mathbf{A} \in \mathbb{R}^{n \times d}.$$

- Aim is to recover $\mathbf{x}_0$ by observing $\mathbf{A}$ and $\mathbf{b}$.

Linear Least Squares Problems

- Linear systems of equations:

$$\mathbf{A}\mathbf{x}_0 + \mathbf{w} = \mathbf{b}, \qquad \mathbf{A} \in \mathbb{R}^{n \times d}.$$

- Aim is to recover $\mathbf{x}_0$ by observing $\mathbf{A}$ and $\mathbf{b}$.
- My studies focus on the LS solutions:

$$\mathbf{x}_{\mathsf{LS}} = \underset{\mathbf{x} \in \mathbb{R}^d}{\operatorname{argmin}} \; \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b}$$

Linear Least Squares Problems

- Linear systems of equations:

$$\mathbf{A}\mathbf{x}_0 + \mathbf{w} = \mathbf{b}, \qquad \mathbf{A} \in \mathbb{R}^{n \times d}.$$

- Aim is to recover $\mathbf{x}_0$ by observing $\mathbf{A}$ and $\mathbf{b}$.
- My studies focus on the LS solutions:

$$\mathbf{x}_{\mathsf{LS}} = \underset{\mathbf{x} \in \mathbb{R}^d}{\operatorname{argmin}} \; \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b}$$

- In practise due to ill conditioned nature of $\mathbf{A}$, $\mathbf{x}_{\mathsf{LS}}$ may not be acceptable.

Linear Least Squares Problems

- Linear systems of equations:

$$\mathbf{A}\mathbf{x}_0 + \mathbf{w} = \mathbf{b}, \qquad \mathbf{A} \in \mathbb{R}^{n \times d}.$$

- Aim is to recover $\mathbf{x}_0$ by observing $\mathbf{A}$ and $\mathbf{b}$.
- My studies focus on the LS solutions:

$$\mathbf{x}_{\mathsf{LS}} = \underset{\mathbf{x} \in \mathbb{R}^d}{\operatorname{argmin}} \; \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = \left(\mathbf{A}^T\mathbf{A}\right)^{-1} \mathbf{A}^T\mathbf{b}$$

- In practise due to ill conditioned nature of $\mathbf{A}$, $\mathbf{x}_{\mathsf{LS}}$ may not be acceptable.
- Generally, it is used with an additional penalty:

$$\mathbf{x}(\lambda) = \underset{\mathbf{x} \in \mathbb{R}^d}{\operatorname{argmin}} \; \underbrace{\frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \frac{\lambda}{2}\|\mathbf{x}\|_2^2}_{f(\mathbf{x},\lambda)} = \underset{\mathbf{x} \in \mathbb{R}}{\operatorname{argmin}} \; \left\| \begin{bmatrix} \mathbf{A} \\ \sqrt{\lambda}\mathbf{I}_d \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix} \right\|_2^2 = \left(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}\right)^{-1} \mathbf{A}^T\mathbf{b}$$

Linear Least Squares Problems

- Linear systems of equations:

$$\mathbf{A}\mathbf{x}_0 + \mathbf{w} = \mathbf{b}, \qquad \mathbf{A} \in \mathbb{R}^{n\times d}.$$

- Aim is to recover $\mathbf{x}_0$ by observing $\mathbf{A}$ and $\mathbf{b}$.
- My studies focus on the LS solutions:

$$\mathbf{x}_{\text{LS}} = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b}$$

- In practise due to ill conditioned nature of $\mathbf{A}$, $\mathbf{x}_{\text{LS}}$ may not be acceptable.
- Generally, it is used with an additional penalty:

$$\mathbf{x}(\lambda) = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \ \underbrace{\frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \frac{\lambda}{2}\|\mathbf{x}\|_2^2}_{f(\mathbf{x},\lambda)} = \underset{\mathbf{x}\in\mathbb{R}}{\text{argmin}} \ \left\| \begin{bmatrix} \mathbf{A} \\ \sqrt{\lambda}\mathbf{I}_d \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix} \right\|_2^2 = \left(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}\right)^{-1}\mathbf{A}^T\mathbf{b}$$

1. Find a proper estimate for λ
2. Construct the solution $\mathbf{x}(\lambda)$

Traditional Approaches: Solution Reconstruction for a Given λ

- Closed form solution: $\mathbf{x}(\lambda) = (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}_d)^{-1}\mathbf{A}^T\mathbf{b}$ where $\mathbf{A} \in \mathbb{R}^{n \times d}$
  - ! $O(nd^2)$ complexity of multiplication
  - ! Squares the condition number

Traditional Approaches: Solution Reconstruction for a Given λ

- Closed form solution: $\mathbf{x}(\lambda) = (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}_d)^{-1}\mathbf{A}^T\mathbf{b}$ where $\mathbf{A} \in \mathbb{R}^{n \times d}$
  - ! $O(nd^2)$ complexity of multiplication
  - ! Squares the condition number
- Direct methods: $\mathbf{x}(\lambda) = \mathbf{R}^{-T}\mathbf{Q}^T\widetilde{\mathbf{b}}$ where $[\mathbf{A}^T \ \sqrt{\lambda}\mathbf{I}]^T = \mathbf{QR}$ and $\widetilde{\mathbf{b}}^T = [\mathbf{b}^T \ \mathbf{0}^T]$
  - ▶ Cholesky Dec., SVD, QR Dec. etc.[1]
  - ! $O(nd^2)$ complexity of full decomposition

Traditional Approaches: Solution Reconstruction for a Given λ

- Closed form solution: $\mathbf{x}(\lambda) = (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}_d)^{-1}\mathbf{A}^T\mathbf{b}$ where $\mathbf{A} \in \mathbb{R}^{n \times d}$
  - ! $O(nd^2)$ complexity of multiplication
  - ! Squares the condition number
- Direct methods: $\mathbf{x}(\lambda) = \mathbf{R}^{-T}\mathbf{Q}^T\widetilde{\mathbf{b}}$ where $[\mathbf{A}^T \ \sqrt{\lambda}\mathbf{I}]^T = \mathbf{QR}$ and $\widetilde{\mathbf{b}}^T = [\mathbf{b}^T \ \mathbf{0}^T]$
  - ▶ Cholesky Dec., SVD, QR Dec. etc.[1]
  - ! $O(nd^2)$ complexity of full decomposition
- First order iterative solvers
  - ▶ CG, LSQR, ART, Chebyshev, GMRES, LSMR etc.[2]
  - ✓ Requires a few matrix-vector or vector-vector multiplications per iteration
  - ✓ $O(nd)$ complexity per iteration

Traditional Approaches: Solution Reconstruction for a Given λ

- Closed form solution: $\mathbf{x}(\lambda) = (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}_d)^{-1}\mathbf{A}^T\mathbf{b}$ where $\mathbf{A} \in \mathbb{R}^{n\times d}$
  - ! $O(nd^2)$ complexity of multiplication
  - ! Squares the condition number
- Direct methods: $\mathbf{x}(\lambda) = \mathbf{R}^{-T}\mathbf{Q}^T\widetilde{\mathbf{b}}$ where $[\mathbf{A}^T \ \sqrt{\lambda}\mathbf{I}]^T = \mathbf{QR}$ and $\widetilde{\mathbf{b}}^T = [\mathbf{b}^T \ \mathbf{0}^T]$
  - ▶ Cholesky Dec., SVD, QR Dec. etc.[1]
  - ! $O(nd^2)$ complexity of full decomposition
- First order iterative solvers
  - ▶ CG, LSQR, ART, Chebyshev, GMRES, LSMR etc.[2]
  - ✓ Requires a few matrix-vector or vector-vector multiplications per iteration
  - ✓ $O(nd)$ complexity per iteration
  - ! Slow convergence:

$$\left\|\mathbf{x}^i - \mathbf{x}(\lambda)\right\|_2 \leq \left(\frac{\sqrt{\kappa(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}_d)} - 1}{\sqrt{\kappa(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}_d)} + 1}\right)^i \left\|\mathbf{x}^1 - \mathbf{x}(\lambda)\right\|_2, \ 1 < i,$$

Traditional Approaches: Computational Bottlenecks

- For the feasibility of the algorithms, in addition to the number of operations, there are two factors related to the number of iterations:

Traditional Approaches: Computational Bottlenecks

- For the feasibility of the algorithms, in addition to the number of operations, there are two factors related to the number of iterations:
    - Distributed matrix-vector multiplications:

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \sum_{\ell}^{N} \mathbf{A}_\ell^T \mathbf{A}_\ell \mathbf{x}, \text{ where } \mathbf{A} = [\mathbf{A}_1^T \ \dots \ \mathbf{A}_N^T]^T$$

Traditional Approaches: Computational Bottlenecks

- For the feasibility of the algorithms, in addition to the number of operations, there are two factors related to the number of iterations:
  - Distributed matrix-vector multiplications:

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \sum_{\ell}^{N} \mathbf{A}_\ell^T \mathbf{A}_\ell \mathbf{x}, \text{ where } \mathbf{A} = [\mathbf{A}_1^T \ \dots \ \mathbf{A}_N^T]^T$$

  - Synchronization steps induced by inner products:

$$\|\mathbf{b}\|_2^2 = \sum_{\ell}^{N} \|\mathbf{b}_\ell\|_2^2 \, , \text{ where } \mathbf{b} = [\mathbf{b}_1^T, \ \dots, \ \mathbf{b}_N^T]^T$$

Traditional Approaches: Computational Bottlenecks

- For the feasibility of the algorithms, in addition to the number of operations, there are two factors related to the number of iterations:
  - Distributed matrix-vector multiplications:

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \sum_{\ell}^{N} \mathbf{A}_\ell^T \mathbf{A}_\ell \mathbf{x}, \text{ where } \mathbf{A} = [\mathbf{A}_1^T \ \dots \ \mathbf{A}_N^T]^T$$

  - Synchronization steps induced by inner products:

$$\|\mathbf{b}\|_2^2 = \sum_{\ell}^{N} \|\mathbf{b}_\ell\|_2^2, \text{ where } \mathbf{b} = [\mathbf{b}_1^T, \ \dots, \ \mathbf{b}_N^T]^T$$

- Preconditioning could be a remedy: $\kappa(\mathbf{N}^T\mathbf{A}) \ll \kappa(\mathbf{A})$ or $\kappa(\mathbf{A}\mathbf{N}) \ll \kappa(\mathbf{A})$

$$\text{Left Preconditioning: } \mathbf{x}_{left} = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \ \left\|\mathbf{N}^T\mathbf{A}\mathbf{x} - \mathbf{N}^T\mathbf{b}\right\|_2^2,$$

$$\text{Right Preconditioning: } \mathbf{x}_{right} = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \ \|\mathbf{A}\mathbf{N}\mathbf{x} - \mathbf{b}\|_2^2,$$

Traditional Approaches: Computational Bottlenecks

- For the feasibility of the algorithms, in addition to the number of operations, there are two factors related to the number of iterations:
  - Distributed matrix-vector multiplications:

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \sum_{\ell}^{N} \mathbf{A}_\ell^T \mathbf{A}_\ell \mathbf{x}, \text{ where } \mathbf{A} = [\mathbf{A}_1^T \ \dots \ \mathbf{A}_N^T]^T$$

  - Synchronization steps induced by inner products:

$$\|\mathbf{b}\|_2^2 = \sum_{\ell}^{N} \|\mathbf{b}_\ell\|_2^2, \text{ where } \mathbf{b} = [\mathbf{b}_1^T, \ \dots, \ \mathbf{b}_N^T]^T$$

- Preconditioning could be a remedy: $\kappa(\mathbf{N}^T\mathbf{A}) \ll \kappa(\mathbf{A})$ or $\kappa(\mathbf{A}\mathbf{N}) \ll \kappa(\mathbf{A})$

$$\text{Left Preconditioning: } \mathbf{x}_{left} = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \ \left\|\mathbf{N}^T\mathbf{A}\mathbf{x} - \mathbf{N}^T\mathbf{b}\right\|_2^2,$$

$$\text{Right Preconditioning: } \mathbf{x}_{right} = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \ \|\mathbf{A}\mathbf{N}\mathbf{x} - \mathbf{b}\|_2^2,$$

$\mathbf{x}_{left} = \mathbf{x}_{\text{LS}}$ if $\mathcal{R}(\mathbf{N}\mathbf{N}^T\mathbf{A}) = \mathcal{R}(\mathbf{A})$ or $\mathbf{N}\mathbf{x}_{right} = \mathbf{x}_{\text{LS}}$ if $\mathcal{R}(\mathbf{N}\mathbf{N}^T\mathbf{A}) = \mathcal{R}(\mathbf{A}^T)$.

Traditional Approaches: Estimation of λ

- If $\mathbf{x}_0$ was available

$$\lambda = \underset{\lambda \in \mathbb{R}}{\text{argmin}} \ \|\mathbf{x}_0 - \mathbf{x}(\lambda)\|_2 \ \text{ or } \lambda = \underset{\lambda \in \mathbb{R}}{\text{argmin}} \ \|\mathbf{A}(\mathbf{x}_0 - \mathbf{x}(\lambda)\|_2$$

Traditional Approaches: Estimation of λ

- If $\mathbf{x}_0$ was available

$$\lambda = \underset{\lambda \in \mathbb{R}}{\text{argmin}} \; \|\mathbf{x}_0 - \mathbf{x}(\lambda)\|_2 \text{ or } \lambda = \underset{\lambda \in \mathbb{R}}{\text{argmin}} \; \|\mathbf{A}(\mathbf{x}_0 - \mathbf{x}(\lambda)\|_2$$

- Discrepancy Principle, UPRE, GSURE and GCV select λ as the minimizer of $T(\lambda)$ where

$$\mathbb{E}_{\mathbf{w}}\left[T(\lambda)\right] = \mathbb{E}_{\mathbf{w}}\left[\|\mathbf{x}_0 - \mathbf{x}(\lambda)\|_2\right] \text{ or } \mathbb{E}_{\mathbf{w}}\left[T(\lambda)\right] = \mathbb{E}_{\mathbf{w}}\left[\|\mathbf{A}(\mathbf{x}_0 - \mathbf{x}(\lambda)\|_2\right]$$

Traditional Approaches: Estimation of λ

- If $\mathbf{x}_0$ was available

$$\lambda = \underset{\lambda \in \mathbb{R}}{\text{argmin}} \; \|\mathbf{x}_0 - \mathbf{x}(\lambda)\|_2 \text{ or } \lambda = \underset{\lambda \in \mathbb{R}}{\text{argmin}} \; \|\mathbf{A}(\mathbf{x}_0 - \mathbf{x}(\lambda)\|_2$$

- Discrepancy Principle, UPRE, GSURE and GCV select λ as the minimizer of $T(\lambda)$ where

$$\mathbb{E}_{\mathbf{w}}\left[T(\lambda)\right] = \mathbb{E}_{\mathbf{w}}\left[\|\mathbf{x}_0 - \mathbf{x}(\lambda)\|_2\right] \text{ or } \mathbb{E}_{\mathbf{w}}\left[T(\lambda)\right] = \mathbb{E}_{\mathbf{w}}\left[\|\mathbf{A}(\mathbf{x}_0 - \mathbf{x}(\lambda)\|_2\right]$$

- Generalized Cross Validation[3] uses following unbiased estimator of the predictive risk

$$G_{full}(\lambda) = \frac{\|\mathbf{b} - \mathbf{A}\mathbf{x}(\lambda)\|_2}{\text{tr}\left(\mathbf{I} - P_{\mathbf{A}}(\lambda)\right)},$$

where $P_{\mathbf{A}}(\lambda) = \mathbf{A}\left(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}_d\right)^{-1}\mathbf{A}^T$ and $\text{sd}_\lambda(\mathbf{A}) = \text{tr}\left(P_{\mathbf{A}}(\lambda)\right)$.

Traditional Approaches: Estimation of λ

- If $\mathbf{x}_0$ was available

$$\lambda = \underset{\lambda \in \mathbb{R}}{\text{argmin}} \ \|\mathbf{x}_0 - \mathbf{x}(\lambda)\|_2 \ \text{ or } \lambda = \underset{\lambda \in \mathbb{R}}{\text{argmin}} \ \|\mathbf{A}(\mathbf{x}_0 - \mathbf{x}(\lambda)\|_2$$

- Discrepancy Principle, UPRE, GSURE and GCV select λ as the minimizer of $T(\lambda)$ where

$$\mathbb{E}_{\mathbf{w}}\left[T(\lambda)\right] = \mathbb{E}_{\mathbf{w}}\left[\|\mathbf{x}_0 - \mathbf{x}(\lambda)\|_2\right] \ \text{ or } \mathbb{E}_{\mathbf{w}}\left[T(\lambda)\right] = \mathbb{E}_{\mathbf{w}}\left[\|\mathbf{A}(\mathbf{x}_0 - \mathbf{x}(\lambda)\|_2\right]$$

- Generalized Cross Validation[3] uses following unbiased estimator of the predictive risk

$$G_{full}(\lambda) = \frac{\|\mathbf{b} - \mathbf{A}\mathbf{x}(\lambda)\|_2}{\text{tr}\left(\mathbf{I} - P_{\mathbf{A}}(\lambda)\right)},$$

where $P_{\mathbf{A}}(\lambda) = \mathbf{A}\left(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}_d\right)^{-1}\mathbf{A}^T$ and $\text{sd}_\lambda(\mathbf{A}) = \text{tr}\left(P_{\mathbf{A}}(\lambda)\right)$.

! Search for minimizer of $G_{full}(\lambda)$ is a major issue

Traditional Approaches: Hybrid Methods

- At the i^{th} iteration, LSQR[4] finds the solution of the following lower dimensional sub-problem: $(\beta_1 = \|\mathbf{b}\|_2)$

$$\mathbf{y}^i(\lambda) = \underset{\mathbf{y}\in\mathbb{R}^i}{\text{argmin}} \quad \|\mathbf{B}_i\mathbf{y} - \beta_1\mathbf{e}_1\|_2^2 + \lambda\,\|\mathbf{y}\|_2^2\,, \text{ where}$$

where $\mathbf{B}_i \in \mathbb{R}^{i+1\times i}$, $\mathbf{x}^i = \mathbf{Q}_i\mathbf{y}^i$ and $\text{span}(\mathbf{Q}_i) = \mathcal{K}_i(\mathbf{A}^T\mathbf{A},\ \mathbf{A}^T\mathbf{b})$

Traditional Approaches: Hybrid Methods

- At the i^{th} iteration, LSQR[4] finds the solution of the following lower dimensional sub-problem: $(\beta_1 = \|\mathbf{b}\|_2)$

$$\mathbf{y}^i(\lambda) = \underset{\mathbf{y}\in\mathbb{R}^i}{\text{argmin}} \quad \|\mathbf{B}_i\mathbf{y} - \beta_1\mathbf{e}_1\|_2^2 + \lambda\,\|\mathbf{y}\|_2^2\,, \text{ where}$$

where $\mathbf{B}_i \in \mathbb{R}^{i+1\times i}$, $\mathbf{x}^i = \mathbf{Q}_i\mathbf{y}^i$ and span$(\mathbf{Q}_i) = \mathcal{K}_i(\mathbf{A}^T\mathbf{A},\ \mathbf{A}^T\mathbf{b})$

- Hybrid-LSQR[5] selects λ that minimizes:

Traditional Approaches: Hybrid Methods

- At the i^{th} iteration, LSQR[4] finds the solution of the following lower dimensional sub-problem: $(\beta_1 = \|\mathbf{b}\|_2)$

$$\mathbf{y}^i(\lambda) = \underset{\mathbf{y}\in\mathbb{R}^i}{\text{argmin}} \quad \|\mathbf{B}_i\mathbf{y} - \beta_1\mathbf{e}_1\|_2^2 + \lambda\,\|\mathbf{y}\|_2^2\,, \text{ where}$$

where $\mathbf{B}_i \in \mathbb{R}^{i+1\times i}$, $\mathbf{x}^i = \mathbf{Q}_i\mathbf{y}^i$ and $\text{span}(\mathbf{Q}_i) = \mathcal{K}_i(\mathbf{A}^T\mathbf{A},\ \mathbf{A}^T\mathbf{b})$

- Hybrid-LSQR[5] selects λ that minimizes:

$$G_{proj}(\lambda) = \frac{\left\|\mathbf{B}_i\mathbf{y}^i(\lambda) - \beta_1\mathbf{e}_1\right\|_2}{\text{tr}\left(\mathbf{I}_{i+1} - P_{\mathbf{B}_i}(\lambda)\right)}$$

✓ Minimization of $G_{proj}(\lambda)$ requires $O(i)$ operations

Traditional Approaches: Hybrid Methods

- At the i^{th} iteration, LSQR[4] finds the solution of the following lower dimensional sub-problem: $(\beta_1 = \|\mathbf{b}\|_2)$

$$\mathbf{y}^i(\lambda) = \underset{\mathbf{y}\in\mathbb{R}^i}{\text{argmin}} \quad \|\mathbf{B}_i\mathbf{y} - \beta_1\mathbf{e}_1\|_2^2 + \lambda\,\|\mathbf{y}\|_2^2\,, \text{ where}$$

where $\mathbf{B}_i \in \mathbb{R}^{i+1\times i}$, $\mathbf{x}^i = \mathbf{Q}_i\mathbf{y}^i$ and $\text{span}(\mathbf{Q}_i) = \mathcal{K}_i(\mathbf{A}^T\mathbf{A},\ \mathbf{A}^T\mathbf{b})$

- Hybrid-LSQR[5] selects λ that minimizes:

$$G_{proj}(\lambda) = \frac{\left\|\mathbf{B}_i\mathbf{y}^i(\lambda) - \beta_1\mathbf{e}_1\right\|_2}{\text{tr}\left(\mathbf{I}_{i+1} - P_{\mathbf{B}_i}(\lambda)\right)}$$

✓ Minimization of $G_{proj}(\lambda)$ requires $O(i)$ operations

! To select a proper λ for the full problem, number of iterations i must be larger than k^*

! k^* scales with the dimension of the problem

Random Projection - I

- Reduces the dimension
- Bounds the number of iterations
- Convenient for parallel and distributed computations[6]

Random Projection - I

- Reduces the dimension
- Bounds the number of iterations
- Convenient for parallel and distributed computations[6]

Definition (Oblivious $\ell 2$ Subspace Embedding)

If a distribution $\mathcal{D}$ over $\mathbb{R}^{m\times n}$ satisfies the following concentration inequality

$$\mathbb{P}_{\mathbf{S}\sim\mathcal{D}}\left(\left\|\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U}-\mathbf{I}\right\|_2 > \epsilon\right) < \delta,$$

with $\forall\mathbf{U}\in\mathbb{R}^{n\times k},\ \mathbf{U}^T\mathbf{U}=\mathbf{I}_k,\ \mathbf{S}\in\mathbb{R}^{m\times n}$, then it is called (ϵ,δ,k)-OSE.

Random Projection - I

- Reduces the dimension
- Bounds the number of iterations
- Convenient for parallel and distributed computations[6]

Definition (Oblivious $\ell 2$ Subspace Embedding)

If a distribution $\mathcal{D}$ over $\mathbb{R}^{m\times n}$ satisfies the following concentration inequality

$$\mathbb{P}_{\mathbf{S}\sim\mathcal{D}}\left(\left\|\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U}-\mathbf{I}\right\|_2 > \epsilon\right) < \delta,$$

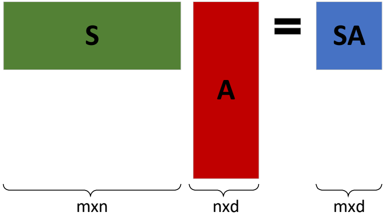
with $\forall \mathbf{U}\in\mathbb{R}^{n\times k},\ \mathbf{U}^T\mathbf{U}=\mathbf{I}_k,\ \mathbf{S}\in\mathbb{R}^{m\times n}$, then it is called (ϵ,δ,k)-OSE.

If the entries of $\mathbf{S}$ are drawn from $\mathcal{N}(0,1/m)$ and $m=O(\epsilon^{-2}\log(1/\delta))$, then $\mathbf{S}$ is an (ϵ,δ,n)-OSE[7], i.e., $\forall \mathbf{a}\in\mathbb{R}^n$, with probability of at least $1-\delta$:

$$(1-\epsilon)\left\|\mathbf{a}\right\|_2 \leq \left\|\mathbf{S}\mathbf{a}\right\|_2 \leq (1+\epsilon)\left\|\mathbf{a}\right\|_2$$

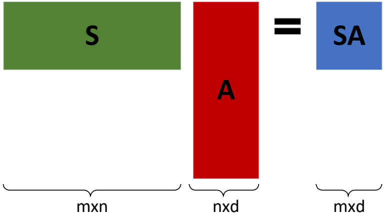
Random Projection - II

- Gaussian Sketches $\sim O(mnd)$

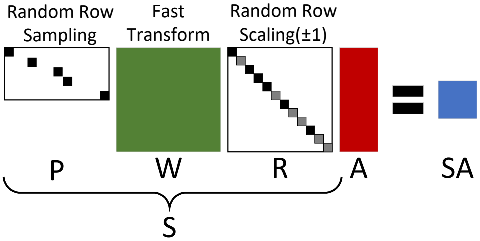


Random Projection - II

- Gaussian Sketches $\sim O(mnd)$

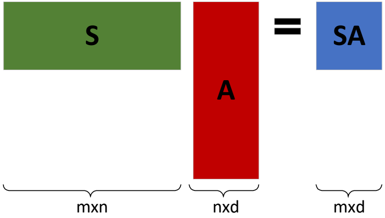


- Randomized Orthogonal Systems $\sim O(nd\log(m))$

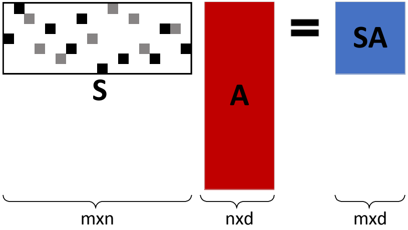


Random Projection - II

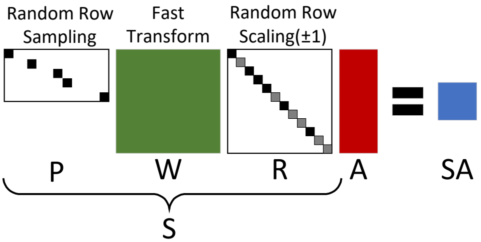
- Gaussian Sketches $\sim O(mnd)$



- CountSketch $\sim O(\text{nnz}(\mathbf{A}))$

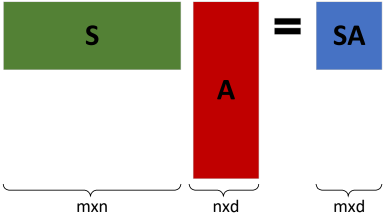


- Randomized Orthogonal Systems $\sim O(nd\log(m))$

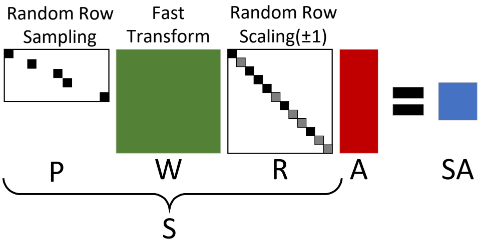


Random Projection - II

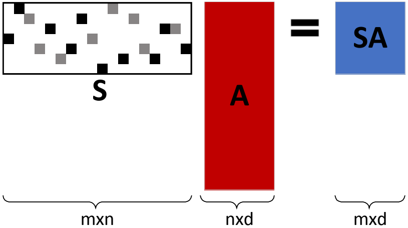
- Gaussian Sketches $\sim O(mnd)$



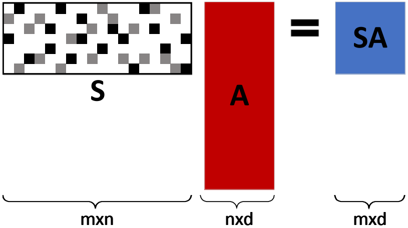
- Randomized Orthogonal Systems $\sim O(nd\log(m))$



- CountSketch $\sim O(\text{nnz}(\mathbf{A}))$



- Sparse Sketches $\sim O(s \cdot \text{nnz}(\mathbf{A}))$



RP-based Methods: Randomized Preconditioning

- Used for highly over-determined $(n \gg d)$ or higly under-determined $(n \ll d)$ problems

$$\mathbf{x}_{right} = \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^d} \ \|\mathbf{A}\mathbf{N}\mathbf{x} - \mathbf{b}\|_2^2$$

RP-based Methods: Randomized Preconditioning

- Used for highly over-determined $(n \gg d)$ or higly under-determined $(n \ll d)$ problems

$$\mathbf{x}_{right} = \underset{\mathbf{x} \in \mathbb{R}^d}{\text{argmin}} \quad \|\mathbf{ANx} - \mathbf{b}\|_2^2$$

- Blendenpik[8] sets $\mathbf{N} = \mathbf{R}_s^{-1}$ in LSQR where $\mathbf{SA} = \mathbf{Q}_s\mathbf{R}_s$ and $\mathbf{S}$ is ROS
- LSRN[9] sets $\mathbf{N} = \mathbf{V}_s\mathbf{\Sigma}_s^{-1}$ in LSQR and CS where $\mathbf{SA} = \mathbf{U}_s\mathbf{\Sigma}_s\mathbf{V}_s^T$ and $\mathbf{S}$ is Gaussian

RP-based Methods: Randomized Preconditioning

- Used for highly over-determined $(n \gg d)$ or higly under-determined $(n \ll d)$ problems

$$\mathbf{x}_{right} = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \quad \|\mathbf{A}\mathbf{N}\mathbf{x} - \mathbf{b}\|_2^2$$

- Blendenpik[8] sets $\mathbf{N} = \mathbf{R}_s^{-1}$ in LSQR where $\mathbf{S}\mathbf{A} = \mathbf{Q}_s\mathbf{R}_s$ and $\mathbf{S}$ is ROS
- LSRN[9] sets $\mathbf{N} = \mathbf{V}_s\mathbf{\Sigma}_s^{-1}$ in LSQR and CS where $\mathbf{S}\mathbf{A} = \mathbf{U}_s\mathbf{\Sigma}_s\mathbf{V}_s^T$ and $\mathbf{S}$ is Gaussian
- Iterative Hessian Sketch (IHS)[10] follows a different path

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x}\|_2^2 + \langle \mathbf{A}^T\mathbf{b},\ \mathbf{x}\rangle \approx \frac{1}{2}\|\mathbf{S}\mathbf{A}\mathbf{x}\|_2^2 + \langle \mathbf{A}^T\mathbf{b},\ \mathbf{x}\rangle$$

increases accuracy over iterations by using the true gradient:

$$\mathbf{x}^{i+1} = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}}\ \frac{1}{2}\left\|\mathbf{S}_i\mathbf{A}(\mathbf{x} - \mathbf{x}^i)\right\|_2^2 - \langle \mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x}^i),\ \mathbf{x}\rangle$$

$$= \mathbf{x}^i + \left(\mathbf{A}^T\mathbf{S}_i^T\mathbf{S}_i\mathbf{A}\right)^{-1}\mathbf{A}^T\left(\mathbf{b} - \mathbf{A}\mathbf{x}^i\right)$$

RP-based Methods: Randomized Preconditioning

- Used for highly over-determined $(n \gg d)$ or higly under-determined $(n \ll d)$ problems

$$\mathbf{x}_{right} = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \quad \|\mathbf{A}\mathbf{N}\mathbf{x} - \mathbf{b}\|_2^2$$

- Blendenpik[8] sets $\mathbf{N} = \mathbf{R}_s^{-1}$ in LSQR where $\mathbf{S}\mathbf{A} = \mathbf{Q}_s\mathbf{R}_s$ and $\mathbf{S}$ is ROS
- LSRN[9] sets $\mathbf{N} = \mathbf{V}_s\mathbf{\Sigma}_s^{-1}$ in LSQR and CS where $\mathbf{S}\mathbf{A} = \mathbf{U}_s\mathbf{\Sigma}_s\mathbf{V}_s^T$ and $\mathbf{S}$ is Gaussian
- Iterative Hessian Sketch (IHS)[10] follows a different path

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x}\|_2^2 + \langle \mathbf{A}^T\mathbf{b},\ \mathbf{x}\rangle \approx \frac{1}{2}\|\mathbf{S}\mathbf{A}\mathbf{x}\|_2^2 + \langle \mathbf{A}^T\mathbf{b},\ \mathbf{x}\rangle$$

increases accuracy over iterations by using the true gradient:

$$\begin{aligned}\mathbf{x}^{i+1} &= \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}}\ \frac{1}{2}\left\|\mathbf{S}_i\mathbf{A}(\mathbf{x}-\mathbf{x}^i)\right\|_2^2 - \langle \mathbf{A}^T(\mathbf{b}-\mathbf{A}\mathbf{x}^i),\ \mathbf{x}\rangle \\ &= \mathbf{x}^i + \left(\mathbf{A}^T\mathbf{S}_i^T\mathbf{S}_i\mathbf{A}\right)^{-1}\mathbf{A}^T\left(\mathbf{b}-\mathbf{A}\mathbf{x}^i\right)\end{aligned}$$

  - $\mathbf{S}_i = \mathbf{S}$ can be used for all iterations, but might cause divergence[11].

RP-based Methods: Randomized Preconditioning

- Used for highly over-determined ($n \gg d$) or higly under-determined ($n \ll d$) problems

$$\mathbf{x}_{right} = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \quad \|\mathbf{A}\mathbf{N}\mathbf{x} - \mathbf{b}\|_2^2$$

- Blendenpik[8] sets $\mathbf{N} = \mathbf{R}_s^{-1}$ in LSQR where $\mathbf{S}\mathbf{A} = \mathbf{Q}_s\mathbf{R}_s$ and $\mathbf{S}$ is ROS
- LSRN[9] sets $\mathbf{N} = \mathbf{V}_s\mathbf{\Sigma}_s^{-1}$ in LSQR and CS where $\mathbf{S}\mathbf{A} = \mathbf{U}_s\mathbf{\Sigma}_s\mathbf{V}_s^T$ and $\mathbf{S}$ is Gaussian
- Iterative Hessian Sketch (IHS)[10] follows a different path

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x}\|_2^2 + \langle \mathbf{A}^T\mathbf{b},\ \mathbf{x}\rangle \approx \frac{1}{2}\|\mathbf{S}\mathbf{A}\mathbf{x}\|_2^2 + \langle \mathbf{A}^T\mathbf{b},\ \mathbf{x}\rangle$$

increases accuracy over iterations by using the true gradient:

$$\mathbf{x}^{i+1} = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}}\ \frac{1}{2}\left\|\mathbf{S}_i\mathbf{A}(\mathbf{x} - \mathbf{x}^i)\right\|_2^2 - \langle \mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x}^i),\ \mathbf{x}\rangle$$
$$= \mathbf{x}^i + \left(\mathbf{A}^T\mathbf{S}_i^T\mathbf{S}_i\mathbf{A}\right)^{-1}\mathbf{A}^T\left(\mathbf{b} - \mathbf{A}\mathbf{x}^i\right)$$

  - $\mathbf{S}_i = \mathbf{S}$ can be used for all iterations, but might cause divergence[11].
  - Accelerated-IHS (A-IHS)[12] uses CG instead of GD to prevent divergence.

Outline

1. Review of Linear Least Squares Problems
   - Problem Formulation
   - Traditional Approaches
     - Reconstruction for a given regularization parameter
     - Methods for estimation of the unknown regularization parameter
   - Random Projection Based Approaches for the LS Problems
2. Proposed Momentum Iterative Hessian Sketch (M-IHS) Techniques
   - Techniques for a Given Regularization Parameter
   - Hybrid Techniques to Estimate Unknown Regularization Parameter
3. Conclusions and Future Work

Proposed M-IHS: Derivation

$$\mathbf{x}^{i+1} = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \; \left\|\mathbf{S}_i\mathbf{A}(\mathbf{x}-\mathbf{x}^i)\right\|_2^2 + \lambda\,\|\mathbf{x}\|_2^2 - 2\langle\mathbf{A}^T(\mathbf{b}-\mathbf{A}\mathbf{x}^i) - \lambda\mathbf{x}^i,\; \mathbf{x}\rangle$$

- ? Can we avoid change of $\mathbf{S}$ at every iteration?
- ? Can we accelerate the convergence of the iterations?

Proposed M-IHS: Derivation

$$\mathbf{x}^{i+1} = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \left\|\mathbf{S}_i\mathbf{A}(\mathbf{x}-\mathbf{x}^i)\right\|_2^2 + \lambda\left\|\mathbf{x}\right\|_2^2 - 2\langle\mathbf{A}^T(\mathbf{b}-\mathbf{A}\mathbf{x}^i) - \lambda\mathbf{x}^i,\ \mathbf{x}\rangle$$

- ? Can we avoid change of $\mathbf{S}$ at every iteration?
- ? Can we accelerate the convergence of the iterations?
- Yes, both can be realizable via Heavy Ball Method (HBM):

$$\mathbf{x}^{i+1} = \mathbf{x}^i + \alpha_i\nabla f(\mathbf{x}^i) + \beta_i(\mathbf{x}^i - \mathbf{x}^{i-1})$$

Proposed M-IHS: Derivation

$$\mathbf{x}^{i+1} = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \left\|\mathbf{S}_i\mathbf{A}(\mathbf{x}-\mathbf{x}^i)\right\|_2^2 + \lambda\|\mathbf{x}\|_2^2 - 2\langle\mathbf{A}^T(\mathbf{b}-\mathbf{A}\mathbf{x}^i) - \lambda\mathbf{x}^i,\ \mathbf{x}\rangle$$

? Can we avoid change of $\mathbf{S}$ at every iteration?

? Can we accelerate the convergence of the iterations?

- Yes, both can be realizable via Heavy Ball Method (HBM):

$$\mathbf{x}^{i+1} = \mathbf{x}^i + \alpha_i \nabla f(\mathbf{x}^i) + \beta_i(\mathbf{x}^i - \mathbf{x}^{i-1})$$

- The optimal fixed momentum parameters for LS problems are

$$\alpha^* = \frac{4}{(\sqrt{\sigma_1}+\sqrt{\sigma_d})^2}, \qquad \beta^* = \frac{\sqrt{\sigma_1}-\sqrt{\sigma_d}}{\sqrt{\sigma_1}+\sqrt{\sigma_d}}$$

Proposed M-IHS: Derivation

$$\mathbf{x}^{i+1} = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \left\|\mathbf{S}_i\mathbf{A}(\mathbf{x}-\mathbf{x}^i)\right\|_2^2 + \lambda\|\mathbf{x}\|_2^2 - 2\langle\mathbf{A}^T(\mathbf{b}-\mathbf{A}\mathbf{x}^i) - \lambda\mathbf{x}^i,\ \mathbf{x}\rangle$$

? Can we avoid change of $\mathbf{S}$ at every iteration?

? Can we accelerate the convergence of the iterations?

- Yes, both can be realizable via Heavy Ball Method (HBM):

$$\mathbf{x}^{i+1} = \mathbf{x}^i + \alpha_i \nabla f(\mathbf{x}^i) + \beta_i(\mathbf{x}^i - \mathbf{x}^{i-1})$$

- The optimal fixed momentum parameters for LS problems are

$$\alpha^* = \frac{4}{(\sqrt{\sigma_1}+\sqrt{\sigma_d})^2}, \qquad \beta^* = \frac{\sqrt{\sigma_1}-\sqrt{\sigma_d}}{\sqrt{\sigma_1}+\sqrt{\sigma_d}}$$

- Momentum-IHS is obtained by incorporating the HBM into the IHS updates:

$$\Delta\mathbf{x}^i = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \ \|\mathbf{S}\mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_2^2 - 2\left\langle\mathbf{A}^T(\mathbf{b}-\mathbf{A}\mathbf{x}^i) - \lambda\mathbf{x}^i,\ \mathbf{x}\right\rangle,$$

$$\mathbf{x}^{i+1} = \mathbf{x}^i + \alpha\Delta\mathbf{x}^i + \beta\left(\mathbf{x}^i - \mathbf{x}^{i-1}\right),$$

Proposed M-IHS: Extension to Under-determined Regime

- A dual of the regularized LS problem is:

$$\boldsymbol{\nu}(\lambda) = \underset{\boldsymbol{\nu} \in \mathbb{R}^n}{\text{argmin}} \quad \underbrace{\frac{1}{2}\left\|\mathbf{A}^T\boldsymbol{\nu}\right\|_2^2 + \frac{\lambda}{2}\left\|\boldsymbol{\nu}\right\|_2^2 - \langle \mathbf{b},\ \boldsymbol{\nu}\rangle}_{g(\boldsymbol{\nu},\lambda)},$$

and the relation between the solutions is

$$\boldsymbol{\nu}(\lambda) = (\mathbf{b} - \mathbf{A}\mathbf{x}(\lambda))/\lambda \iff \mathbf{x}(\lambda) = \mathbf{A}^T\boldsymbol{\nu}(\lambda).$$

Proposed M-IHS: Extension to Under-determined Regime

- A dual of the regularized LS problem is:

$$\boldsymbol{\nu}(\lambda) = \underset{\boldsymbol{\nu}\in\mathbb{R}^n}{\text{argmin}} \ \underbrace{\frac{1}{2}\left\|\mathbf{A}^T\boldsymbol{\nu}\right\|_2^2 + \frac{\lambda}{2}\left\|\boldsymbol{\nu}\right\|_2^2 - \langle \mathbf{b},\ \boldsymbol{\nu}\rangle}_{g(\boldsymbol{\nu},\lambda)},$$

and the relation between the solutions is

$$\boldsymbol{\nu}(\lambda) = (\mathbf{b} - \mathbf{A}\mathbf{x}(\lambda))/\lambda \Longleftrightarrow \mathbf{x}(\lambda) = \mathbf{A}^T\boldsymbol{\nu}(\lambda).$$

- The `Dual M-IHS` uses following updates:

$$\Delta\boldsymbol{\nu}^i = \underset{\boldsymbol{\nu}\in\mathbb{R}^n}{\text{argmin}} \ \left\|\mathbf{S}\mathbf{A}^T\boldsymbol{\nu}\right\|_2^2 + \lambda\left\|\boldsymbol{\nu}\right\|_2^2 - 2\left\langle \mathbf{b} - \mathbf{A}\mathbf{A}^T\boldsymbol{\nu}^i - \lambda\boldsymbol{\nu}^i),\ \boldsymbol{\nu}\right\rangle,$$
$$\boldsymbol{\nu}^{i+1} = \boldsymbol{\nu}^i + \alpha\Delta\boldsymbol{\nu}^i + \beta\left(\boldsymbol{\nu}^i - \boldsymbol{\nu}^{i-1}\right).$$

Proposed M-IHS: Convergence Properties

Theorem (Non-asymptotic Analysis)

Let $\mathbf{U}_1 \in \mathbb{R}^{n \times d}$ *consists of the first* n *rows of an orthogonal basis for* $[\mathbf{A}^T \ \sqrt{\lambda}\mathbf{I}_d]^T$. *Let the sketching matrix* $\mathbf{S} \in \mathbb{R}^{m \times n}$ *be drawn from a distribution* $\mathcal{D}$ *such that*

$$\mathbb{P}_{\mathbf{S}\sim\mathcal{D}}\left(\left\|\mathbf{U}_1^T\mathbf{S}^T\mathbf{S}\mathbf{U}_1 - \mathbf{U}_1^T\mathbf{U}_1\right\|_2 \geq \epsilon\right) < \delta, \quad \epsilon \in (0,1).$$

Then, the `M-IHS` *with the following momentum parameters*

$$\beta^* = \left(\epsilon \Big/ \left(1 + \sqrt{1-\epsilon^2}\right)\right)^2, \qquad \alpha^* = (1-\beta^*)\sqrt{1-\epsilon^2},$$

converges to the optimal solution $\mathbf{x}(\lambda)$ *at the following rate with a probability of at least* $(1-\delta)$:

$$\left\|\mathbf{x}^{i+1} - \mathbf{x}(\lambda)\right\|_{\mathbf{D}_\lambda^{-1}} \leq \frac{\epsilon}{1+\sqrt{1-\epsilon^2}} \left\|\mathbf{x}^{i} - \mathbf{x}(\lambda)\right\|_{\mathbf{D}_\lambda^{-1}},$$

where $\mathbf{D}_\lambda^{-1}$ *is the diagonal matrix whose diagonal entries are* $\sqrt{\sigma_i^2 + \lambda}$, $1 \leq i \leq d$.

Proposed M-IHS: Total Number of Iterations

Corollary

For some $\epsilon \in (0, 1/2)$ *and arbitrary* η*, the number of iterations for the* `M-IHS` *to obtain an* η*-optimal solution approximation in* $\ell 2$*-norm is upper bounded by*

$$N = \left\lceil \frac{\log(\eta)\log(C)}{\log(\epsilon) - \log(1 + \sqrt{1 - \epsilon^2})} \right\rceil$$

where the constant $C = \sqrt{\kappa(\mathbf{A}^T\mathbf{A} + \lambda \mathbf{I}_d)}$

$$\left\| \mathbf{x}^N - \mathbf{x}(\lambda) \right\|_{\mathbf{2}} \leq \eta \left\| \mathbf{x}(\lambda) \right\|_{\mathbf{2}}$$

Proposed M-IHS: Sketch Size

Lemma (Lower Bounds on the Sketch Size)

If the sketching matrix $\mathbf{S} \in \mathbb{R}^{m \times n}$ *is chosen in one of the following cases, then the condition in the theorem*

$$\mathbb{P}_{\mathbf{S}\sim\mathcal{D}}\left(\left\|\mathbf{U}_1^T\mathbf{S}^T\mathbf{S}\mathbf{U}_1 - \mathbf{U}_1^T\mathbf{U}_1\right\|_2 \geq \epsilon\right) < \delta, \quad \epsilon \in (0,1)$$

is satisfied.

1. $\mathbf{S}$ *is a CountSketch with*
$$m = \Omega\left(\mathbf{sd}_\lambda(\mathbf{A})^2/(\epsilon^2\delta)\right)$$
2. $\mathbf{S}$ *is a Sub-Gaussian sketching matrix with*
$$m = \Omega(\mathsf{sd}_\lambda(\mathbf{A})/\epsilon^2)$$
3. $\mathbf{S}$ *is a ROS matrix with*
$$m = \Omega\left(\left(\mathbf{sd}_\lambda(\mathbf{A}) + \log(1/\epsilon\delta)\log(\mathbf{sd}_\lambda(\mathbf{A})/\delta)\right)/\epsilon^2\right)$$
4. $\mathbf{S}$ *is a Sparse Sketching with*
$$s = \Omega(\log_\alpha(\mathsf{sd}_\lambda(\mathbf{A})/\delta)/\epsilon)$$
non-zero elements in each column and
$$m = \Omega(\alpha \cdot \mathsf{sd}_\lambda(\mathbf{A})\log(\mathsf{sd}_\lambda(\mathbf{A})/\delta)/\epsilon^2)$$
where $\alpha > 2$, $\delta < 1/2$, $\epsilon < 1/2$

Proposed M-IHS: Empirical Convergence

Remark (Asymptotic Analysis)

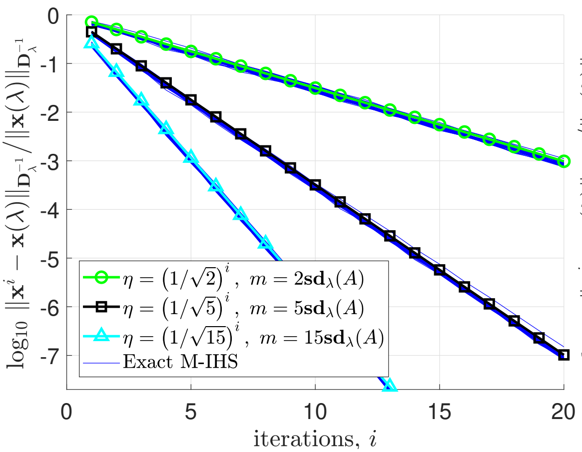
If the entries of the sketching matrix are independent, zero mean, unit variance with bounded higher order moments, then the `M-IHS` and the `Dual M-IHS` with the following momentum parameters

$$\beta = \frac{\mathsf{sd}_\lambda(\mathbf{A})}{m}, \qquad \alpha = (1-\beta)^2$$

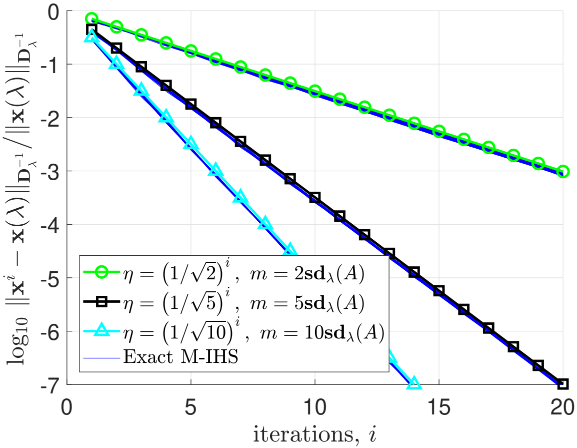
will converge to the optimal solutions with a convergence rate of $\sqrt{\beta}$ as $m \to \infty$ while $\mathsf{sd}_\lambda(\mathbf{A})/m$ remains constant. Any sketch size $m > \mathsf{sd}_\lambda(\mathbf{A})$ can be chosen to obtain an η-optimal solution approximation in at most $\dfrac{\log(\eta)}{\log(\sqrt{\beta})}$ iterations.

$$\left\|\mathbf{x}^i - \mathbf{x}(\lambda)\right\|_{\mathbf{D}_\lambda^{-1}} \leq \left(\sqrt{\frac{\mathsf{sd}_\lambda(\mathbf{A})}{m}}\right)^i \|\mathbf{x}(\lambda)\|_{\mathbf{D}_\lambda^{-1}}$$

Proposed M-IHS: Theoretical vs Numerical Convergence



(a) Dense problem with size 32000×1000 $\kappa(\mathbf{A}) = 10^8$, $\mathrm{sd}_\lambda(\mathbf{A}) = 119$, and ROS matrix via DCT



(b) Sparse problem with size 24000×1200, $\kappa(\mathbf{A}) = 10^7$, sparsity ratio 0.1%, $\mathrm{sd}_\lambda(\mathbf{A}) = 410$, and CountSketch

Proposed M-IHS: Inexact Sub-solver

- The next step in the `M-IHS` updates:

$$\Delta \mathbf{x}^i = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \quad \|\mathbf{SAx}\|_2^2 + \lambda \|\mathbf{x}\|_2^2 + 2\left\langle \nabla f(\mathbf{x}^i, \lambda),\ \mathbf{x}\right\rangle$$

can be obtained by solving the following lower dimensional sub-problems

$$\left((\mathbf{SA})^T(\mathbf{SA}) + \lambda \mathbf{I}_d\right)\Delta \mathbf{x}^i = -\nabla f(\mathbf{x}^i, \lambda).$$

Proposed M-IHS: Inexact Sub-solver

- The next step in the `M-IHS` updates:

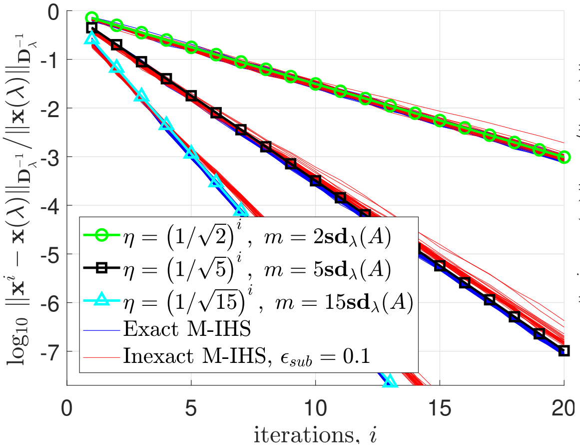
$$\Delta \mathbf{x}^i = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \quad \|\mathbf{SAx}\|_2^2 + \lambda \|\mathbf{x}\|_2^2 + 2\left\langle \nabla f(\mathbf{x}^i, \lambda),\ \mathbf{x}\right\rangle$$

  can be obtained by solving the following lower dimensional sub-problems

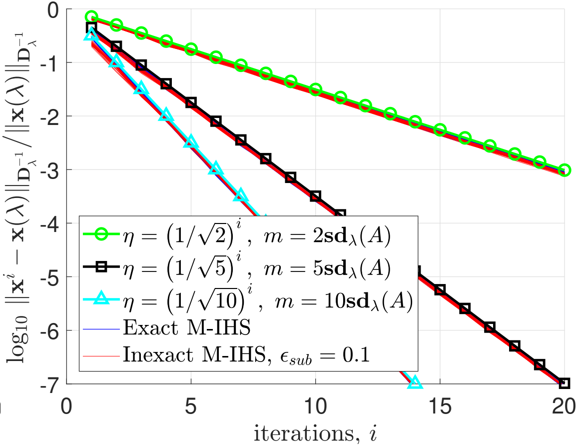
$$\left((\mathbf{SA})^T(\mathbf{SA}) + \lambda \mathbf{I}_d\right)\Delta \mathbf{x}^i = -\nabla f(\mathbf{x}^i, \lambda).$$

- We introduced `AAb_Solver` for the the problems in the form of $\mathbf{A}^T\mathbf{Ax} = \mathbf{b}$.
  - Does not square the condition number
  - More stable than symmetric CG or Lanczos Tridiagonalization algorithms
  - Stopping criterion: $\epsilon_{sub} \geq \frac{\left\|\mathbf{A}^T\mathbf{Ax}^i - \mathbf{b}\right\|_2}{\|\mathbf{b}\|_2}$
  - Computes the solution in $O(md)$ operations

Proposed M-IHS: Theoretical vs Numerical Convergence



(a) Dense problem with size 32000×1000, $\kappa(\mathbf{A}) = 10^8$, $\mathsf{sd}_\lambda(\mathbf{A}) = 119$, and ROS matrix via DCT



(b) Sparse problem with size 24000×1200, $\kappa(\mathbf{A}) = 10^7$, sparsity ratio 0.1%, $\mathsf{sd}_\lambda(\mathbf{A}) = 410$, and CountSketch

Proposed M-IHS: Overall Algorithms

M-IHS (for $n \geq d$)

Input: $\mathbf{A}$, $\mathbf{b}$, m, λ, $\mathbf{x}^1$, $\mathsf{sd}_\lambda(\mathbf{A})$, ϵ_{sub}

$\mathbf{SA} = \texttt{RP\_fun}(\mathbf{A}, m)$

$\beta = \mathsf{sd}_\lambda(\mathbf{A})/m, \quad \alpha = (1-\beta)^2$

while *until stopping criteria* **do**

$\quad \mathbf{g}^i = \mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x}^i) - \lambda\mathbf{x}^i$

$\quad \Delta\mathbf{x}^i = \texttt{AAb\_Solver}(\mathbf{SA},\ \mathbf{g}^i,\ \lambda,\ \epsilon_{sub})$

$\quad \mathbf{x}^{i+1} = \mathbf{x}^i + \alpha\Delta\mathbf{x}^i + \beta(\mathbf{x}^i - \mathbf{x}^{i-1})$

end while

Dual M-IHS (for $n \leq d$)

Input: $\mathbf{A}$, $\mathbf{b}$, m, λ, $\mathsf{sd}_\lambda(\mathbf{A})$, ϵ_{sub}

$\mathbf{SA}^T = \texttt{RP\_fun}(\mathbf{A}^T, m)$

$\beta = \mathsf{sd}_\lambda(\mathbf{A})/m, \ \alpha = (1-\beta)^2, \boldsymbol{\nu}^0 = 0$

while *until stopping criteria* **do**

$\quad \mathbf{g}^i = \mathbf{b} - \mathbf{A}\mathbf{A}^T\boldsymbol{\nu}^i - \lambda\boldsymbol{\nu}^i$

$\quad \Delta\boldsymbol{\nu}^i = \texttt{AAb\_Solver}(\mathbf{SA}^T,\ \mathbf{g}^i,\ \lambda,\ \epsilon_{sub})$

$\quad \boldsymbol{\nu}^{i+1} = \boldsymbol{\nu}^i + \alpha\Delta\boldsymbol{\nu}^i + \beta(\boldsymbol{\nu}^i - \boldsymbol{\nu}^{i-1})$

end while

Proposed M-IHS: An Observation

The following linear systems is $d \times d$ dimensional

$$\left((\mathbf{SA})^T(\mathbf{SA}) + \lambda \mathbf{I}_d\right) \Delta \mathbf{x}^i = -\nabla f(\mathbf{x}^i, \lambda).$$

where $\mathbf{SA} \in \mathbb{R}^{m \times n}$ with $m \sim \text{sd}_\lambda(\mathbf{A}) \ll n, d$

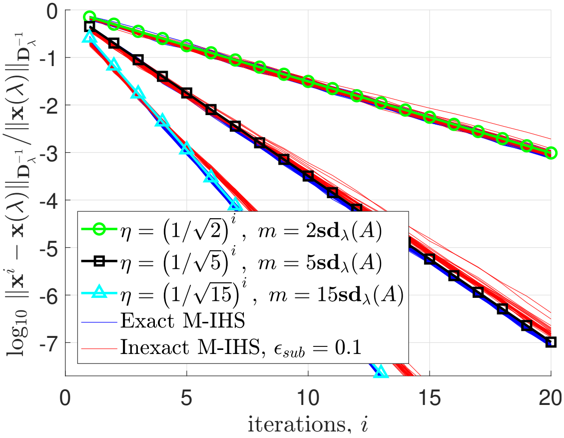


Figure: Dense problem with size 32000×1000 $\kappa(\mathbf{A}) = 10^8$, $\text{sd}_\lambda(\mathbf{A}) = 119$, and ROS matrix via DCT

Proposed M-IHS: Two-Stage Sketching

The dual of the problem

$$\Delta \mathbf{x}^i = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \quad \|\mathbf{S}\mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_2^2 + 2\left\langle \nabla f(\mathbf{x}^i, \lambda),\ \mathbf{x}\right\rangle$$

is a highly over-determined $d \times m$ dimensional problem:

$$\mathbf{z}^* = \underset{\mathbf{z}\in\mathbb{R}^m}{\text{argmin}} \quad \underbrace{\frac{1}{2}\left\|\mathbf{A}^T\mathbf{S}^T\mathbf{z} + \nabla f(\mathbf{x}^i, \lambda)\right\|_2^2 + \frac{\lambda}{2}\|\mathbf{z}\|_2^2}_{h(\mathbf{z},\mathbf{x}^i,\lambda)},$$

with $\Delta \mathbf{x}^i = (\nabla f(\mathbf{x}^i, \lambda) - \mathbf{A}^T\mathbf{S}^T\mathbf{z}^*)/\lambda$.

Proposed M-IHS: Two-Stage Sketching

The dual of the problem

$$\Delta \mathbf{x}^i = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^d} \ \|\mathbf{S}\mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_2^2 + 2\left\langle \nabla f(\mathbf{x}^i,\lambda),\ \mathbf{x}\right\rangle$$

is a highly over-determined $d \times m$ dimensional problem:

$$\mathbf{z}^* = \operatorname*{argmin}_{\mathbf{z}\in\mathbb{R}^m} \ \underbrace{\frac{1}{2}\left\|\mathbf{A}^T\mathbf{S}^T\mathbf{z} + \nabla f(\mathbf{x}^i,\lambda)\right\|_2^2 + \frac{\lambda}{2}\|\mathbf{z}\|_2^2}_{h(\mathbf{z},\mathbf{x}^i,\lambda)},$$

with $\Delta \mathbf{x}^i = (\nabla f(\mathbf{x}^i,\lambda) - \mathbf{A}^T\mathbf{S}^T\mathbf{z}^*)/\lambda$. Another RP can be applied through $\mathbf{W} \in \mathbb{R}^{m_2 \times d}$ as

$$\Delta \mathbf{z}^j = \operatorname*{argmin}_{z\in\mathbb{R}^m} \ \left\|\mathbf{W}\mathbf{A}^T\mathbf{S}^T\mathbf{z}\right\|_2^2 + \lambda\|\mathbf{z}\|_2^2 + 2\left\langle \nabla_{\mathbf{z}} h(\mathbf{z}^j,\mathbf{x}^i,\lambda),\ \mathbf{z}\right\rangle,$$

$$\mathbf{z}^{j+1} = \mathbf{z}^j + \alpha_2 \Delta \mathbf{z}^j + \beta_2\left(\mathbf{z}^j - \mathbf{z}^{j-1}\right),$$

where $\beta_2 = \mathrm{sd}_\lambda(\mathbf{A})/m_2$ and $\alpha_2 = (1-\beta_2)^2$.

Proposed M-IHS: Two-Stage Sketching

The dual of the problem

$$\Delta \mathbf{x}^i = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \quad \|\mathbf{S}\mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_2^2 + 2\left\langle \nabla f(\mathbf{x}^i, \lambda),\ \mathbf{x}\right\rangle$$

is a highly over-determined $d \times m$ dimensional problem:

$$\mathbf{z}^* = \underset{\mathbf{z}\in\mathbb{R}^m}{\text{argmin}} \quad \underbrace{\frac{1}{2}\left\|\mathbf{A}^T\mathbf{S}^T\mathbf{z} + \nabla f(\mathbf{x}^i, \lambda)\right\|_2^2 + \frac{\lambda}{2}\|\mathbf{z}\|_2^2,}_{h(\mathbf{z},\mathbf{x}^i,\lambda)}$$

with $\Delta \mathbf{x}^i = (\nabla f(\mathbf{x}^i, \lambda) - \mathbf{A}^T\mathbf{S}^T\mathbf{z}^*)/\lambda$. Another RP can be applied through $\mathbf{W} \in \mathbb{R}^{m_2 \times d}$ as

$$\Delta \mathbf{z}^j = \underset{z\in\mathbb{R}^m}{\text{argmin}} \quad \left\|\mathbf{W}\mathbf{A}^T\mathbf{S}^T\mathbf{z}\right\|_2^2 + \lambda\|\mathbf{z}\|_2^2 + 2\left\langle \nabla_{\mathbf{z}} h(\mathbf{z}^j, \mathbf{x}^i, \lambda),\ \mathbf{z}\right\rangle,$$

$$\mathbf{z}^{j+1} = \mathbf{z}^j + \alpha_2\Delta\mathbf{z}^j + \beta_2\left(\mathbf{z}^j - \mathbf{z}^{j-1}\right),$$

where $\beta_2 = \text{sd}_\lambda(\mathbf{A})/m_2$ and $\alpha_2 = (1-\beta_2)^2$. `AAb_Solver` can be used for the sub-problems:

$$\left((\mathbf{W}\mathbf{A}^T\mathbf{S}^T)^T(\mathbf{W}\mathbf{A}^T\mathbf{S}^T) + \lambda\mathbf{I}\right)\Delta\mathbf{z}^j = -\nabla h(\mathbf{z}^j, \mathbf{x}^i, \lambda)$$

Primal Dual M-IHS (for $n \geq d$)

Input: $\mathbf{A},\ \mathbf{b},\ m_1,\ m_2,\ \lambda,\ \mathsf{sd}_\lambda(\mathbf{A}),\ \epsilon_{sub}$

$\mathbf{SA} = \texttt{RP\_fun}(\mathbf{A}, m_1)$

$\mathbf{WA}^T\mathbf{S}^T = \texttt{RP\_fun}(\mathbf{A}^T\mathbf{S}^T, m_2)$

$\beta_\ell = \mathsf{sd}_\lambda(\mathbf{A})/m_\ell, \quad \ell = 1, 2$

$\alpha_\ell = (1-\beta_\ell)^2 \quad \ell = 1, 2$

$\mathbf{x}^0 = 0,\ \mathbf{z}^{1,0} = 0$

for i=1:N **do**

  $\mathbf{b}^i = \mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x}^i) - \lambda\mathbf{x}^i$

  for j=1:M **do**

    $g^{i,j} = \mathbf{SA}(\mathbf{b}^i - \mathbf{A}^T\mathbf{S}^T\mathbf{z}^{i,j}) - \lambda\mathbf{z}^{i,j}$

    $\Delta\mathbf{z}^{i,j} = \texttt{AAb\_Solver}(\mathbf{WA}^T\mathbf{S}^T, \mathbf{g}^{i,j},\ \lambda,\ \epsilon_{sub})$

    $\mathbf{z}^{i,j+1} = \mathbf{z}^{i,j} + \alpha_2\Delta\mathbf{z}^{i,j} + \beta_2(\mathbf{z}^{i,j} - \mathbf{z}^{i,j-1})$

  end for

  $\Delta\mathbf{x}^i = (\mathbf{b}^i - \mathbf{A}^T\mathbf{S}^T\mathbf{z}^{i,M+1})/\lambda,\ \mathbf{z}^{1,0} = \mathbf{z}^{M+1,M}$

  $\mathbf{x}^{i+1} = \mathbf{x}^i + \alpha_1\Delta\mathbf{x}^i + \beta_1(\mathbf{x}^i - \mathbf{x}^{i-1})$

end for

Primal Dual M-IHS (for $n \leq d$)

Input: $\mathbf{A},\ \mathbf{b},\ m_1,\ m_2,\ \lambda,\ \mathsf{sd}_\lambda(\mathbf{A}),\ \epsilon_{sub}$

$\mathbf{SA}^T = \texttt{RP\_fun}(\mathbf{A}^T, m_1)$

$\mathbf{WAS}^T = \texttt{RP\_fun}(\mathbf{SA}^T, m_2)$

$\beta_\ell = \mathsf{sd}_\lambda(\mathbf{A})/m_\ell, \quad \ell = 1, 2$

$\alpha_\ell = (1-\beta_\ell)^2, \quad \ell = 1, 2$

$\boldsymbol{\nu}^{1,0} = 0,\ z^{1,0} = 0$

for i=1:N **do**

  $\mathbf{b}^i = \mathbf{b} - \mathbf{AA}^T\boldsymbol{\nu}^i - \lambda\boldsymbol{\nu}^i$

  for j=1:M **do**

    $\mathbf{g}^{i,j} = \mathbf{SA}^T(\mathbf{b}^i - \mathbf{AS}^T\mathbf{z}^{i,j}) - \lambda\mathbf{z}^{i,j}$

    $\Delta\mathbf{z}^{i,j} = \texttt{AAb\_Solver}(\mathbf{WAS}^T, \mathbf{g}^{i,j},\ \lambda,\ \epsilon_{sub})$

    $\mathbf{z}^{i,j+1} = \mathbf{z}^{i,j} + \alpha_2\Delta\mathbf{z}^{i,j} + \beta_2(\mathbf{z}^{i,j} - \mathbf{z}^{i,j-1})$

  end for

  $\Delta\boldsymbol{\nu}^i = (\mathbf{b}^i - \mathbf{AS}^T\mathbf{z}^{i,M+1})/\lambda, \quad \mathbf{z}^{1,0} = \mathbf{z}^{M+1,M}$

  $\boldsymbol{\nu}^{i+1} = \boldsymbol{\nu}^i + \alpha_1\Delta\boldsymbol{\nu}^i + \beta_1(\boldsymbol{\nu}^i - \boldsymbol{\nu}^{i-1})$

end for

Experiments: Un-regularized Problems

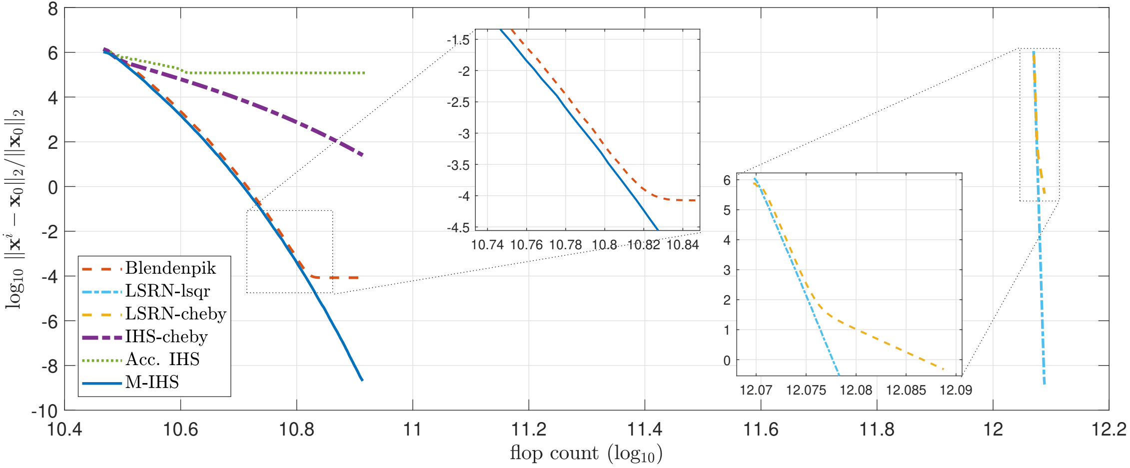


Figure: Performance comparison on an un-regularized LS problem with size $2^{16} \times 2000$ and $\kappa(\mathbf{A}) = 10^8$. In order to compare the convergence rates, number of iterations for all solvers are set to $N = 100$ with the same sketch size: $m = 4000$.

Experiments: Over-determined Regularized Problems

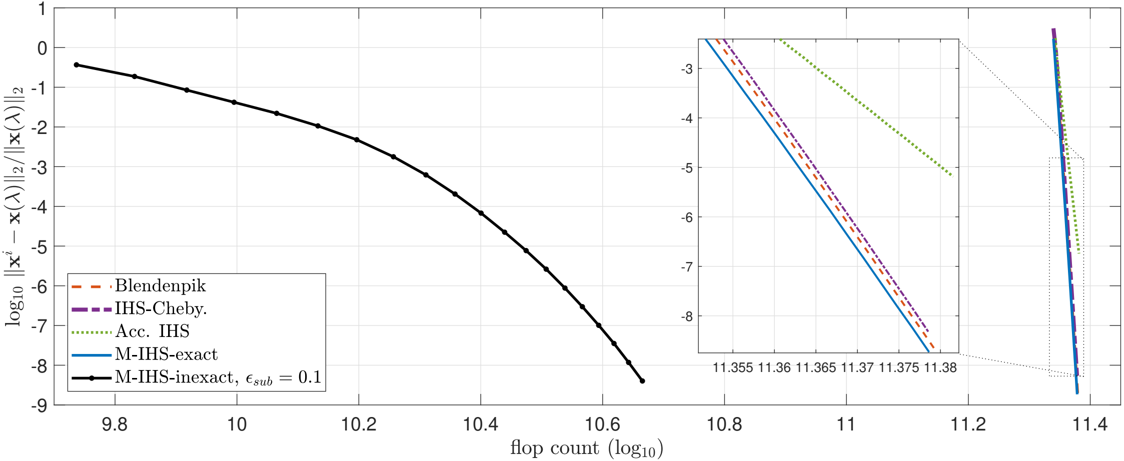
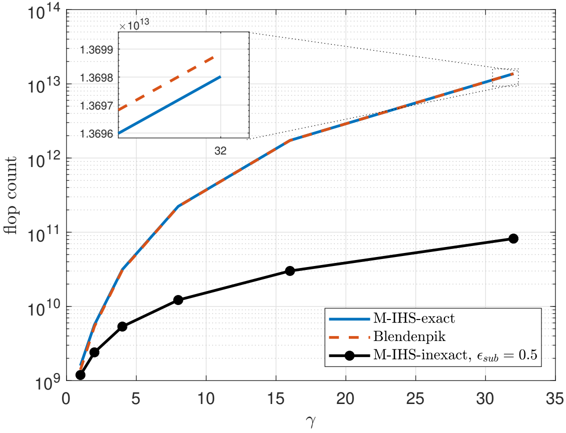


Figure: Performance comparison on a regularized LS problem ($n \gg d$) with dimensions $(n, d, m, \mathsf{sd}_\lambda(\mathbf{A})) = (2^{16}, 4000, 4000, 443)$.

Experiments: Scalability to Larger Size Problems



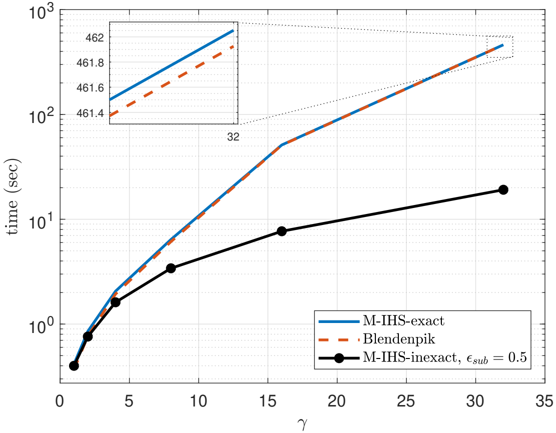


Figure: Complexity of the algorithms in terms of operation count and computation time on a set of $5 \cdot 10^4 \times 500 \cdot \gamma$ dimensional over-determined problems with $m = d$ and $\mathrm{sd}_\lambda(\mathbf{A}) = d/10$.

Proposed Hybrid M-IHS - I

- The `Hybrid M-IHS` uses the following update at the i^{th} iteration:

$$\begin{aligned}\left((\mathbf{SA})^T(\mathbf{SA}) + \lambda_i \mathbf{I}_d\right) \Delta\mathbf{x}^i(\lambda_i) &= \mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x}^i) - \lambda_i \mathbf{x}^i, \\ \mathbf{x}^{i+1} &= \mathbf{x}^i + \alpha_i \Delta\mathbf{x}^i(\lambda_i) + \beta_i(\mathbf{x}^i - \mathbf{x}^{i-1}),\end{aligned}$$

with varying λ_i, α_i and β_i parameters.

Proposed Hybrid M-IHS - I

- The `Hybrid M-IHS` uses the following update at the i^{th} iteration:

$$\left((\mathbf{SA})^T(\mathbf{SA}) + \lambda_i \mathbf{I}_d\right) \Delta\mathbf{x}^i(\lambda_i) = \mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x}^i) - \lambda_i \mathbf{x}^i,$$
$$\mathbf{x}^{i+1} = \mathbf{x}^i + \alpha_i \Delta\mathbf{x}^i(\lambda_i) + \beta_i(\mathbf{x}^i - \mathbf{x}^{i-1}),$$

  with varying λ_i, α_i and β_i parameters.
- After obtaining a proper estimate for the λ_i, the momentum parameters α_i and β_i can be selected as: $(\mathbf{SA} = \mathbf{U}_s \mathbf{\Sigma}_s \mathbf{V}_s^T)$

$$\beta_i = \mathsf{sd}_{\lambda_i}(\mathbf{\Sigma}_s)/m, \qquad \alpha_i = (1 - \beta_i)^2.$$

Proposed Hybrid M-IHS - I

- The `Hybrid M-IHS` uses the following update at the i^{th} iteration:

$$\left((\mathbf{SA})^T(\mathbf{SA}) + \lambda_i \mathbf{I}_d\right)\Delta\mathbf{x}^i(\lambda_i) = \mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x}^i) - \lambda_i \mathbf{x}^i,$$
$$\mathbf{x}^{i+1} = \mathbf{x}^i + \alpha_i \Delta\mathbf{x}^i(\lambda_i) + \beta_i(\mathbf{x}^i - \mathbf{x}^{i-1}),$$

  with varying λ_i, α_i and β_i parameters.

- After obtaining a proper estimate for the λ_i, the momentum parameters α_i and β_i can be selected as: $(\mathbf{SA} = \mathbf{U}_s \boldsymbol{\Sigma}_s \mathbf{V}_s^T)$

$$\beta_i = \mathrm{sd}_{\lambda_i}(\boldsymbol{\Sigma}_s)/m, \qquad \alpha_i = (1 - \beta_i)^2.$$

- To find a proper λ_i for the i^{th} sub-problem, we can utilize the GCV as[13]:

$$G_{full}(\lambda) = \frac{\|\mathbf{b} - \mathbf{A}\mathbf{x}(\lambda)\|_2}{\mathrm{tr}\left(\mathbf{I} - P_{\mathbf{A}}(\lambda)\right)} \longrightarrow \lambda_i = \underset{\lambda \in \mathbb{R}}{\mathrm{argmin}} \frac{\left\|\mathbf{b} - \mathbf{A}\left(\mathbf{x}^i + \Delta\mathbf{x}^i(\lambda)\right)\right\|_2}{\mathrm{tr}\left(\mathbf{I} - P_{\boldsymbol{\Sigma}_s}(\lambda)\right)}$$

! Converges very fast but requires access to $\mathbf{A}$ for each λ

Proposed Hybrid M-IHS - II

- To avoid access to $\mathbf{A}$, we can give up on the noise components outside $\mathcal{R}(\mathbf{A})$:

$$\lambda \mathbf{A}^{\ddagger}\mathbf{x}(\lambda) = \mathbf{U}^T(\mathbf{b} - \mathbf{A}\mathbf{x}(\lambda)), \quad (\mathbf{A}^{\ddagger} = \mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{V}^T)$$

Proposed Hybrid M-IHS - II

- To avoid access to $\mathbf{A}$, we can give up on the noise components outside $\mathcal{R}(\mathbf{A})$:

$$\lambda \mathbf{A}^{\ddagger}\mathbf{x}(\lambda) = \mathbf{U}^T(\mathbf{b} - \mathbf{A}\mathbf{x}(\lambda)), \quad (\mathbf{A}^{\ddagger} = \mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{V}^T)$$

- If $\mathbf{A}^{\ddagger}$ is replaced by $(\mathbf{SA})^{\ddagger}$, then the following biased estimate is obtained:

$$\lambda \left\| (\mathbf{SA})^{\ddagger}\mathbf{x}(\lambda) \right\|_2 = \lambda \left\| \mathbf{\Sigma}_s^{-1}\mathbf{V}_s^T\mathbf{x}(\lambda) \right\|_2 = \left\| (\mathbf{SA})^{\ddagger}\mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x}(\lambda)) \right\|_2, \tag{1}$$

where $\mathbf{SA} = \mathbf{U}_s\mathbf{\Sigma}_s\mathbf{V}_s^T$ and the bias is given by[14]

$$\mathbb{E}_{\mathbf{S}}\left[\left\| (\mathbf{SA})^{\ddagger}\mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x}(\lambda)) \right\|_2 \right] = \theta \left\| \mathbf{U}^T(\mathbf{b} - \mathbf{A}\mathbf{x}(\lambda)) \right\|_2.$$

Proposed Hybrid M-IHS - II

- To avoid access to $\mathbf{A}$, we can give up on the noise components outside $\mathcal{R}(\mathbf{A})$:

$$\lambda \mathbf{A}^{\ddagger}\mathbf{x}(\lambda) = \mathbf{U}^T(\mathbf{b} - \mathbf{A}\mathbf{x}(\lambda)), \quad (\mathbf{A}^{\ddagger} = \mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{V}^T)$$

- If $\mathbf{A}^{\ddagger}$ is replaced by $(\mathbf{SA})^{\ddagger}$, then the following biased estimate is obtained:

$$\lambda \left\| (\mathbf{SA})^{\ddagger}\mathbf{x}(\lambda) \right\|_2 = \lambda \left\| \mathbf{\Sigma}_s^{-1}\mathbf{V}_s^T\mathbf{x}(\lambda) \right\|_2 = \left\| (\mathbf{SA})^{\ddagger}\mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x}(\lambda)) \right\|_2, \qquad (1)$$

where $\mathbf{SA} = \mathbf{U}_s\mathbf{\Sigma}_s\mathbf{V}_s^T$ and the bias is given by[14]

$$\mathbb{E}_{\mathbf{S}}\left[\left\| (\mathbf{SA})^{\ddagger}\mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x}(\lambda)) \right\|_2 \right] = \theta \left\| \mathbf{U}^T(\mathbf{b} - \mathbf{A}\mathbf{x}(\lambda)) \right\|_2.$$

- To get a λ estimate for the i^{th} sub-problem, we substitute $\mathbf{x}^i + \Delta\mathbf{x}^i(\lambda)$ for $\mathbf{x}(\lambda)$ in eq. (1)

$$\lambda_i = \underset{\lambda \in \mathbb{R}}{\text{argmin}} \frac{\lambda \left\| \mathbf{\Sigma}_s^{-1}\mathbf{V}_s^T\left(\mathbf{x}^i + \Delta\mathbf{x}^i(\lambda)\right) \right\|_2}{d - \text{tr}\left(P_{\mathbf{\Sigma}_s}(\lambda)\right)}.$$

Proposed Hybrid Dual M-IHS

- The `Hybrid Dual M-IHS` uses the following update at the i^{th} iteration:

$$\left((\mathbf{A}\mathbf{S}^T)^T(\mathbf{S}\mathbf{A}^T) + \lambda_i \mathbf{I}_n\right) \Delta\boldsymbol{\nu}^i(\lambda_i) = \mathbf{b} - \mathbf{A}\mathbf{A}^T\boldsymbol{\nu}^i - \lambda_i\boldsymbol{\nu}^i$$
$$\boldsymbol{\nu}^{i+1} = \boldsymbol{\nu}^i + \alpha_i \Delta\boldsymbol{\nu}^i(\lambda_i) + \beta_i(\boldsymbol{\nu}^i - \boldsymbol{\nu}^{i-1})$$

with varying λ_i, α_i and β_i parameters.

Proposed Hybrid Dual M-IHS

- The `Hybrid Dual M-IHS` uses the following update at the i^{th} iteration:

$$\left((\mathbf{A}\mathbf{S}^T)^T(\mathbf{S}\mathbf{A}^T) + \lambda_i \mathbf{I}_n\right)\Delta\boldsymbol{\nu}^i(\lambda_i) = \mathbf{b} - \mathbf{A}\mathbf{A}^T\boldsymbol{\nu}^i - \lambda_i\boldsymbol{\nu}^i$$
$$\boldsymbol{\nu}^{i+1} = \boldsymbol{\nu}^i + \alpha_i\Delta\boldsymbol{\nu}^i(\lambda_i) + \beta_i(\boldsymbol{\nu}^i - \boldsymbol{\nu}^{i-1})$$

  with varying λ_i, α_i and β_i parameters.
- Momentum parameters can be chosen in the same fashion as the `Hybrid M-IHS` after estimating a proper λ_i.

Proposed Hybrid Dual M-IHS

- The `Hybrid Dual M-IHS` uses the following update at the i^{th} iteration:

$$\left((\mathbf{A}\mathbf{S}^T)^T(\mathbf{S}\mathbf{A}^T) + \lambda_i \mathbf{I}_n\right)\Delta\boldsymbol{\nu}^i(\lambda_i) = \mathbf{b} - \mathbf{A}\mathbf{A}^T\boldsymbol{\nu}^i - \lambda_i\boldsymbol{\nu}^i$$
$$\boldsymbol{\nu}^{i+1} = \boldsymbol{\nu}^i + \alpha_i\Delta\boldsymbol{\nu}^i(\lambda_i) + \beta_i(\boldsymbol{\nu}^i - \boldsymbol{\nu}^{i-1})$$

  with varying λ_i, α_i and β_i parameters.
- Momentum parameters can be chosen in the same fashion as the `Hybrid M-IHS` after estimating a proper λ_i.
- $\lambda\boldsymbol{\nu}(\lambda) = \mathbf{b} - \mathbf{A}\mathbf{x}(\lambda)$, so the GCV can be written as

$$G_{full}(\lambda) = \frac{\lambda\left\|\boldsymbol{\nu}(\lambda)\right\|_2}{\text{tr}\left(\mathbf{I}_n - P_{\mathbf{A}}(\lambda)\right)}. \quad (2)$$

- To find a proper λ_i estimate, we substitute $\boldsymbol{\nu}^i + \Delta\boldsymbol{\nu}^i(\lambda)$ for $\boldsymbol{\nu}(\lambda)$ in eq. (2)

$$\lambda_i = \underset{\lambda\in\mathbb{R}}{\text{argmin}} \frac{\lambda\left\|\boldsymbol{\nu}^i + \Delta\boldsymbol{\nu}^i(\lambda)\right\|_2}{\text{tr}\left(\mathbf{I}_n - P_{\boldsymbol{\Sigma}_s}(\lambda)\right)}.$$

`Hybrid M-IHS` (for $n \gg d$)

Input: $\mathbf{A} \in \mathbb{R}^{n\times d},\ \mathbf{b},\ m,\ \mathbf{x}^0$

$\quad \mathbf{SA} = \texttt{RP\_fun}(\mathbf{A},\ m)$

$[\mathbf{\Sigma}_s, \mathbf{V}_s] = \texttt{svd}(\mathbf{SA})$

while *until stopping criteria* **do**

$\quad \mathbf{g}^i = \mathbf{V}_s^T\mathbf{A}^T\left(\mathbf{b} - \mathbf{A}\mathbf{x}^i\right)$

$\quad \mathbf{f}^i = \mathbf{\Sigma}_s^{-1}\mathbf{g}^i + \mathbf{\Sigma}_s\mathbf{V}_s^T\mathbf{x}^i$

$\quad \lambda_i = \underset{\lambda}{\text{argmin}}\ \dfrac{\left\|\left(\mathbf{\Sigma}_s^2 + \lambda\mathbf{I}\right)^{-1}\mathbf{f}^i\right\|_2}{\text{tr}\left(\left(\mathbf{\Sigma}_s^2 + \lambda\mathbf{I}\right)^{-1}\right)}$

$\quad \Delta\mathbf{x}^i = \mathbf{V}_s\left(\mathbf{\Sigma}_s^2 + \lambda_i\mathbf{I}\right)^{-1}\left(\mathbf{g}^i - \lambda_i\mathbf{V}_s^T\mathbf{x}^i\right)$

$\quad \widehat{k} = d - \lambda_i\text{tr}\left(\left(\mathbf{\Sigma}_s^2 + \lambda_i\mathbf{I}\right)^{-1}\right)$

$\quad \beta_i = \widehat{k}/m$

$\quad \alpha_i = (1 - \beta_i)^2$

$\quad \mathbf{x}^{i+1} = \mathbf{x}^i + \alpha_i\Delta\mathbf{x}^i + \beta_i(\mathbf{x}^i - \mathbf{x}^{i-1})$

end while

`Hybrid Dual M-IHS` (for $n \ll d$)

Input: $\mathbf{A} \in \mathbb{R}^{n\times d},\ \mathbf{b},\ m$

$\quad \mathbf{SA}^T = \texttt{RP\_fun}(\mathbf{A}^T, m)$

$[\mathbf{\Sigma}_s,\ \mathbf{V}_s] = \texttt{svd}(\mathbf{SA}^T,\ n)$

while *until stopping criteria* **do**

$\quad \widetilde{\mathbf{h}}^i = \mathbf{V}_s^T\left(\mathbf{b} - \mathbf{A}\mathbf{A}^T\boldsymbol{\nu}^i\right)$

$\quad \mathbf{f}^i = \widetilde{\mathbf{h}}^i + \mathbf{\Sigma}_s^2\mathbf{V}_s^T\boldsymbol{\nu}^i$

$\quad \lambda_i = \underset{\lambda}{\text{argmin}}\ \dfrac{\left\|\left(\mathbf{\Sigma}_s^2 + \lambda\mathbf{I}\right)^{-1}\mathbf{f}^i\right\|_2}{\text{tr}\left(\left(\mathbf{\Sigma}_s^2 + \lambda\mathbf{I}\right)^{-1}\right)}$

$\quad \Delta\boldsymbol{\nu}^i = \mathbf{V}_s\left(\mathbf{\Sigma}_s^2 + \lambda_i\mathbf{I}_d\right)^{-1}\left(\widetilde{\mathbf{h}}^i - \lambda_i\mathbf{V}_s^T\boldsymbol{\nu}^i\right)$

$\quad \widehat{k} = d - \lambda_i\text{tr}\left(\left(\mathbf{\Sigma}_s^2 + \lambda_i\mathbf{I}\right)^{-1}\right)$

$\quad \beta_i = \widehat{k}/m$

$\quad \alpha_i = (1 - \beta_i)^2$

$\quad \boldsymbol{\nu}^{i+1} = \boldsymbol{\nu}^i + \alpha_i\Delta\boldsymbol{\nu}^i + \beta_i(\boldsymbol{\nu}^i - \boldsymbol{\nu}^{i-1})$

end while

Proposed Hybrid Primal Dual M-IHS - I

- In main (outer) iterations, it uses `Hybrid Dual M-IHS` update

$$\Delta\boldsymbol{\nu}^i(\lambda_i) = \underset{\boldsymbol{\nu}\in\mathbb{R}^n}{\text{argmin}} \;\; \frac{1}{2}\left\|\mathbf{S}\mathbf{A}^T\boldsymbol{\nu}\right\|_2^2 + \frac{\lambda}{2}\left\|\boldsymbol{\nu}\right\|_2^2 + \langle\nabla g(\boldsymbol{\nu}^i,\lambda_i),\; \boldsymbol{\nu}\rangle \tag{3}$$

$$\boldsymbol{\nu}^{i+1} = \boldsymbol{\nu}^i + \alpha_i\Delta\boldsymbol{\nu}^i(\lambda_i) + \beta_i(\boldsymbol{\nu}^i - \boldsymbol{\nu}^{i-1})$$

- Instead of eq. (3), the dual problem:

$$\mathbf{z}^i(\lambda) = \underset{\mathbf{z}\in\mathbb{R}^{m_1}}{\text{argmin}} \;\; \underbrace{\left\|\mathbf{A}\mathbf{S}^T\mathbf{z} + \nabla g(\boldsymbol{\nu}^i,\lambda)\right\|_2^2 + \lambda\left\|\mathbf{z}\right\|_2^2}_{h(\mathbf{z},\boldsymbol{\nu}^i,\lambda)}, \tag{4}$$

is solved by using following inner iterations:

$$\Delta\mathbf{z}^{i,j}(\lambda_{i,j}) = \underset{\mathbf{z}\in\mathbb{R}^{m_1}}{\text{argmin}} \;\; \left\|\mathbf{W}\mathbf{A}\mathbf{S}^T\mathbf{z}\right\|_2^2 + \lambda_{i,j}\left\|\mathbf{z}\right\|_2^2 + 2\langle\nabla_{\mathbf{z}} h(\mathbf{z}^{i,j},\boldsymbol{\nu}^i,\lambda_{i,j}),\; \mathbf{z}\rangle,$$

$$\mathbf{z}^{i,j+1} = \mathbf{z}^{i,j} + \alpha_j\Delta\mathbf{z}^{i,j}(\lambda_{i,j}) + \beta_j(\mathbf{z}^{i,j} - \mathbf{z}^{i,j-1}),$$

Proposed Hybrid Primal Dual M-IHS - II

- By using the following relation

$$\mathbf{SA}^T\boldsymbol{\nu}^i + \mathbf{SA}^T\Delta\boldsymbol{\nu}^i(\lambda_i) \xleftarrow[\text{inner loop}]{\texttt{Hybrid PD M-IHS}} \mathbf{SA}^T\boldsymbol{\nu}^i + \mathbf{z}^{i,j} + \Delta\mathbf{z}^{i,j}(\lambda_{i,j})$$

Proposed Hybrid Primal Dual M-IHS - II

- By using the following relation

$$\mathbf{SA}^T\boldsymbol{\nu}^i + \mathbf{SA}^T\Delta\boldsymbol{\nu}^i(\lambda_i) \xleftarrow[\text{inner loop}]{\texttt{Hybrid PD M-IHS}} \mathbf{SA}^T\boldsymbol{\nu}^i + \mathbf{z}^{i,j} + \Delta\mathbf{z}^{i,j}(\lambda_{i,j})$$

- We combined risk functions used in `Hybrid M-IHS` and `Hybrid Dual M-IHS`:

$$\lambda_i = \underset{\lambda\in\mathbb{R}}{\text{argmin}} \frac{\lambda\left\|\boldsymbol{\Sigma}_s^{-1}\mathbf{V}_s^T\left(\mathbf{x}^i + \Delta\mathbf{x}^i(\lambda)\right)\right\|_2}{d - \text{tr}\left(P_{\boldsymbol{\Sigma}_s}(\lambda)\right)} \quad \text{and} \quad \lambda_i = \underset{\lambda\in\mathbb{R}}{\text{argmin}} \frac{\lambda\left\|\boldsymbol{\nu}^i + \Delta\boldsymbol{\nu}^i(\lambda)\right\|_2}{\text{tr}\left(\mathbf{I}_n - P_{\boldsymbol{\Sigma}_s}(\lambda)\right)}$$

- Obtained the following risk function:

$$\lambda_{i,j} = \underset{\lambda\in\mathbb{R}}{\text{argmin}} \frac{\lambda\left\|\boldsymbol{\Sigma}_w^{-1}\mathbf{V}_w^T(\mathbf{SA}^T\boldsymbol{\nu}^i + \mathbf{z}^{i,j} + \Delta\mathbf{z}^{i,j}(\lambda))\right\|_2}{m_1 - \text{tr}\left(P_{\boldsymbol{\Sigma}_w}(\lambda)\right)}$$

where $\mathbf{WAS}^T = \mathbf{U}_w\boldsymbol{\Sigma}_w\mathbf{V}_w^T$.

`Hybrid Primal Dual M-IHS` (for $n \leq d$ or $n \geq d$)

Input: $\mathbf{A} \in \mathbb{R}^{n \times d}, \ \mathbf{b}, \ m_1, \ m_2$

$[\mathbf{SA}^T] = \texttt{RP\_fun}(\mathbf{A}^T, \ m_1)$

$[\mathbf{WAS}^T] = \texttt{RP\_fun}(\mathbf{AS}^T, \ m_2)$

$[\mathbf{\Sigma}_w, \mathbf{V}_w] = \texttt{svd}(\mathbf{WAS}^T, \ m_1)$

$\tau = -\infty, \ i = -1, \ \boldsymbol{\nu}^0 = \mathbf{x}^0 = \mathbf{0}, \ \mathbf{z}^{0,0} = \mathbf{0}$

while *until first stopping criteria* **do**

$i = i + 1$

$\mathbf{h}^i = \mathbf{b} - \mathbf{A}\mathbf{x}^i$

$\widetilde{\boldsymbol{\nu}}^i = \mathbf{SA}^T \boldsymbol{\nu}^i$

$\mathbf{z}^{i,0} = \mathbf{z}^{i-1,j}, \ j = -1$

while *until second stopping criteria* **do**

$j = j + 1;$

$\mathbf{g}^{i,j} = \mathbf{V}_w^T \mathbf{SA}^T (\mathbf{h}^i - \mathbf{AS}^T \mathbf{z}^{i,j})$

$\widetilde{\mathbf{z}}^{i,j} = \mathbf{V}_w^T (\mathbf{z}^{i,j} + \widetilde{\boldsymbol{\nu}}^i)$

$\mathbf{f}^{i,j} = \mathbf{\Sigma}_w^{-1} \mathbf{g}^{i,j} + \mathbf{\Sigma}_w \widetilde{\mathbf{z}}^{i,j}$

$\lambda_{i,j} = \underset{\lambda \geq \tau}{\text{argmin}} \ \dfrac{\left\| \left(\mathbf{\Sigma}_w^2 + \lambda \mathbf{I} \right)^{-1} \mathbf{f}^i \right\|_2}{\text{tr}\left((\mathbf{\Sigma}_w^2 + \lambda \mathbf{I})^{-1} \right)}$

$\Delta \mathbf{z}^{i,j} = \mathbf{V}_w \left(\mathbf{\Sigma}_w^2 + \lambda_{i,j} \mathbf{I} \right)^{-1} \left(\mathbf{g}^{i,j} - \lambda_{i,j} \widetilde{\mathbf{z}}^{i,j} \right)$

$\widehat{k} = m_1 - \lambda_{i,j} \text{tr} \left(\left(\mathbf{\Sigma}_w^2 + \lambda_{i,j} \mathbf{I} \right)^{-1} \right)$

$\beta_{1,j} = \widehat{k} / m_2$

$\alpha_{1,j} = (1 - \beta_{1,j})^2$

$\mathbf{z}^{i,j+1} = \mathbf{z}^{i,j} + \alpha_{1,j} \Delta \mathbf{z}^{i,j} + \beta_{1,j} (\mathbf{z}^{i,j} - \mathbf{z}^{i,j-1})$

end while

$\Delta \boldsymbol{\nu}^i = (\mathbf{h}^i - \lambda_{i,j} \boldsymbol{\nu}^i - \mathbf{AS}^T \mathbf{z}^{i,j+1}) / \lambda_{i,j}$

$\beta_{2,i} = \widehat{k} / m_1$

$\alpha_{2,i} = (1 - \beta_{2,i})^2$

$\boldsymbol{\nu}^{i+1} = \boldsymbol{\nu}^i + \alpha_{2,i} \Delta \boldsymbol{\nu}^i + \beta_{2,i} (\boldsymbol{\nu}^i - \boldsymbol{\nu}^{i-1})$

$\mathbf{x}^{i+1} = \mathbf{A}^T \boldsymbol{\nu}^{i+1}$

$\tau = \max(\lambda_{i,j}, \ \tau)$

end while

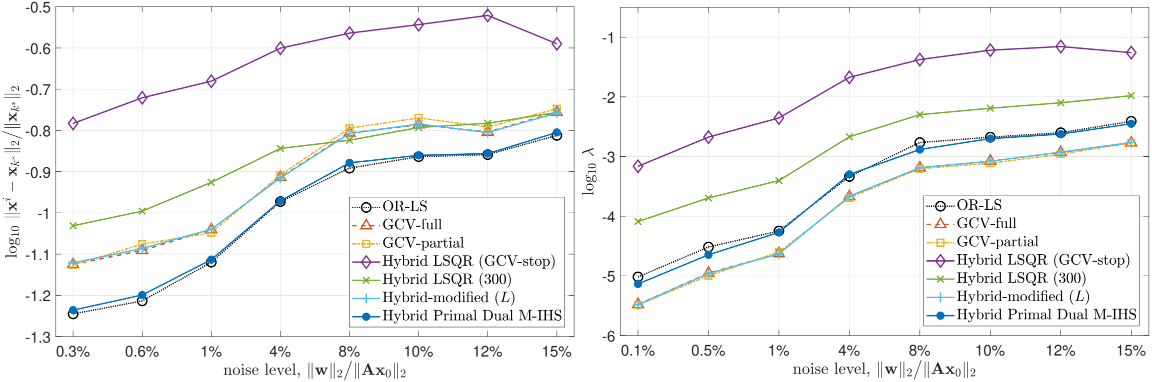


Figure: Error and parameter estimation performances on an image de-blurring problem with Gaussian psf. $(n, d, m_1, m_2) = (10^4, 10^4, 2k^*, 5k^*)$

Table: Effective ranks and the number of iterations that the iterative algorithms need to obtain the results.

Techniques	**0.3%**	**0.6%**	**1%**	**4%**	**8%**	**10%**	**12%**	**15%**
k^*	293	259	245	195	163	164	162	158
Hybrid LSQR	39	27	23	8	4	4	3	38
Hybrid-modified	593	559	545	495	463	464	462	458
Hybrid M-IHS	14	15	14	11	13	12	12	10

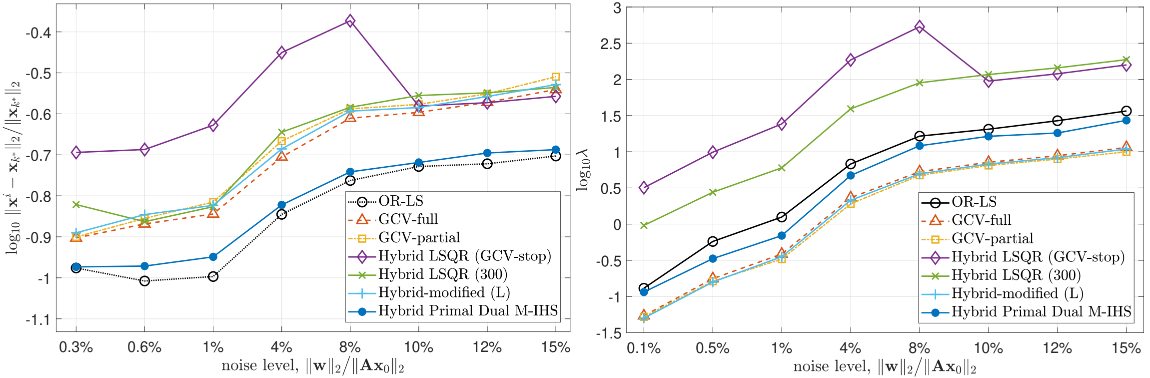


Figure: Error and parameter estimation performances on a seismic travel-time tomography problem with Fresnel wave model. $(n, d, m_1, m_2) = (2 \cdot 10^4, 10^4, 2k^*, 5k^*)$

Table: Effective ranks and the number of iterations that the iterative algorithms need to obtain the results.

Techniques	**0.3%**	**0.6%**	**1%**	**4%**	**8%**	**10%**	**12%**	**15%**
k^*	1324	1006	759	417	261	224	188	176
Hybrid LSQR	43	33	27	11	7	69	63	57
Hybrid-modified	2260	1879	1676	1386	1266	1256	1227	1994
Hybrid M-IHS	10	10	9	9	10	10	12	10

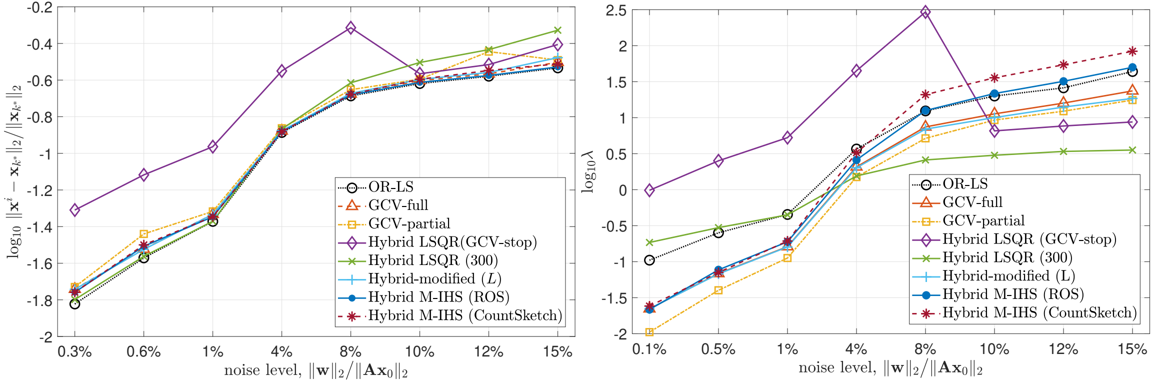


Figure: Error and parameter estimation performances on X-ray tomography problem with parallel beam geometry. $(n, d, m) = (12780, 2500, 5000)$

Table: Effective ranks and the number of iterations that the iterative algorithms need to obtain the results.

Techniques	**0.3%**	**0.6%**	**1%**	**4%**	**8%**	**10%**	**12%**	**15%**
k^*	2495	2489	2480	2460	2356	2306	2260	2106
Hybrid LSQR	38	29	22	9	6	133	124	126
Hybrid-modified	2498	2492	2483	2463	2359	2309	2263	2109
Hybrid M-IHS	18	17	16	13	12	9	10	9

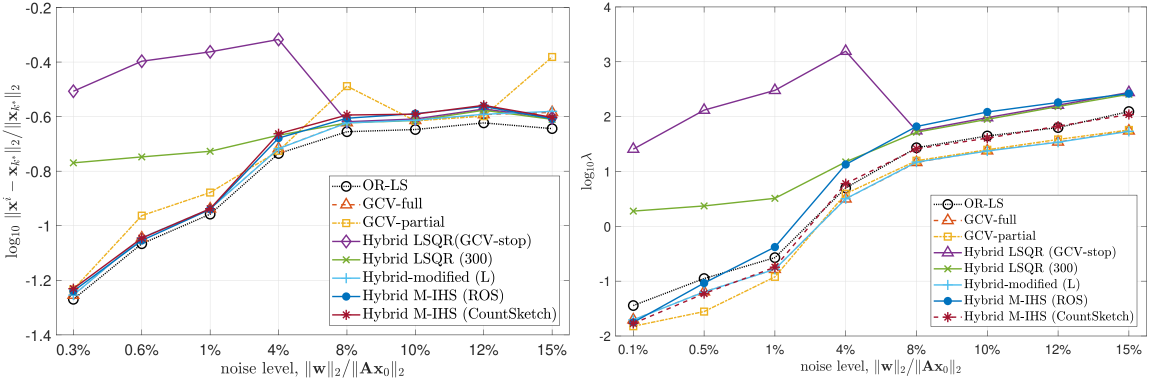


Figure: Error and parameter estimation performances on seismic travel-time tomography problem with straight-line wave model. $(n, d, m) = (64000, 1600, 3200)$

Table: Effective ranks and the number of iterations that the iterative algorithms need to obtain the results.

Techniques	0.3%	0.6%	1%	4%	8%	10%	12%	15%
k^*	1590	1581	1565	1473	1226	1221	1214	1180
Hybrid LSQR	48	24	22	6	284	280	276	256
Hybrid-modified	1600	1600	1600	1600	1600	1600	1593	1534
Hybrid M-IHS	18	18	17	13	10	8	9	8

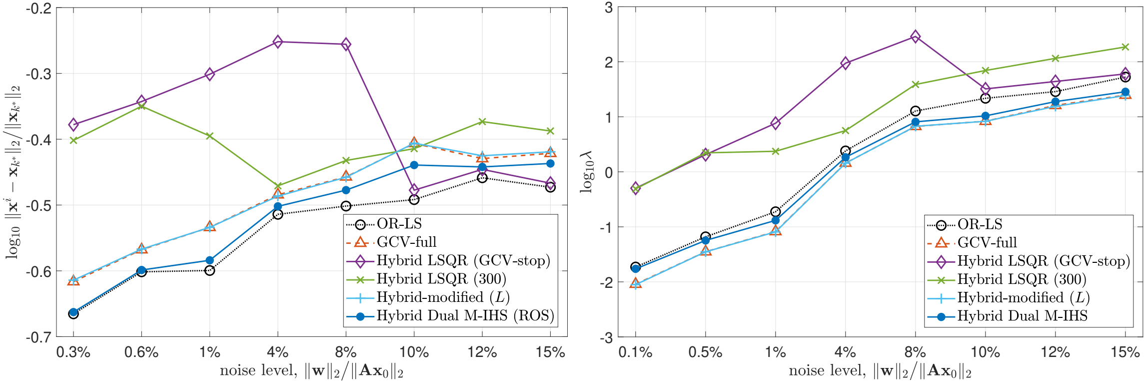


Figure: Error and parameter estimation performances on a randomly generated data. $(n, d, m) = (1500, 10^4, 3000)$

Table: Effective ranks and the number of iterations that the iterative algorithms need to obtain the results.

Techniques	0.3%	0.6%	1%	4%	8%	10%	12%	15%
k^*	879	832	791	679	603	579	563	527
Hybrid LSQR	177	109	58	17	10	98	82	70
Hybrid-modified	1179	1132	1091	979	903	879	863	827
Hybrid M-IHS	7	7	7	8	10	10	10	10

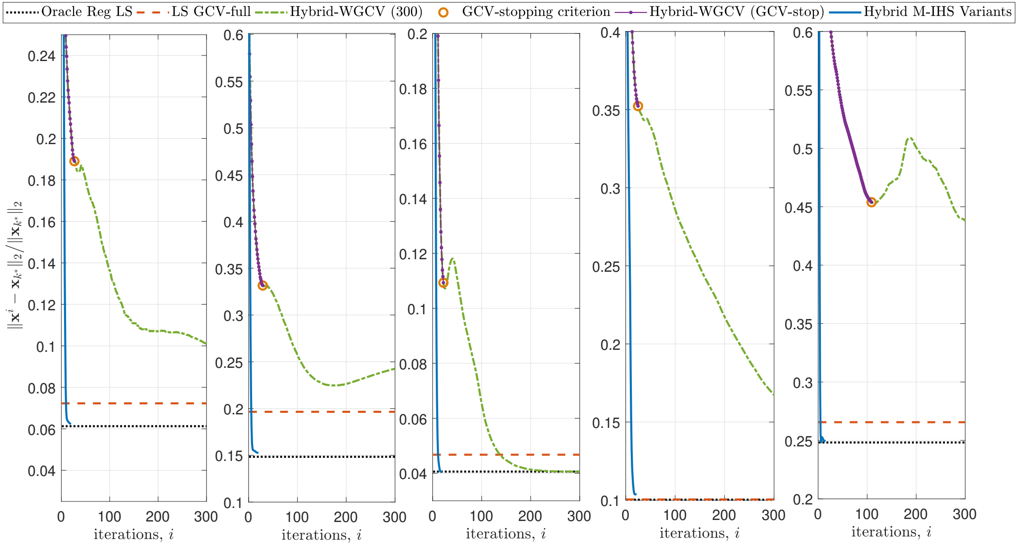
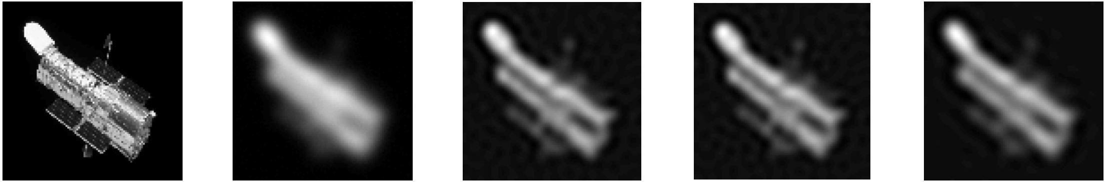
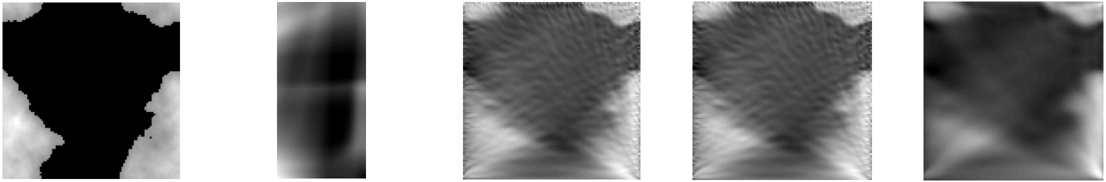


Figure: Convergence behaviour of the hybrid methods in each previous example at a noise level of 1%.



(a) $\mathbf{x}_0$ (b) $\mathbf{b}$ (c) $\mathbf{x}_{Oracle}$ (d) $\mathbf{x}_{M\text{-}IHS}$ (e) $\mathbf{x}_{Hybrid\text{-}LSQR}$

Figure: Example 1 ($n = d$): image deblurring problem with Gaussian psf

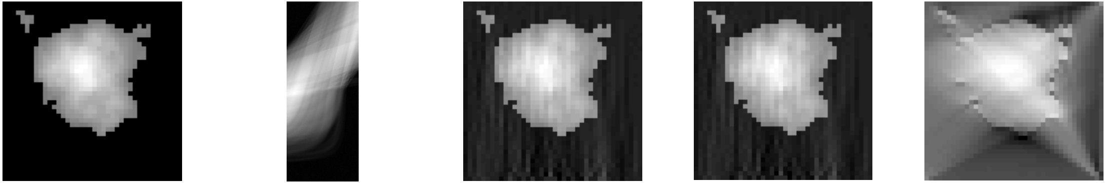


(a) $\mathbf{x}_0$ (b) $\mathbf{b}$ (c) $\mathbf{x}_{Oracle}$ (d) $\mathbf{x}_{M\text{-}IHS}$ (e) $\mathbf{x}_{Hybrid\text{-}LSQR}$

Figure: Example 2 ($n \geq d$): seismic travel-time tomography problem with Fresnel wave model

(a) $\mathbf{x}_0$ (b) $\mathbf{b}$ (c) $\mathbf{x}_{\texttt{Oracle}}$ (d) $\mathbf{x}_{\textit{M-IHS}}$ (e) $\mathbf{x}_{\textit{Hybrid-LSQR}}$

Figure: Example 3 ($n \gg d$): X-ray tomography problem with parallel beam geometry



(a) $\mathbf{x}_0$ (b) $\mathbf{b}$ (c) $\mathbf{x}_{\texttt{Oracle}}$ (d) $\mathbf{x}_{\textit{M-IHS}}$ (e) $\mathbf{x}_{\textit{Hybrid-LSQR}}$

Figure: Example 4 ($n \gg d$): seismic travel-time tomography problem with straight-line wave model

Conclusions and Future Work

- ✓ We introduced a group of solver for large scale linear least squares problems.
- ✓ The proposed algorithms are effective as long as the statistical dimension is sufficiently smaller than at least one size of the coefficient matrix.
- ✓ They have various desirable properties for modern computing devices that are prevalent in large scale applications.
- ✓ In regularized problems, if the regularization parameters are unknown, the `Hybrid M-IHS` algorithms have capability of finding better parameters than direct methods in far fewer number of iterations than the conventional hybrid methods.

Conclusions and Future Work

- ✓ We introduced a group of solver for large scale linear least squares problems.
- ✓ The proposed algorithms are effective as long as the statistical dimension is sufficiently smaller than at least one size of the coefficient matrix.
- ✓ They have various desirable properties for modern computing devices that are prevalent in large scale applications.
- ✓ In regularized problems, if the regularization parameters are unknown, the `Hybrid M-IHS` algorithms have capability of finding better parameters than direct methods in far fewer number of iterations than the conventional hybrid methods.
- The effect of the inexact sub-solvers on the convergence rate of the `M-IHS` algorithms can be studied as a future direction.
- Classical sketching methods can be investigated to estimate proper regularization parameter and to construct regularized solution.

Bibliography I

[1] Å. Björck, *Numerical methods for least squares problems*. Philadelphia, PA, USA: SIAM, 1996, ISBN: 978-0-89871-360-2.

[2] A. Greenbaum, *Iterative methods for solving linear systems*. Siam, 1997, vol. 17.

[3] G. H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, no. 2, pp. 215–223, 1979.

[4] C. C. Paige and M. A. Saunders, "Lsqr: An algorithm for sparse linear equations and sparse least squares," *ACM Trans. Math. Softw.*, vol. 8, no. 1, pp. 43–71, 1982.

[5] M. E. Kilmer and D. P. O'Leary, "Choosing regularization parameters in iterative methods for ill-posed problems," *SIAM J. Matrix Anal. Appl.*, vol. 22, no. 4, pp. 1204–1221, 2001.

[6] J. Yang, X. Meng, and M. W. Mahoney, "Implementing randomized matrix algorithms in parallel and distributed environments," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 58–92, 2015.

[7] D. P. Woodruff *et al.*, "Sketching as a tool for numerical linear algebra," *Found. Trends Theor. Comput. Sci.*, vol. 10, no. 1–2, pp. 1–157, 2014.

[8] H. Avron, P. Maymounkov, and S. Toledo, "Blendenpik: Supercharging lapack's least-squares solver," *SIAM J. Sci. Comput.*, vol. 32, no. 3, pp. 1217–1236, 2010.

[9] X. Meng, M. A. Saunders, and M. W. Mahoney, "Lsrn: A parallel iterative solver for strongly over-or underdetermined systems," *SIAM J. Sci. Comput.*, vol. 36, no. 2, pp. C95–C118, 2014.

[10] M. Pilanci and M. J. Wainwright, "Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1842–1879, 2016.

[11] J. Wang, J. D. Lee, M. Mahdavi, M. Kolar, N. Srebro, *et al.*, "Sketching meets random projection in the dual: A provable recovery algorithm for big and high-dimensional data," *Electron. J. Stat.*, vol. 11, no. 2, pp. 4896–4944, 2017.

[12] I. K. Ozaslan, M. Pilanci, and O. Arikan, "Fast and robust solution techniques for large scale linear system of equations," in *2019 27th Signal Processing and Communications Applications Conference (SIU)*, IEEE, 2019, pp. 1–4.

Bibliography II

[13] J. Lacotte and M. Pilanci, "Faster least squares optimization," *arXiv preprint arXiv:1911.02675*, 2019.

[14] M. Pilanci and M. J. Wainwright, "Randomized sketches of convex programs with sharp guarantees," *IEEE Trans. Inform. Theory*, vol. 61, no. 9, pp. 5096–5115, 2015.

[15] P. C. Hansen, "Regularization tools: A matlab package for analysis and solution of discrete ill-posed problems," *Numer. Algorithms*, vol. 6, no. 1, pp. 1–35, 1994.

[16] S. Gazzola, P. C. Hansen, and J. G. Nagy, "Ir tools: A matlab package of iterative regularization methods and large-scale test problems," *Numer. Algorithms*, vol. 81, no. 3, pp. 773–811, 2019.

RP-based Methods: Classical Sketching

- Based on observing $(\mathbf{SA}, \mathbf{Sb})$ pair instead of $(\mathbf{A}, \mathbf{b})$

$$\widehat{\mathbf{x}}(\lambda) = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \;\; \frac{1}{2}\|\mathbf{SAx} - \mathbf{Sb}\|_2^2 + \frac{\lambda}{2}\|\mathbf{x}\|_2^2$$

- Seeks ζ-optimal cost approximation[15]:

$$f(\widehat{\mathbf{x}}(\lambda), \lambda) \leq (1+\zeta) f(\mathbf{x}(\lambda), \lambda)$$

- $O(nd\log(m) + md^2)$ vs $O(nd^2)$
- Sub-optimal for obtaining a η-optimal solution approximation[16]:

$$\|\widehat{\mathbf{x}}(\lambda) - \mathbf{x}(\lambda)\|_{\mathbf{W}} \leq \eta \|\mathbf{x}(\lambda)\|_{\mathbf{W}},$$

for example, if $\mathbf{w} \sim \mathcal{N}(0, \; \sigma_{\mathbf{w}}^2 \mathbf{I}_n)$, then:

$$\mathbb{E}_{\mathbf{w}}\left[\|\mathbf{x}_{\text{LS}} - \mathbf{x}_0\|_{\mathbf{A}}\right] \preceq \frac{\sigma_{\mathbf{w}}^2 d}{n} \qquad \text{whereas} \qquad \mathbb{E}_{\mathbf{S},\mathbf{w}}\left[\|\widehat{\mathbf{x}}(\lambda) - \mathbf{x}_0\|_{\mathbf{A}}\right] \succeq \frac{\sigma_{\mathbf{w}}^2 d}{\min(m,n)}$$

Estimation of the Statistical Dimension

The statistical dimension of $\mathbf{A} \in \mathbb{R}^{n \times d}$ can be estimated as

$$\text{sd}_\lambda(\mathbf{A}) = \text{tr}\left(\mathbf{A}\left(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}\right)^{-1}\mathbf{A}^T\right) = \text{tr}\left(\mathbf{I} - \lambda\left(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}\right)^{-1}\right)$$

$$= d - \lambda\mathbb{E}_{\mathbf{v}}\left[\text{tr}\left(\mathbf{v}^T\left(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}\right)^{-1}\mathbf{v}\right)\right] \approx d - \frac{\lambda}{T}\sum_{i=1}^{T}\langle\mathbf{v}^i,\ \mathbf{z}^i\rangle$$

where $\left(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}\right)\mathbf{z} = \mathbf{v}^i$ and $\mathbf{v}^i$'s are Rademacher r.v.'s with covariance $\mathbb{E}\left[\mathbf{v}\mathbf{v}^T\right] = \mathbf{I}_d$.

Inexact Hutchinson Trace Estimator

1: **Input:** $\mathbf{SA} \in \mathbb{R}^{m \times d},\ \lambda,\ T,\ \epsilon_{tr}$
2: $\mathbf{v}^\ell = \{-1, +1\}^d, \quad \ell = 1, \ldots, T$
3: $\tau = 0$
4: **for** i $= 1{:}T$ **do**
5: $\quad \mathbf{z}^i = \texttt{AAb\_Solver}(\mathbf{SA}, \mathbf{v}^i, \lambda, \epsilon_{tr})$
6: $\quad \tau = \tau + \lambda\langle\mathbf{v}^i,\ \mathbf{z}^i\rangle$
7: **end for**
8: **Output**: $\widehat{\text{sd}_\lambda} = d - \tau/T$

Numerical Experiments and Comparisons

Data is generated syntactically as following:

1. The entries of $\mathbf{A}$ were drawn from the distribution $\mathcal{N}(1_d, \mathbf{\Gamma})$ where $\Gamma_{ij} = 5 \cdot 0.9^{|i-j|}$.
2. Singular values were replaced with *philips* profile provided in RegTool[17].
3. Condition number $\kappa(\mathbf{A})$ was set to 10^8.
4. For un-regularized problems, the entries of $\mathbf{x}_0$ were sampled from Uni[-1,1].
5. For regularized problems, the inputs provided by RegTool were used.
6. Additive i.i.d. Gaussian noise at level of $\|\mathbf{w}\|_2 \,/\, \|\mathbf{A}\mathbf{x}\|_2 = 1\%$ was used for regularized problems.

Results were averaged over 32 MC simulations.

Experiments: Under-determined Regularized Problems

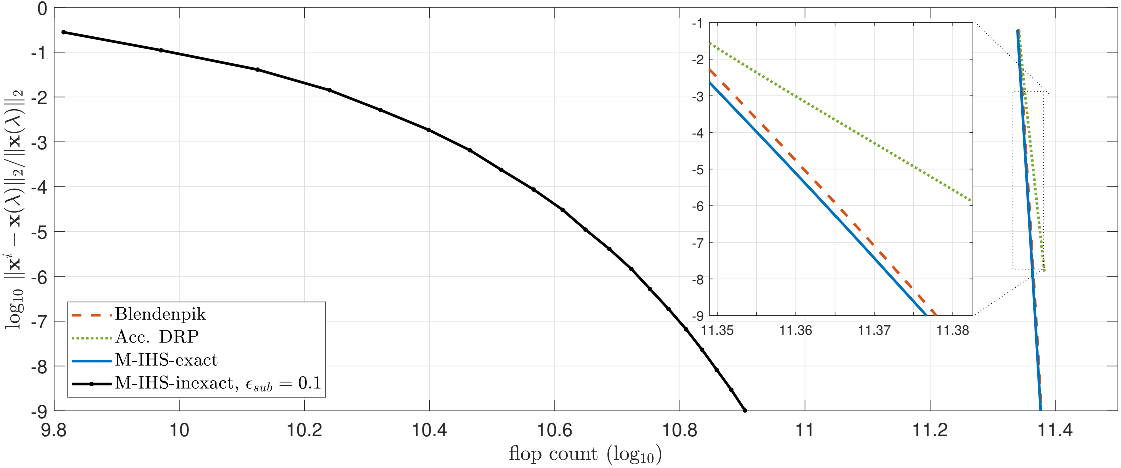


Figure: Performance comparison on a regularized LS problem ($n \ll d$) with dimensions $(n, d, m, \mathsf{sd}_\lambda(\mathbf{A})) = (4000, 2^{16}, 4000, 462)$.

Experiments: Scalability to Larger Size Problems - II

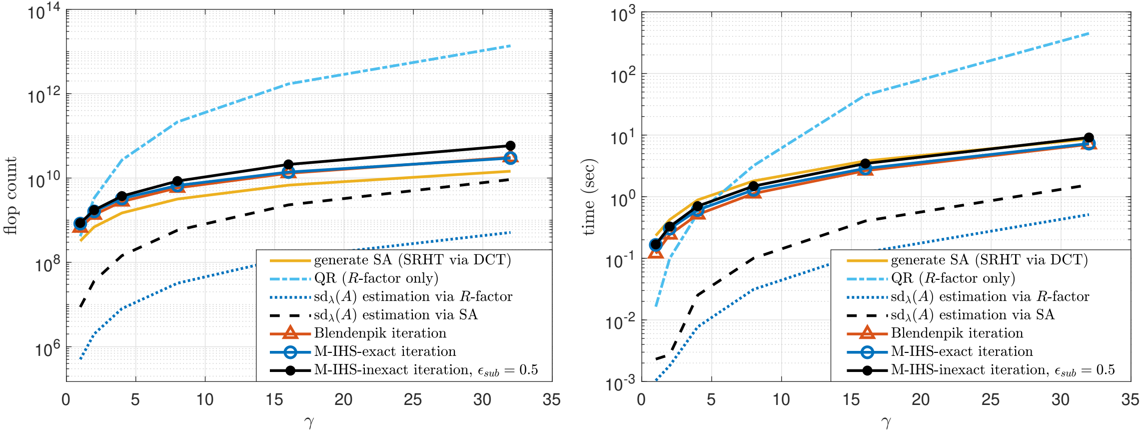


Figure: Complexity of the each stage in terms of operation count and computation time on a set of $5 \cdot 10^4 \times 500 \cdot \gamma$ dimensional over-determined problems with $m = d$ and $\mathrm{sd}_\lambda(\mathbf{A}) = d/10$.

Experiments: Effect of $\text{sd}_\lambda(\mathbf{A})$ on Performance of the Inexact Schemes

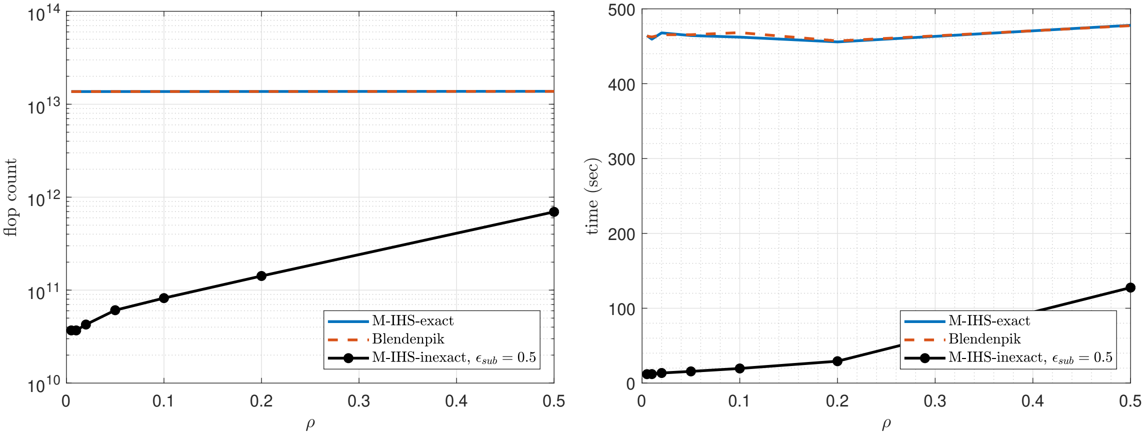


Figure: Complexity of the algorithms in terms of operation count and computation time on a $5 \cdot 10^4 \times 4 \cdot 10^3$ dimensional problem for different $\rho = \text{sd}_\lambda(A)/d$ ratios.

Experiments: Effect of $\text{sd}_\lambda(\mathbf{A})$ on Performance of the Inexact Schemes - II

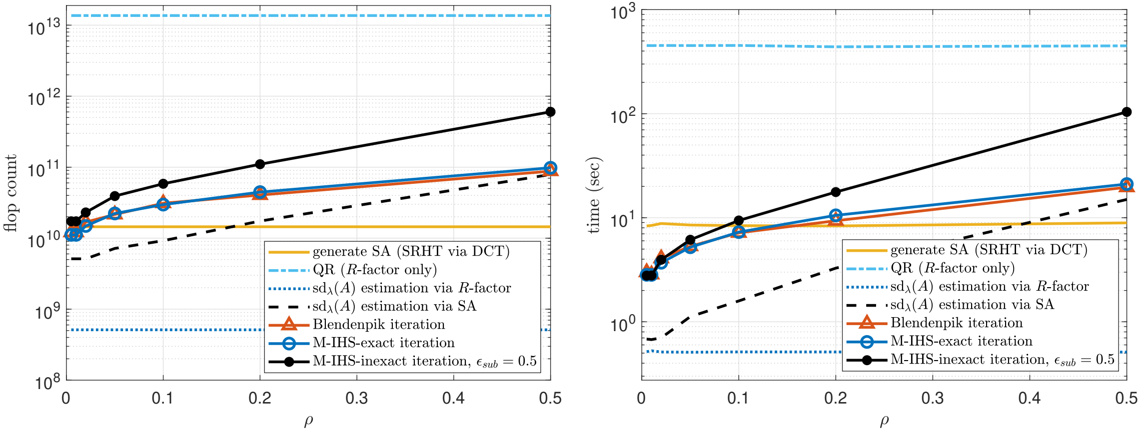


Figure: Complexity of each stage in terms of operation count and computation time on a $5 \cdot 10^4 \times 4 \cdot 10^3$ dimensional problem for different $\rho = \text{sd}_\lambda(\mathbf{A})/d$ ratios.

Experiments: Un-regularized LS Problem

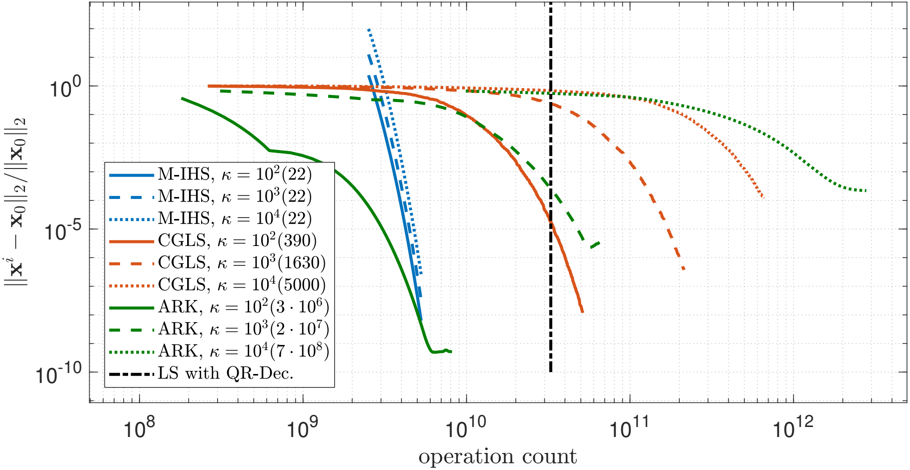


Figure: Performance comparison of the M-IHS, ARK and CGLS on a set of un-regularized LS problem with size $2^{16} \times 500$ and different condition numbers.

Numerical Experiments and Comparisons for Hybrid Methods

- We used IR tools[18] to generate realistic examples:
  1. Image de-blurring problem with Gaussian psf: $10^4 \times 10^4$
  2. Seismic travel-time tomography problem with Fresnel wave model: $2 \cdot 10^4 \times 10^4$
  3. X-ray tomography problem with parallel beam geometry: 12780×2500
  4. Seismic travel-time tomography with Straight-Line wave model: 6400×1600
  5. Randomly generated $\mathbf{A}$ and $\mathbf{x}_0$ as earlier: $1500 \times 4 \cdot 10^4$
- We calculated relative error with respect to the effective true input $\mathbf{x}_{k^*} = \mathbf{V}_{k^*}\mathbf{V}_{k^*}^T\mathbf{x}_0$
- Additive Gaussian noise with 8 different levels was used. Noise level is determined by the ratio $\frac{\|\mathbf{w}\|_2}{\|\mathbf{A}\mathbf{x}_0\|_2}$.
- Results were averaged over 20 noise realizations.

Numerical Experiments and Comparisons for Hybrid Methods

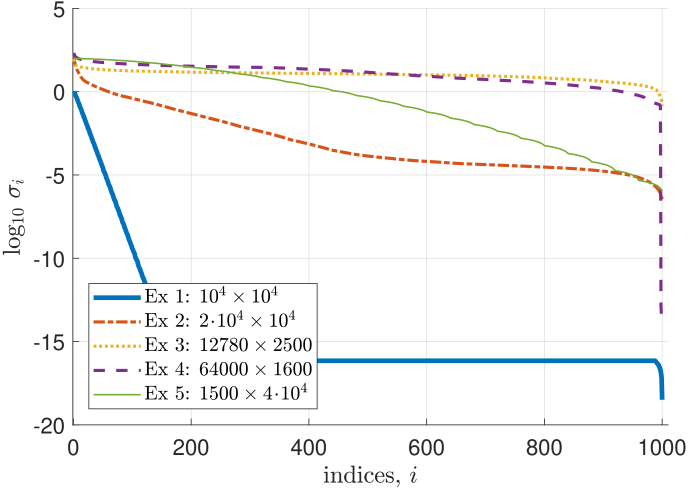


Figure: The size and the singular value profiles of the coefficient matrices used in the numerical experiments.

Numerical Experiments and Comparisons

Table: PSNR (in dB) values of the reconstructed images measured with respect to the effective true input $\mathbf{x}_{k^*}$.

ex. no	**ex. 1**			**ex. 2**			**ex. 3**			**ex. 4**		
$\|\mathbf{w}\|/\|\mathbf{A}\mathbf{x}_0\|$	0.3%	1%	10%	0.3%	1%	10%	0.3%	1%	10%	0.3%	1%	10%
OR-LS	36.00	35.71	31.48	28.46	23.65	22.89	49.20	39.79	24.93	35.92	28.99	22.56
Hybrid M-IHS	35.95	35.6	29.27	28.44	23.60	22.89	47.70	39.49	24.82	36.02	28.92	22.56
Hybrid LSQR	30.57	29.80	24.93	22.58	16.09	19.37	38.40	31.43	24.02	15.95	15.93	22.04

Linear Least Squares Problems

- Linear Systems of equations:

$$\mathbf{A}\mathbf{x}_0 + \mathbf{w} = \mathbf{b}, \qquad \mathbf{A} \in \mathbb{R}^{n \times d}.$$

- Aim is to recover $\mathbf{x}_0$ by observing $\mathbf{A}$ and $\mathbf{b}$:

Linear Least Squares Problems

- Linear Systems of equations:

$$\mathbf{A}\mathbf{x}_0 + \mathbf{w} = \mathbf{b}, \qquad \mathbf{A} \in \mathbb{R}^{n \times d}.$$

- Aim is to recover $\mathbf{x}_0$ by observing $\mathbf{A}$ and $\mathbf{b}$:

$$\mathbf{x}_{\mathsf{LS}} = \underset{\mathbf{x} \in \mathbb{R}^d}{\operatorname{argmin}} \; \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$$

Linear Least Squares Problems

- Linear Systems of equations:

$$\mathbf{A}\mathbf{x}_0 + \mathbf{w} = \mathbf{b}, \qquad \mathbf{A} \in \mathbb{R}^{n \times d}.$$

- Aim is to recover $\mathbf{x}_0$ by observing $\mathbf{A}$ and $\mathbf{b}$: $(\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T)$

$$\mathbf{x}_{\text{LS}} = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = \sum_{i=1}^{d} \frac{\mathbf{u}_i^T\mathbf{b}}{\sigma_i}\mathbf{v}_i$$

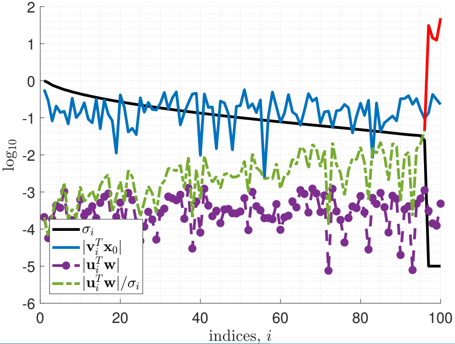
Linear Least Squares Problems

- Linear Systems of equations:

$$\mathbf{A}\mathbf{x}_0 + \mathbf{w} = \mathbf{b}, \qquad \mathbf{A} \in \mathbb{R}^{n\times d}.$$

- Aim is to recover $\mathbf{x}_0$ by observing $\mathbf{A}$ and $\mathbf{b}$: $(\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T)$

$$\mathbf{x}_{\mathsf{LS}} = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \; \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = \sum_{i=1}^{d} \frac{\mathbf{u}_i^T\mathbf{b}}{\sigma_i}\mathbf{v}_i = \sum_{i=1}^{k^*}\left(\mathbf{v}_i^T\mathbf{x}_0 + \frac{\mathbf{u}_i^T\mathbf{w}}{\sigma_i}\right)\mathbf{v}_i + \sum_{i=k^*+1}^{d}\left(\mathbf{v}_i^T\mathbf{x}_0 + \frac{\mathbf{u}_i^T\mathbf{w}}{\sigma_i}\right)\mathbf{v}_i$$



Linear Least Squares Problems

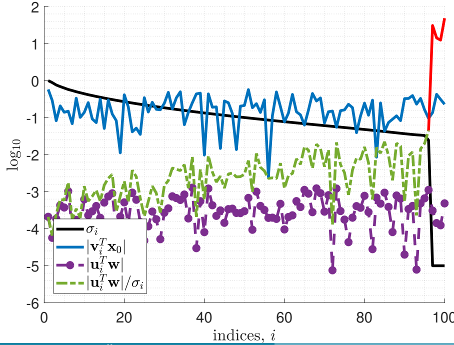
- Linear Systems of equations:

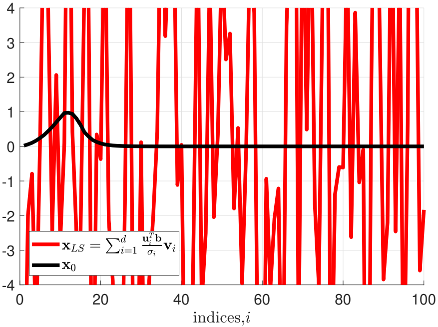
$$\mathbf{A}\mathbf{x}_0 + \mathbf{w} = \mathbf{b}, \qquad \mathbf{A} \in \mathbb{R}^{n\times d}.$$

- Aim is to recover $\mathbf{x}_0$ by observing $\mathbf{A}$ and $\mathbf{b}$: $(\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T)$

Noise Enhancement

$$\mathbf{x}_{\mathsf{LS}} = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = \sum_{i=1}^{d} \frac{\mathbf{u}_i^T\mathbf{b}}{\sigma_i}\mathbf{v}_i = \sum_{i=1}^{k^*}\left(\mathbf{v}_i^T\mathbf{x}_0 + \frac{\mathbf{u}_i^T\mathbf{w}}{\sigma_i}\right)\mathbf{v}_i + \underbrace{\sum_{i=k^*+1}^{d}\left(\mathbf{v}_i^T\mathbf{x}_0 + \frac{\mathbf{u}_i^T\mathbf{w}}{\sigma_i}\right)\mathbf{v}_i}_{\text{Noise Enhancement}}$$





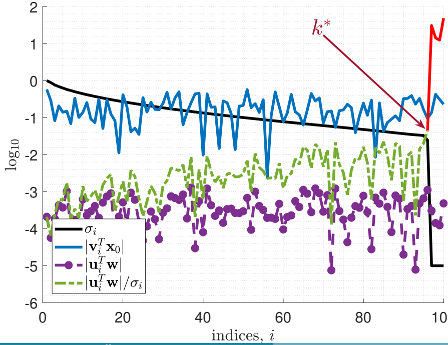
Linear Least Squares Problems

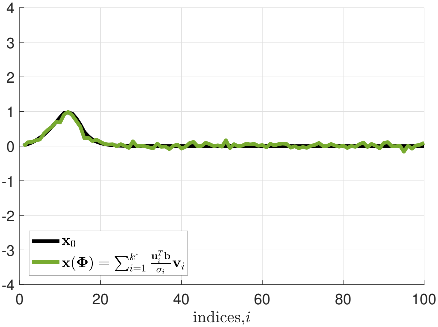
- Linear Systems of equations:

$$\mathbf{A}\mathbf{x}_0 + \mathbf{w} = \mathbf{b}, \qquad \mathbf{A} \in \mathbb{R}^{n\times d}.$$

- Aim is to recover $\mathbf{x}_0$ by observing $\mathbf{A}$ and $\mathbf{b}$: $(\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T)$

$$\mathbf{x}_{\mathsf{LS}} = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \; \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = \sum_{i=1}^{d} {\color{red}\phi_i} \frac{\mathbf{u}_i^T\mathbf{b}}{\sigma_i}\mathbf{v}_i = \sum_{i=1}^{k^*} \left(\mathbf{v}_i^T\mathbf{x}_0 + \frac{\mathbf{u}_i^T\mathbf{w}}{\sigma_i} \right) \mathbf{v}_i$$





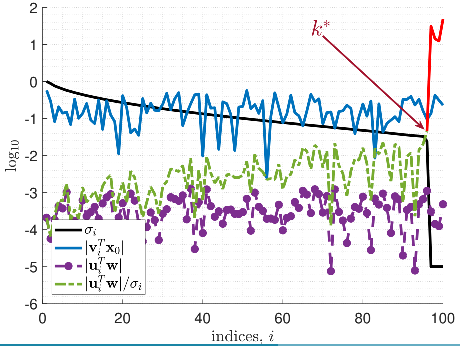
Linear Least Squares Problems

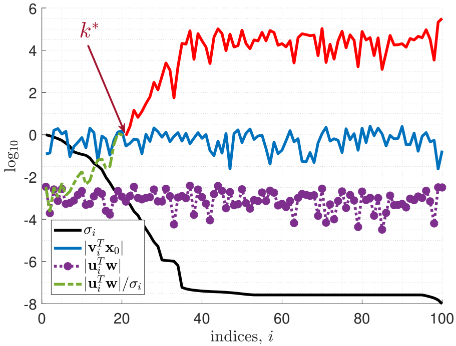
- Linear Systems of equations:

$$\mathbf{A}\mathbf{x}_0 + \mathbf{w} = \mathbf{b}, \qquad \mathbf{A} \in \mathbb{R}^{n \times d}.$$

- Aim is to recover $\mathbf{x}_0$ by observing $\mathbf{A}$ and $\mathbf{b}$: $(\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T)$

$$\mathbf{x}_{\text{LS}} = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}} \; \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = \sum_{i=1}^{d} {\color{red}\phi_i} \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i = \sum_{i=1}^{k^*} \left(\mathbf{v}_i^T \mathbf{x}_0 + \frac{\mathbf{u}_i^T \mathbf{w}}{\sigma_i} \right) \mathbf{v}_i$$





Regularized LS Problems

$$\mathbf{x}(\boldsymbol{\Phi}) = \mathbf{V}\boldsymbol{\Phi}\boldsymbol{\Sigma}^{-1}\mathbf{U}^T\mathbf{b} = \sum_{i=1}^{d} \phi_i \frac{\mathbf{u}_i^T\mathbf{b}}{\sigma_i}\mathbf{v}_i, \quad \text{assume } |\mathbf{v}_i^T\mathbf{x}_0| \leq \frac{|\mathbf{u}_i^T\mathbf{w}|}{\sigma_i} \text{ for } i \in [k^*]$$

- Hard thresholding: $\phi_i = \begin{cases} 1, & 0 < i < k^* \\ 0, & \text{otherwise} \end{cases}$
  - $\mathbf{x}(k^*) = \mathbf{U}_{k^*}\boldsymbol{\Sigma}_{k^*}^{-1}\mathbf{V}_{k^*}^T = \sum_{i=1}^{k^*} \frac{\mathbf{u}_i^T\mathbf{b}}{\sigma_i}\mathbf{v}_i$
- Soft thresholding: $\phi_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \approx \begin{cases} 1, & \sigma_i \gg \lambda \\ 0, & \sigma_i \ll \lambda \end{cases}$
  - $\text{sd}_\lambda(\mathbf{A}) = \sum_{i=1}^{d} \phi_i = \sum_{i=1}^{d} \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \approx k^*$
  - $\mathbf{x}(\lambda) = \mathbf{V}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^2 + \lambda\mathbf{I}_d)^{-1}\mathbf{U}^T\mathbf{b} = (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}_d)^{-1}\mathbf{A}^T\mathbf{b}$
  - $\mathbf{x}(\lambda) = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} \underbrace{\frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \frac{\lambda}{2}\|\mathbf{x}\|_2^2}_{f(\mathbf{x},\lambda)}$