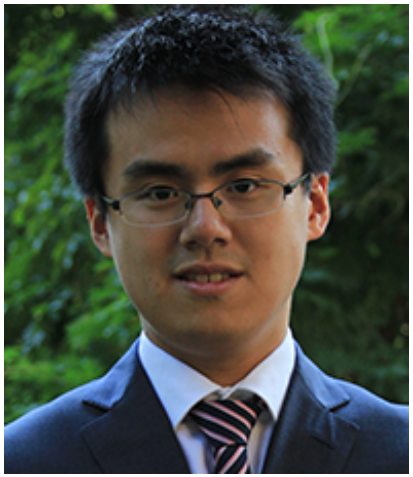
Recursive Importance Sketching for Rank Constrained Least Squares: Algorithms and High-order Convergence

Wen Huang

Joint meeting of Wuhan University and Xiamen University

Yuetian Luo
UW-Madison

Xudong Li
Fudan University



Anru Zhang
UW-Madison

Thank Yuetian Luo for creating these slides.

Problem of Interest

$$\min_{\mathbf{X}\in\mathbb{R}^{p_1\times p_2}} f(\mathbf{X}) := \frac{1}{2}\|\mathbf{y}-\mathcal{A}(\mathbf{X})\|_2^2, \quad \text{subject to} \quad \operatorname{rank}(\mathbf{X}) = r,$$

where $\mathbf{y}\in\mathbb{R}^n, \mathcal{A}(\mathbf{X}) = [\langle \mathbf{A}_1, \mathbf{X}\rangle, \ldots, \langle \mathbf{A}_n, \mathbf{X}\rangle]^\top$.

Problem of Interest

$$\min_{\mathbf{X}\in\mathbb{R}^{p_1\times p_2}} f(\mathbf{X}) := \frac{1}{2}\|\mathbf{y}-\mathcal{A}(\mathbf{X})\|_2^2, \quad \text{subject to} \quad \operatorname{rank}(\mathbf{X}) = r,$$

where $\mathbf{y}\in\mathbb{R}^n, \mathcal{A}(\mathbf{X}) = [\langle \mathbf{A}_1, \mathbf{X}\rangle, \ldots, \langle \mathbf{A}_n, \mathbf{X}\rangle]^\top$.

Motivation: Low rank matrix recovery

- Observe $\mathbf{y}, \mathcal{A}$ from $\mathbf{y} = \mathcal{A}(\mathbf{X}^*) + \epsilon$. Goal: recover $\mathbf{X}^*$ from $\mathbf{y}, \mathcal{A}$

Specific problems:

- Matrix regression: $\mathbf{A}_i \overset{i.i.d.}{\sim} N(0,1)$
  [Candès and Plan, 2011, Recht et al., 2010]
- Matrix Completion: $\mathbf{A}_i$ has one entry to be 1, others are 0
  [Candès and Tao, 2010]
- Phase retrieval: $\mathbf{A}_i = \mathbf{a}_i\mathbf{a}_i^\top$ [Shechtman et al., 2015]
- Rank-one sensing: $\mathbf{A}_i = \mathbf{a}_i\mathbf{b}_i^\top$
  [Cai and Zhang, 2015, Chen et al., 2015]

Prior Work

- Convex relaxation: $\min_{\mathbf{X}} \frac{1}{2}\|\mathbf{y} - \mathcal{A}(\mathbf{X})\|_2^2 + \lambda\|\mathbf{X}\|_*$

  [Recht et al., 2010, Candès and Plan, 2011]

  Theoretical properties ✔ computation can be intensive

- Non-convex methods: enforce rank r constraint
  - Factorize $\mathbf{X} = \mathbf{R}\mathbf{L}^\top$ + Gradient descent or Alternating Minimization on $\mathbf{R} \in \mathbb{R}^{p_1 \times r}, \mathbf{L} \in \mathbb{R}^{p_2 \times r}$

    [Ma et al., 2019, Park et al., 2018, Sun and Luo, 2015, Tu et al., 2016, Wang et al., 2017, Zhao et al., 2015, Zheng and Lafferty, 2015, Jain et al., 2013, Hardt, 2014]...
  - Projected gradient descent (Singular value projection (SVP), Iterative Hard Thresholding (IHT))

    [Goldfarb and Ma, 2011, Jain et al., 2010, Tanner and Wei, 2013]...
  - Manifold optimization [Boumal and Absil, 2011, Keshavan et al., 2009, Vandereycken, 2013, Wei et al., 2016, Huang and Hand, 2018]
  - ...

Prior Work

- Convex relaxation: $\min_{\mathbf{X}} \frac{1}{2}\|\mathbf{y} - \mathcal{A}(\mathbf{X})\|_2^2 + \lambda\|\mathbf{X}\|_*$

  [Recht et al., 2010, Candès and Plan, 2011]

  Theoretical properties ✔ computation can be intensive

- Non-convex methods: enforce rank r constraint
  - Factorize $\mathbf{X} = \mathbf{R}\mathbf{L}^\top$ + Gradient descent or Alternating Minimization on $\mathbf{R} \in \mathbb{R}^{p_1 \times r}, \mathbf{L} \in \mathbb{R}^{p_2 \times r}$

    [Ma et al., 2019, Park et al., 2018, Sun and Luo, 2015, Tu et al., 2016, Wang et al., 2017, Zhao et al., 2015, Zheng and Lafferty, 2015, Jain et al., 2013, Hardt, 2014]...
  - Projected gradient descent (Singular value projection (SVP), Iterative Hard Thresholding (IHT))

    [Goldfarb and Ma, 2011, Jain et al., 2010, Tanner and Wei, 2013]...
  - Manifold optimization [Boumal and Absil, 2011, Keshavan et al., 2009, Vandereycken, 2013, Wei et al., 2016, Huang and Hand, 2018]
  - ...
- Most of existing algorithms
  - require careful tuning or
  - have a convergence rate no faster than linear.

  $\Longrightarrow$ Can we do better?

Our Algorithm: RISRO

Recursive Importance Sketching algorithm for *Rank* constrained least squares *Optimization* (RISRO).

Advantages

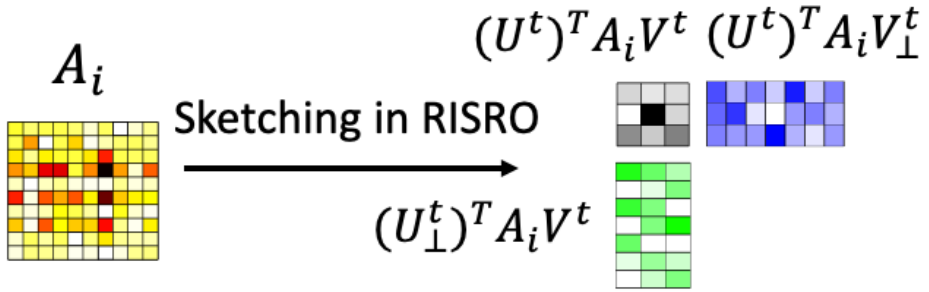
- Tuning free
- High-order convergence guarantees under proper assumptions

RISRO-Procedure

1. Input $\mathbf{y}, \mathcal{A}$, and initialization $\mathbf{X}^0$ with (economic) SVD $\mathbf{U}^0\mathbf{\Sigma}^0\mathbf{V}^{0\top}$
2. For $t = 0, 1, \ldots$
   - Perform importance sketching on $\mathcal{A}$.
   - Solve a dimension reduced least squares.
   - Update sketching matrices.

RISRO-Procedure

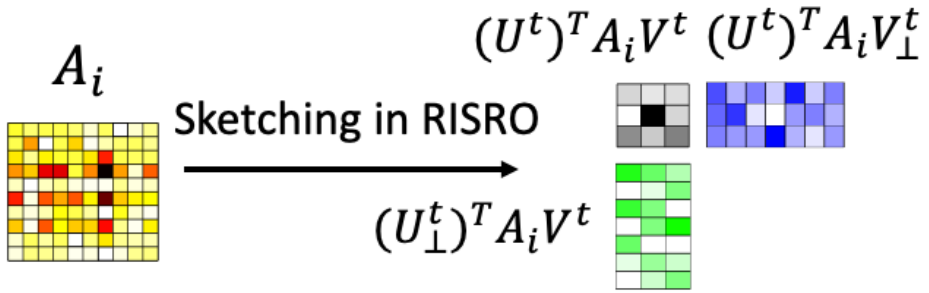
1. Input $\mathbf{y}, \mathcal{A}$, and initialization $\mathbf{X}^0$ with (economic) SVD $\mathbf{U}^0\mathbf{\Sigma}^0\mathbf{V}^{0\top}$
2. For $t = 0, 1, \ldots$
   - Perform importance sketching on $\mathcal{A}$. Construct importance covariates $\mathbf{A}_i^B := \mathbf{U}^{t\top}\mathbf{A}_i\mathbf{V}^t$, $\mathbf{A}_i^{D_1} := \mathbf{U}_\perp^{t\top}\mathbf{A}_i\mathbf{V}^t$, $\mathbf{A}_i^{D_2} := \mathbf{U}^{t\top}\mathbf{A}_i\mathbf{V}_\perp^t$



   - Solve a dimension reduced least squares.
   - Update sketching matrices.

RISRO-Procedure

1. Input $\mathbf{y}, \mathcal{A}$, and initialization $\mathbf{X}^0$ with (economic) SVD $\mathbf{U}^0\mathbf{\Sigma}^0\mathbf{V}^{0\top}$
2. For $t = 0, 1, \ldots$
   - Perform importance sketching on $\mathcal{A}$. Construct importance covariates $\mathbf{A}_i^B := \mathbf{U}^{t\top}\mathbf{A}_i\mathbf{V}^t$, $\mathbf{A}_i^{D_1} := \mathbf{U}_\perp^{t\top}\mathbf{A}_i\mathbf{V}^t$, $\mathbf{A}_i^{D_2} := \mathbf{U}^{t\top}\mathbf{A}_i\mathbf{V}_\perp^t$



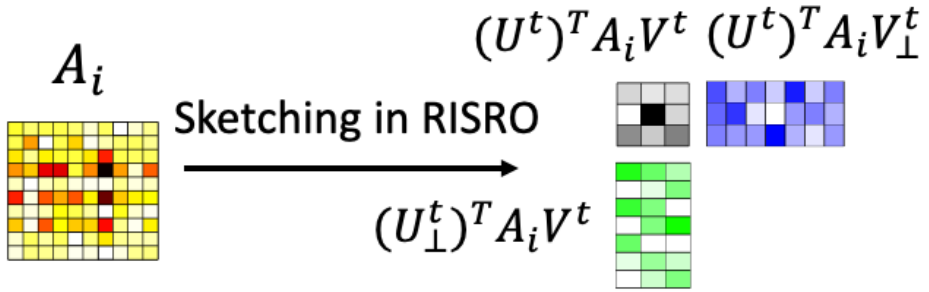
   - Solve a dimension reduced least squares.

$$(\mathbf{B}^{t+1}, \mathbf{D}_1^{t+1}, \mathbf{D}_2^{t+1}) = \underset{\mathbf{B},\mathbf{D}_1,\mathbf{D}_2}{\arg\min} \sum_{i=1}^{n} \left(\mathbf{y}_i - \langle \mathbf{A}_i^B, \mathbf{B}\rangle - \langle \mathbf{A}_i^{D_1}, \mathbf{D}_1\rangle - \langle \mathbf{A}_i^{D_2}, \mathbf{D}_2^\top\rangle\right)^2$$

   - Update sketching matrices.

RISRO-Procedure

1. Input $\mathbf{y}, \mathcal{A}$, and initialization $\mathbf{X}^0$ with (economic) SVD $\mathbf{U}^0\mathbf{\Sigma}^0\mathbf{V}^{0\top}$
2. For $t = 0, 1, \ldots$
   - Perform importance sketching on $\mathcal{A}$. Construct importance covariates $\mathbf{A}_i^B := \mathbf{U}^{t\top}\mathbf{A}_i\mathbf{V}^t$, $\mathbf{A}_i^{D_1} := \mathbf{U}_\perp^{t\top}\mathbf{A}_i\mathbf{V}^t$, $\mathbf{A}_i^{D_2} := \mathbf{U}^{t\top}\mathbf{A}_i\mathbf{V}_\perp^t$



   - Solve a dimension reduced least squares.

$$(\mathbf{B}^{t+1}, \mathbf{D}_1^{t+1}, \mathbf{D}_2^{t+1}) = \underset{\mathbf{B},\mathbf{D}_1,\mathbf{D}_2}{\arg\min} \sum_{i=1}^{n} \left(\mathbf{y}_i - \langle \mathbf{A}_i^B, \mathbf{B}\rangle - \langle \mathbf{A}_i^{D_1}, \mathbf{D}_1\rangle - \langle \mathbf{A}_i^{D_2}, \mathbf{D}_2^\top\rangle\right)^2$$

   - Update sketching matrices. Let $\mathbf{X}_U^{t+1} = (\mathbf{U}^t\mathbf{B}^{t+1} + \mathbf{U}_\perp^t\mathbf{D}_1^{t+1})$, $\mathbf{X}_V^{t+1} = (\mathbf{V}^t\mathbf{B}^{t+1\top} + \mathbf{V}_\perp^t\mathbf{D}_2^{t+1})$. Update $\mathbf{U}^{t+1} = \mathrm{QR}(\mathbf{X}_U^{t+1})$, $\mathbf{V}^{t+1} = \mathrm{QR}(\mathbf{X}_V^{t+1})$.
   - (Optional) $\mathbf{X}^{t+1} = \mathbf{X}_U^{t+1}\left(\mathbf{B}^{t+1}\right)^\dagger \mathbf{X}_V^{t+1\top}$

$\mathrm{QR}(\cdot)$ is the Q part in QR decomposition and $(\cdot)^\dagger$ is the Moore-Penrose inverse

RISRO-Intuition

Suppose $\mathbf{y}_i = \langle \mathbf{A}_i, \bar{\mathbf{X}} \rangle + \bar{\epsilon}_i$ where $\bar{\mathbf{X}}$ is a rank r target matrix. Rewritten

$$\mathbf{y}_i = \langle \mathbf{A}_i^B, \mathbf{U}^{t\top} \bar{\mathbf{X}} \mathbf{V}^t \rangle + \langle \mathbf{A}_i^{D_1}, \mathbf{U}_\perp^{t\top} \bar{\mathbf{X}} \mathbf{V}^t \rangle + \langle \mathbf{A}_i^{D_2}, \mathbf{U}^{t\top} \bar{\mathbf{X}} \mathbf{V}_\perp^t \rangle + \epsilon_i^t,$$

where $\epsilon_i^t = \langle \mathbf{U}_\perp^{t\top} \mathbf{A}_i \mathbf{V}_\perp^t, \mathbf{U}_\perp^{t\top} \bar{\mathbf{X}} \mathbf{V}_\perp^t \rangle + \bar{\epsilon}_i$.

Suppose $\mathbf{y}_i = \langle \mathbf{A}_i, \bar{\mathbf{X}} \rangle + \bar{\epsilon}_i$ where $\bar{\mathbf{X}}$ is a rank r target matrix. Rewritten

$$\mathbf{y}_i = \langle \mathbf{A}_i^B, \mathbf{U}^{t\top} \bar{\mathbf{X}} \mathbf{V}^t \rangle + \langle \mathbf{A}_i^{D_1}, \mathbf{U}_\perp^{t\top} \bar{\mathbf{X}} \mathbf{V}^t \rangle + \langle \mathbf{A}_i^{D_2}, \mathbf{U}^{t\top} \bar{\mathbf{X}} \mathbf{V}_\perp^t \rangle + \epsilon_i^t,$$

where $\epsilon_i^t = \langle \mathbf{U}_\perp^{t\top} \mathbf{A}_i \mathbf{V}_\perp^t, \mathbf{U}_\perp^{t\top} \bar{\mathbf{X}} \mathbf{V}_\perp^t \rangle + \bar{\epsilon}_i$.

If $\epsilon^t = 0$. Then

$$\mathbf{B}^{t+1} = \mathbf{U}^{t\top} \bar{\mathbf{X}} \mathbf{V}^t, \quad \mathbf{D}_1^{t+1} = \mathbf{U}_\perp^{t\top} \bar{\mathbf{X}} \mathbf{V}^t, \quad \mathbf{D}_2^{t+1} = \mathbf{U}^{t\top} \bar{\mathbf{X}} \mathbf{V}_\perp^t$$

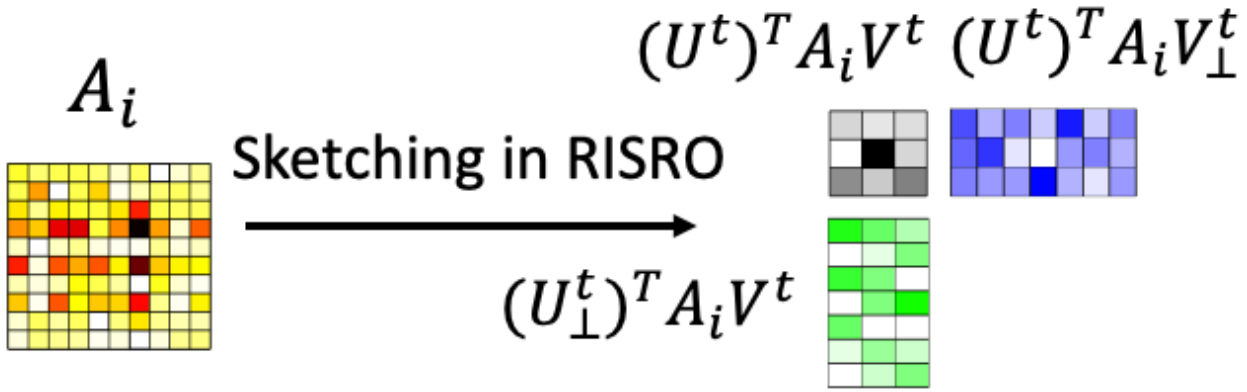
is a solution of the least squares. Moreover if $\mathbf{B}^{t+1}$ is invertible

$$\mathbf{X}^{t+1} = \mathbf{X}_U^{t+1} \left(\mathbf{B}^{t+1}\right)^{-1} \mathbf{X}_V^{t+1\top} = \bar{\mathbf{X}}$$

In general $\epsilon^t \neq 0$, but we hope $\mathbf{X}^t \to \bar{\mathbf{X}}$.

Importance Sketching in RISRO

Sketching: do dimension reduction to speed up the computation



- Comparison of Importance Sketching and Randomized Sketching

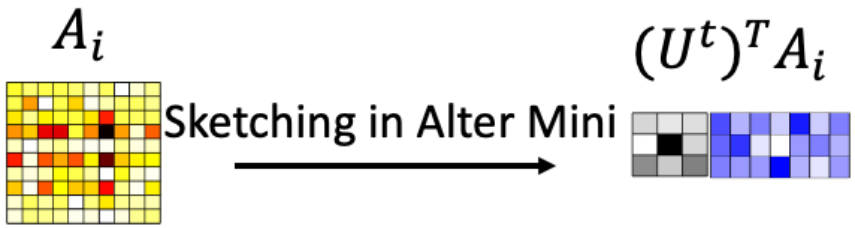
	Importance Sketching	Randomized Sketching [Mahoney, 2011, Woodruff, 2014]
Sketching Matrix	Deterministic, $\mathbf{U}^t$, $\mathbf{V}^t$ (with supervision)	Random
Dimension reduction	Reduce p, hold n	Reduce n, hold p
Statistical efficiency	High	Low

Sketching Interpretations for algorithms in literature

- Alternating Minimization (Alter Mini)
  [Jain et al., 2013, Zhao et al., 2015]

$$\widehat{\mathbf{V}}^{t+1} = \arg\min_{\mathbf{V}\in\mathbb{R}^{p_2\times r}} \sum_{i=1}^{n} \left(\mathbf{y}_i - \langle \mathbf{A}_i, \mathbf{U}^t\mathbf{V}^\top \rangle\right)^2 = \arg\min_{\mathbf{V}\in\mathbb{R}^{p_2\times r}} \sum_{i=1}^{n} \left(\mathbf{y}_i - \langle \mathbf{U}^{t\top}\mathbf{A}_i, \mathbf{V}^\top \rangle\right)^2,$$

$$\mathbf{V}^{t+1} = \mathrm{QR}(\widehat{\mathbf{V}}^{t+1})$$

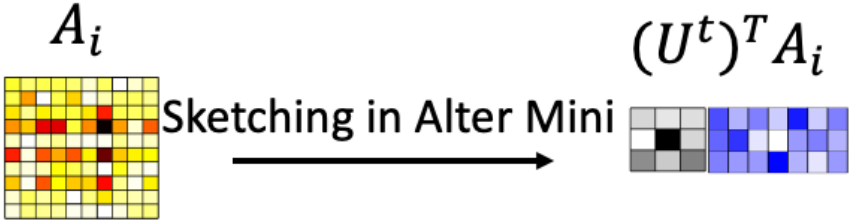


Sketching Interpretations for algorithms in literature

- Alternating Minimization (Alter Mini)
  [Jain et al., 2013, Zhao et al., 2015]

$$\widehat{\mathbf{V}}^{t+1} = \arg\min_{\mathbf{V}\in\mathbb{R}^{p_2\times r}} \sum_{i=1}^{n} \left(\mathbf{y}_i - \langle \mathbf{A}_i, \mathbf{U}^t\mathbf{V}^\top \rangle\right)^2 = \arg\min_{\mathbf{V}\in\mathbb{R}^{p_2\times r}} \sum_{i=1}^{n} \left(\mathbf{y}_i - \langle \mathbf{U}^{t\top}\mathbf{A}_i, \mathbf{V}^\top \rangle\right)^2,$$

$$\mathbf{V}^{t+1} = \mathrm{QR}(\widehat{\mathbf{V}}^{t+1})$$



- Rank 2r iterative least squares (R2RILS) for matrix completion
  [Bauch and Nadler, 2020]

$$\min_{\mathbf{M}\in\mathbb{R}^{p_1\times r},\mathbf{N}\in\mathbb{R}^{p_2\times r}} \sum_{(i,j)\in\Omega} \left\{ \left(\mathbf{U}^t\mathbf{N}^\top + \mathbf{M}\mathbf{V}^{t\top} - \mathbf{Y}\right)_{[i,j]} \right\}^2,$$

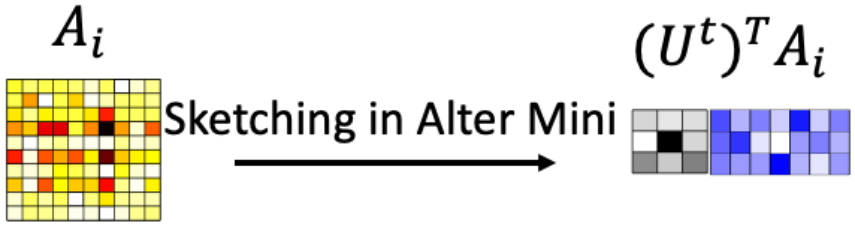
  Ω is the observed entry indices.

Sketching Interpretations for algorithms in literature

- Alternating Minimization (Alter Mini)
  [Jain et al., 2013, Zhao et al., 2015]

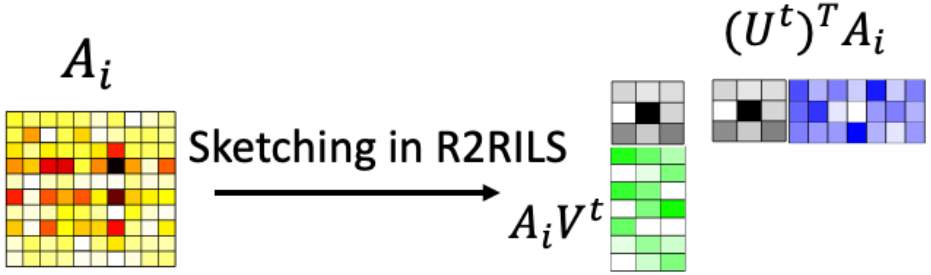
$$\widehat{\mathbf{V}}^{t+1} = \arg\min_{\mathbf{V}\in\mathbb{R}^{p_2\times r}} \sum_{i=1}^{n} \left(\mathbf{y}_i - \langle \mathbf{A}_i, \mathbf{U}^t\mathbf{V}^\top\rangle\right)^2 = \arg\min_{\mathbf{V}\in\mathbb{R}^{p_2\times r}} \sum_{i=1}^{n} \left(\mathbf{y}_i - \langle \mathbf{U}^{t\top}\mathbf{A}_i, \mathbf{V}^\top\rangle\right)^2,$$

$$\mathbf{V}^{t+1} = \mathrm{QR}(\widehat{\mathbf{V}}^{t+1})$$

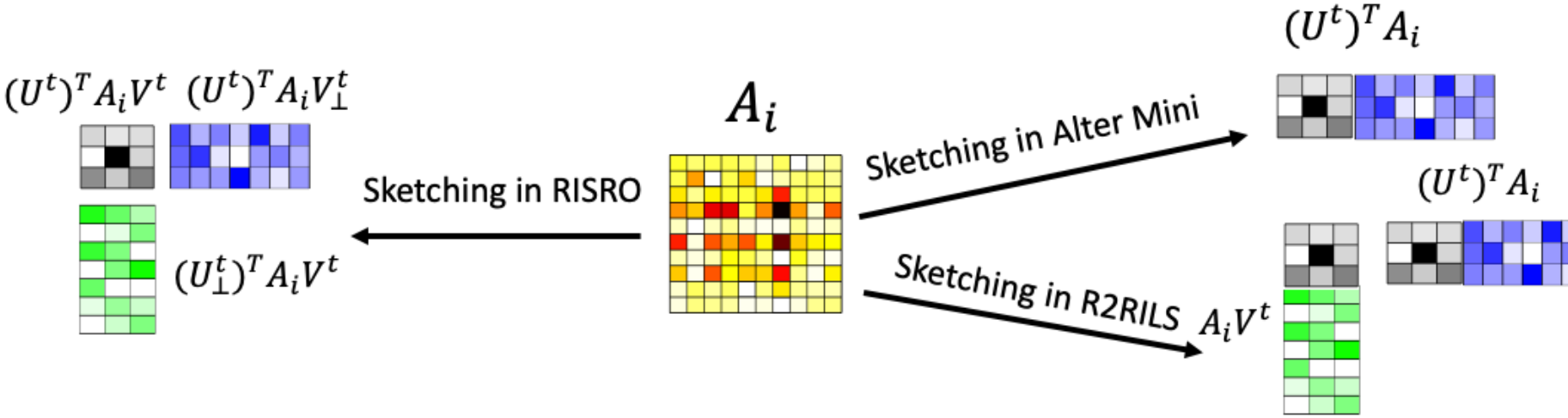


- Rank 2r iterative least squares (R2RILS) for matrix completion
  [Bauch and Nadler, 2020]

$$\sum_{(i,j)\in\Omega} \left\{ \left(\mathbf{U}^t\mathbf{N}^\top + \mathbf{M}\mathbf{V}^{t\top} - \mathbf{Y}\right)_{[i,j]} \right\}^2 \Longleftrightarrow \sum_{(i,j)\in\Omega} \left(\mathbf{Y}_{[i,j]} - \langle \mathbf{U}^{t\top}\mathbf{A}^{ij}, \mathbf{N}^\top\rangle - \langle \mathbf{M}, \mathbf{A}^{ij}\mathbf{V}^t\rangle\right)$$

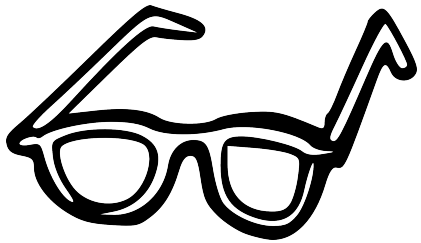


Sketching Interpretations for algorithms in literature



- Alter Mini: Miss one set of covariates $\Longrightarrow$ large iteration error
- R2RILS: Double core sketch
  $\Longrightarrow \begin{cases} \text{Rank deficiency in the least squares} \\ \text{Hard in theory and implementation} \end{cases}$
- ★ RISRO: resolve both issues $\Longrightarrow$ High-order convergence!

Convergence Analysis



Convergence Analysis: Assumption

- Restricted isometry property (RIP)

  [Candès, 2008, Recht et al., 2010] :

  $\mathcal{A} : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^n$ satisfies the $r-$RIP with r-restricted isometry constant $\delta \in [0, 1)$ if

  $$(1 - \delta)\|\mathbf{Z}\|_{\mathbf{F}}^2 \leq \|\mathcal{A}(\mathbf{Z})\|_2^2 \leq (1 + \delta)\|\mathbf{Z}\|_{\mathbf{F}}^2$$

  holds for all $\mathbf{Z}$ of rank at most r.

  - Widely used [Cai and Zhang, 2013, Candès and Plan, 2011]...
  - Easy to satisfy in random subgaussian design [Candès and Plan, 2011]
  - Key quantity in landscape analysis [Bhojanapalli et al., 2016, Ge et al., 2017, Uschmajew and Vandereycken, 2018]

Convergence Analysis: Assumption

- Restricted isometry property (RIP)

  [Candès, 2008, Recht et al., 2010] :

  $\mathcal{A} : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^n$ satisfies the $r-$RIP with r-restricted isometry constant $\delta \in [0, 1)$ if

  $$(1 - \delta)\|\mathbf{Z}\|_{\mathbf{F}}^2 \leq \|\mathcal{A}(\mathbf{Z})\|_2^2 \leq (1 + \delta)\|\mathbf{Z}\|_{\mathbf{F}}^2$$

  holds for all $\mathbf{Z}$ of rank at most r.

  - Widely used [Cai and Zhang, 2013, Candès and Plan, 2011]...
  - Easy to satisfy in random subgaussian design [Candès and Plan, 2011]
  - Key quantity in landscape analysis [Bhojanapalli et al., 2016, Ge et al., 2017, Uschmajew and Vandereycken, 2018]

  Caveat: RIP may not hold in some scenarios such as matrix completion.

RISRO Convergence Analysis

Let $\bar{\mathbf{X}}$ be a rank r stationary point and $\bar{\epsilon} := \mathbf{y} - \mathcal{A}(\bar{\mathbf{X}})$.
Assume

- $\mathcal{A}$ satisfies $3r$-restricted isometry property (RIP) with RIP constant δ
- Initialization condition: $\|\mathbf{X}^0 - \bar{\mathbf{X}}\|_{\mathbf{F}} \leq C(\delta)\sigma_r(\bar{\mathbf{X}})$
- Small residual (gradient) condition: $\|\mathcal{A}^*(\bar{\epsilon})\|_{\mathbf{F}} \leq C'(\delta)\sigma_r(\bar{\mathbf{X}})$.

$\sigma_r(\bar{\mathbf{X}})$ is the r-th largest singular value of $\bar{\mathbf{X}}$. $\mathcal{A}^*(\mathbf{b}) := \sum_{i=1}^{n} \mathbf{b}_i \mathbf{A}_i$ is the adjoint operator of $\mathcal{A}$.

RISRO Convergence Analysis

Let $\bar{\mathbf{X}}$ be a rank r stationary point and $\bar{\epsilon} := \mathbf{y} - \mathcal{A}(\bar{\mathbf{X}})$.

Theorem 1: Under the assumptions above, $\mathbf{X}^t$ generated by RISRO converges Q-linearly to $\bar{\mathbf{X}}$:

$$\|\mathbf{X}^{t+1} - \bar{\mathbf{X}}\|_{\mathbf{F}} \leq \frac{3}{4}\|\mathbf{X}^t - \bar{\mathbf{X}}\|_{\mathbf{F}}, \quad \forall\, t \geq 0.$$

RISRO Convergence Analysis

Let $\bar{\mathbf{X}}$ be a rank r stationary point and $\bar{\epsilon} := \mathbf{y} - \mathcal{A}(\bar{\mathbf{X}})$.

Theorem 1: Under the assumptions above, $\mathbf{X}^t$ generated by RISRO converges Q-linearly to $\bar{\mathbf{X}}$:

$$\|\mathbf{X}^{t+1} - \bar{\mathbf{X}}\|_{\mathbf{F}} \leq \frac{3}{4}\|\mathbf{X}^t - \bar{\mathbf{X}}\|_{\mathbf{F}}, \quad \forall\, t \geq 0.$$

$$\|\mathbf{X}^{t+1} - \bar{\mathbf{X}}\|_{\mathbf{F}}^2 \leq \frac{c_1(\delta)\|\mathbf{X}^t - \bar{\mathbf{X}}\|^2}{\sigma_r^2(\bar{\mathbf{X}})} \left(\|\mathbf{X}^t - \bar{\mathbf{X}}\|_{\mathbf{F}}^2 + \|\mathcal{A}^*(\bar{\epsilon})\|_{\mathbf{F}}\|\mathbf{X}^t - \bar{\mathbf{X}}\|_{\mathbf{F}} + \|\mathcal{A}^*(\bar{\epsilon})\|_{\mathbf{F}}^2\right),$$
$$\forall\, t \geq 0.$$

RISRO Convergence Analysis

Let $\bar{\mathbf{X}}$ be a rank r stationary point and $\bar{\epsilon} := \mathbf{y} - \mathcal{A}(\bar{\mathbf{X}})$.

Theorem 1: Under the assumptions above, $\mathbf{X}^t$ generated by RISRO converges Q-linearly to $\bar{\mathbf{X}}$:

$$\|\mathbf{X}^{t+1} - \bar{\mathbf{X}}\|_{\mathbf{F}} \leq \frac{3}{4}\|\mathbf{X}^t - \bar{\mathbf{X}}\|_{\mathbf{F}}, \quad \forall\, t \geq 0.$$

$$\|\mathbf{X}^{t+1} - \bar{\mathbf{X}}\|_{\mathbf{F}}^2 \leq \frac{c_1(\delta)\|\mathbf{X}^t - \bar{\mathbf{X}}\|^2}{\sigma_r^2(\bar{\mathbf{X}})} \left(\|\mathbf{X}^t - \bar{\mathbf{X}}\|_{\mathbf{F}}^2 + \|\mathcal{A}^*(\bar{\epsilon})\|_{\mathbf{F}}\|\mathbf{X}^t - \bar{\mathbf{X}}\|_{\mathbf{F}} + \|\mathcal{A}^*(\bar{\epsilon})\|_{\mathbf{F}}^2\right),$$
$$\forall\, t \geq 0.$$

If $\bar{\epsilon} = 0$, then $\{\mathbf{X}^t\}$ converges quadratically to $\bar{\mathbf{X}}$ as

$$\|\mathbf{X}^{t+1} - \bar{\mathbf{X}}\|_{\mathbf{F}} \leq \frac{\sqrt{c_1(\delta)}\|\mathbf{X}^t - \bar{\mathbf{X}}\|_{\mathbf{F}}^2}{\sigma_r(\bar{\mathbf{X}})}, \quad \forall\, t \geq 0.$$

RISRO Convergence Analysis

★ Quadratic-linear convergence

$$\|\mathbf{X}^{t+1}-\bar{\mathbf{X}}\|_{\mathbf{F}}^2 \leq \frac{c_1(\delta)\|\mathbf{X}^t-\bar{\mathbf{X}}\|^2}{\sigma_r^2(\bar{\mathbf{X}})}\left(\|\mathbf{X}^t-\bar{\mathbf{X}}\|_{\mathbf{F}}^2+\|\mathcal{A}^*(\bar{\epsilon})\|_{\mathbf{F}}\|\mathbf{X}^t-\bar{\mathbf{X}}\|_{\mathbf{F}}+\|\mathcal{A}^*(\bar{\epsilon})\|_{\mathbf{F}}^2\right).$$

- when $\|\mathbf{X}^t-\bar{\mathbf{X}}\|_{\mathbf{F}} \gg \|\mathcal{A}^*(\bar{\epsilon})\|_{\mathbf{F}} \Longrightarrow$ quadratic convergence
- when $\|\mathbf{X}^t-\bar{\mathbf{X}}\|_{\mathbf{F}} \leq c\|\mathcal{A}^*(\bar{\epsilon})\|_{\mathbf{F}} \Longrightarrow$ reduce to linear convergence

$\bar{\epsilon}\downarrow \Longrightarrow$ Longer period of quadratic convergence.

RISRO Convergence Analysis

★ Quadratic-linear convergence

$$\|\mathbf{X}^{t+1}-\bar{\mathbf{X}}\|_{\mathbf{F}}^2 \leq \frac{c_1(\delta)\|\mathbf{X}^t-\bar{\mathbf{X}}\|^2}{\sigma_r^2(\bar{\mathbf{X}})}\left(\|\mathbf{X}^t-\bar{\mathbf{X}}\|_{\mathbf{F}}^2+\|\mathcal{A}^*(\bar{\epsilon})\|_{\mathbf{F}}\|\mathbf{X}^t-\bar{\mathbf{X}}\|_{\mathbf{F}}+\|\mathcal{A}^*(\bar{\epsilon})\|_{\mathbf{F}}^2\right).$$

- when $\|\mathbf{X}^t-\bar{\mathbf{X}}\|_{\mathbf{F}} \gg \|\mathcal{A}^*(\bar{\epsilon})\|_{\mathbf{F}} \Longrightarrow$ quadratic convergence
- when $\|\mathbf{X}^t-\bar{\mathbf{X}}\|_{\mathbf{F}} \leq c\|\mathcal{A}^*(\bar{\epsilon})\|_{\mathbf{F}} \Longrightarrow$ reduce to linear convergence

$\bar{\epsilon}\downarrow \Longrightarrow$ Longer period of quadratic convergence.

★ $\bar{\epsilon}=0 \quad \Longrightarrow \quad \mathbf{y}=\mathcal{A}(\bar{\mathbf{X}}) \quad \Longrightarrow$ matrix sensing

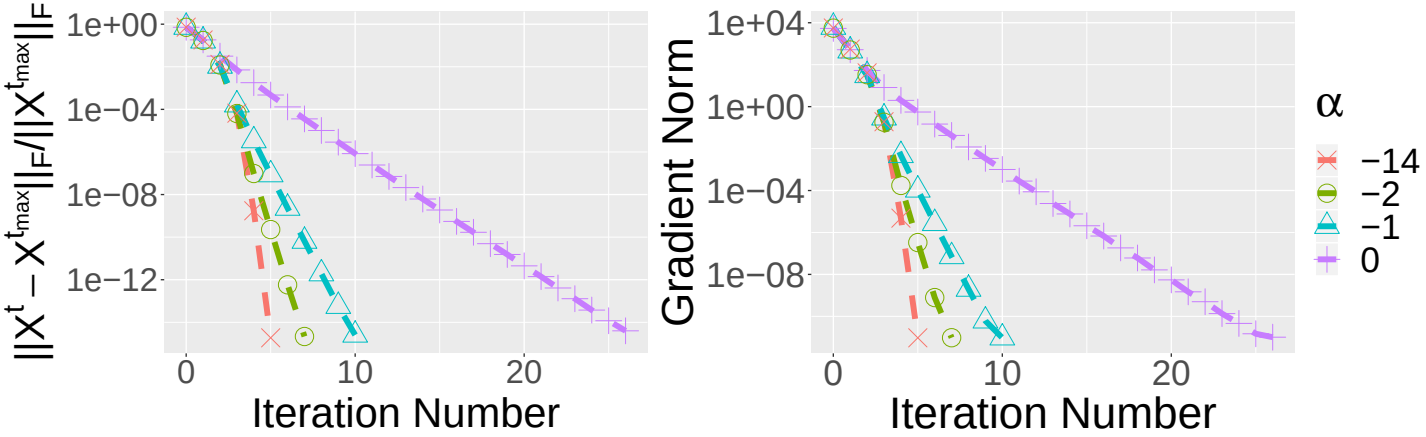
[Recht et al., 2010]

RISRO achieves quadratic convergence

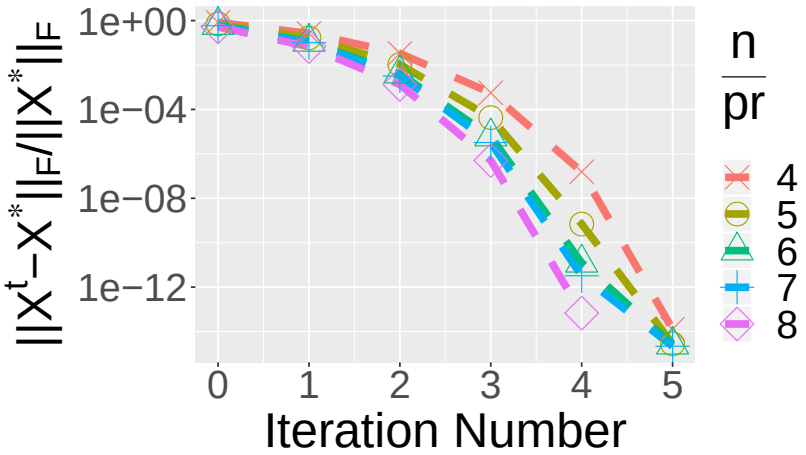
Simulation

$\mathbf{y}_i = \langle \mathbf{A}_i, \mathbf{X}^* \rangle + \epsilon_i$ for $1 \le i \le n$, $\mathbf{A}_i \overset{i.i.d.}{\sim} N(0,1)$ and $\epsilon_i \overset{i.i.d.}{\sim} N(0,\sigma^2)$.
$\mathbf{X}^* \in \mathbb{R}^{p\times p}$ with $p = 100, r = 3, \kappa(\mathbf{X}^*) = 1$ and $\mathbf{X}^0 = \mathrm{SVD}_r(\mathcal{A}^*(\mathbf{y}))$.

- (Quadratic-linear) $n = 5pr$, $\sigma = 10^{\alpha}$ for $\alpha \in \{0, -1, -2, -14\}$



- (Quadratic) $n/(pr) \in \{4, 5, 6, 7, 8\}$, $\sigma = 0$



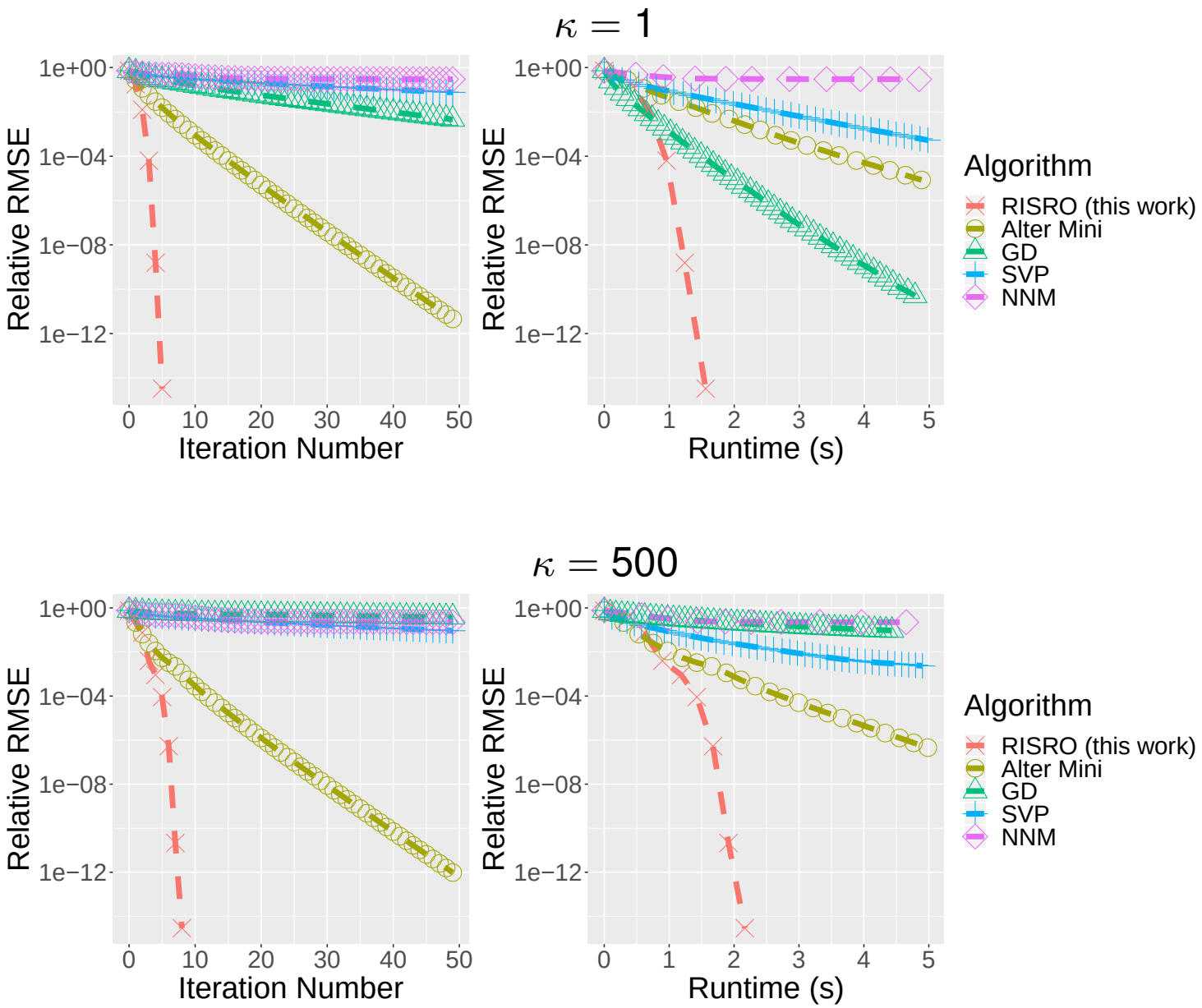
Vs. Other Algorithms

Suppose $p_1 = p_2 = p$ and $n \geq pr$. Under similar assumptions as in Theorem 1:

	GD	PGD (SVP / IHT)	Alter Mini	RISRO (this work)
Iteration Complexity	$O(np^2r)$	$O(np^2)$	$O(np^2r^2)$	$O(np^2r^2)$
Tuning	Yes	Yes	No	No
Convergence	Linear	Linear	Linear	Quadratic-(linear)

★ Improve upon Alter Mini for free

Comparison Simulation $\sigma = 0$



Any connection of RISRO to existing optimization algorithms?

Connection to Riemannian Manifold Optimization

Iteration t of RISRO:

1. Perform importance sketching.
2. Perform a dimension reduced least squares.
3. Update sketching matrices and $\mathbf{X}^{t+1}$.

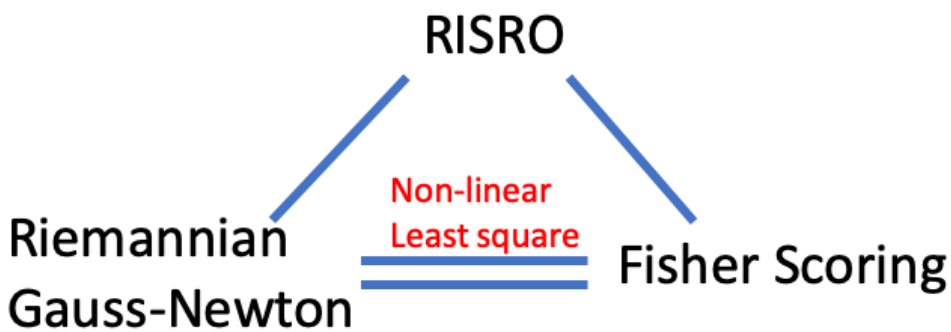
Connection to Riemannian Manifold Optimization

Iteration t of RISRO:

1. Perform importance sketching.
2. Perform a dimension reduced least squares.

   $\Longrightarrow$ Implicitly solves "Fisher Scoring" or "Riemannian Gauss-Newton" equation in Riemannian optimization on fixed rank matrices.
3. Update sketching matrices and $\mathbf{X}^{t+1}$.

   $\Longrightarrow$ Perform a type of retraction in Riemannian optimization literature



Riemannian Manifold Optimization

- Target: optimize a function f defined on a Riemannian manifold $\mathcal{M}$. [Absil et al., 2009]
- Common Riemannian manifolds, embedded submanifold: a smooth subset of $\mathbb{R}^n$ + a Riemannian metric.

Riemannian Manifold Optimization

- Target: optimize a function f defined on a Riemannian manifold $\mathcal{M}$. [Absil et al., 2009]
- Common Riemannian manifolds, embedded submanifold: a smooth subset of $\mathbb{R}^n$ + a Riemannian metric.
- $\mathcal{M}_r = \{\mathbf{X} \in \mathbb{R}^{p_1 \times p_2} : \ \mathrm{rank}(\mathbf{X}) = r\}$

  Riemannian metric: Euclidean inner product,
  $\langle \mathbf{U}, \mathbf{V} \rangle = \mathrm{trace}(\mathbf{U}^\top \mathbf{V})$

Retraction

- Iterative algorithm: $x^{t+1} = x^t + \xi$.

  Manifold optimization: x^{t+1} may not lie in the manifold

  Solution: retraction!

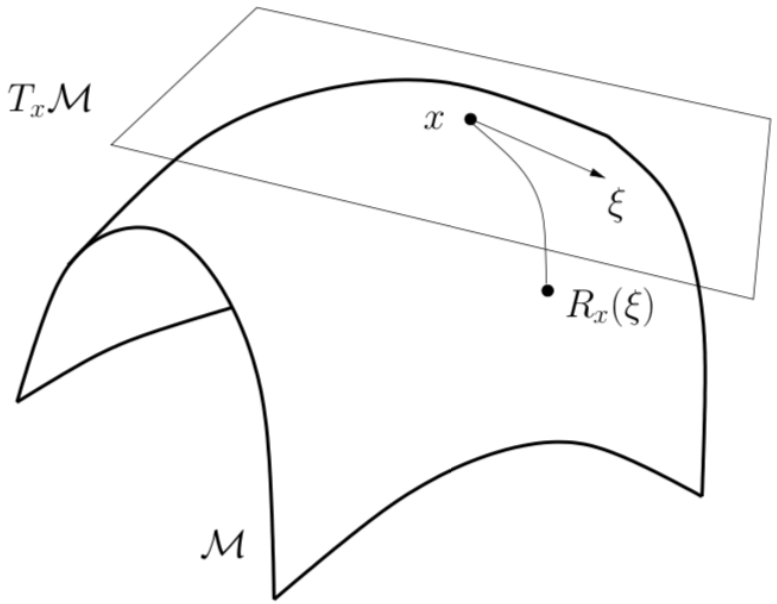
Retraction

- Iterative algorithm: $x^{t+1} = x^t + \xi$.

  Manifold optimization: x^{t+1} may not lie in the manifold

  Solution: retraction!

- Retraction: a smooth map that brings the vector in the tangent space back to the manifold. Denote $T_x\mathcal{M}$ as the tangent space at x



[Absil et al., 2009, Section 4.1]

$R : T\mathcal{M} \to \mathcal{M}, x \times \xi \to R_x(\xi) \in \mathcal{M}$.

Retraction

- Iterative algorithm: $x^{t+1} = x^t + \xi$.

  Manifold optimization: x^{t+1} may not lie in the manifold

  Solution: retraction!

- Retraction: a smooth map that brings the vector in the tangent space back to the manifold. Denote $T_x\mathcal{M}$ as the tangent space at x

★ Let η^t be the update direction such that $\mathbf{X}^t + \eta^t$ has the following representation,

$$\mathbf{X}^t + \eta^t = [\mathbf{U}^t \quad \mathbf{U}_\perp^t] \begin{bmatrix} \mathbf{B}^{t+1} & \mathbf{D}_2^{t+1\top} \\ \mathbf{D}_1^{t+1} & \mathbf{0} \end{bmatrix} [\mathbf{V}^t \quad \mathbf{V}_\perp^t]^\top.$$

★ $\mathbf{X}^t + \eta^t \Longrightarrow \mathbf{X}^{t+1}$. Retraction is:

$$\mathbf{X}^{t+1} = R_{\mathbf{X}^t}(\eta^t) = [\mathbf{U}^t \quad \mathbf{U}_\perp^t] \begin{bmatrix} \mathbf{B}^{t+1} & \mathbf{D}_2^{t+1\top} \\ \mathbf{D}_1^{t+1} & \mathbf{D}_1^{t+1}(\mathbf{B}^{t+1})^{-1}\mathbf{D}_2^{t+1\top} \end{bmatrix} [\mathbf{V}^t \quad \mathbf{V}_\perp^t]^\top$$

★ η^t solves the Fisher Scoring or Riemannian Gauss-Newton

Connection to Riemannian Optimization

Recall $f(\mathbf{X}) := \frac{1}{2}\|\mathbf{y} - \mathcal{A}(\mathbf{X})\|_2^2$.

- Riemannian Gradient: $\operatorname{grad} f(\mathbf{X})$
- Riemannian Hessian: $\operatorname{Hess} f(\mathbf{X})$
- Riemannian Newton direction η_{Newton}

$$-\operatorname{grad} f(\mathbf{X}) = \operatorname{Hess} f(\mathbf{X})[\eta_{\text{Newton}}]$$

Connection to Riemannian Optimization

Recall $f(\mathbf{X}) := \frac{1}{2}\|\mathbf{y} - \mathcal{A}(\mathbf{X})\|_2^2$.

- Riemannian Gradient: $\operatorname{grad} f(\mathbf{X}) = P_{T_{\mathbf{X}}}(\mathcal{A}^*(\mathcal{A}(\mathbf{X}) - \mathbf{y}))$.
  $P_{T_{\mathbf{X}}}(\cdot)$ is the orthogonal projector onto the tangent space at $\mathbf{X}$.
- Riemannian Hessian: $\operatorname{Hess} f(\mathbf{X})$
  $[\eta] = P_{T_{\mathbf{X}}}(\mathcal{A}^*(\mathcal{A}(\eta))) + h(\mathbf{y} - \mathcal{A}(\mathbf{X}))$.
  $h(\cdot)$ depends on $\mathbf{X}, \eta$.
- Riemannian Newton direction η_{Newton}

$$-\operatorname{grad} f(\mathbf{X}) = \operatorname{Hess} f(\mathbf{X})[\eta_{\text{Newton}}]$$

$$\Longleftrightarrow -\operatorname{grad} f(\mathbf{X}) = P_{T_{\mathbf{X}}}(\mathcal{A}^*(\mathcal{A}(\eta_{\text{Newton}}))) + h(\mathbf{y} - \mathcal{A}(\mathbf{X}))$$

Connection to Riemannian Optimization

Recall $f(\mathbf{X}) := \frac{1}{2}\|\mathbf{y} - \mathcal{A}(\mathbf{X})\|_2^2$.

- Riemannian Gradient: $\operatorname{grad} f(\mathbf{X}) = P_{T_\mathbf{X}}(\mathcal{A}^*(\mathcal{A}(\mathbf{X}) - \mathbf{y}))$.
  $P_{T_\mathbf{X}}(\cdot)$ is the orthogonal projector onto the tangent space at $\mathbf{X}$.
- Riemannian Hessian: $\operatorname{Hess} f(\mathbf{X})$
  $[\eta] = P_{T_\mathbf{X}}(\mathcal{A}^*(\mathcal{A}(\eta))) + h(\mathbf{y} - \mathcal{A}(\mathbf{X}))$.
  $h(\cdot)$ depends on $\mathbf{X}, \eta$.
- Riemannian Newton direction η_{Newton}

$$-\operatorname{grad} f(\mathbf{X}) = \operatorname{Hess} f(\mathbf{X})[\eta_{\text{Newton}}]$$

$$\Longleftrightarrow -\operatorname{grad} f(\mathbf{X}) = P_{T_\mathbf{X}}(\mathcal{A}^*(\mathcal{A}(\eta_{\text{Newton}}))) + h(\mathbf{y} - \mathcal{A}(\mathbf{X}))$$

- Update in RISRO:

$$\mathbf{X}^t + \eta^t = [\mathbf{U}^t \quad \mathbf{U}_\perp^t] \begin{bmatrix} \mathbf{B}^{t+1} & \mathbf{D}_2^{t+1\top} \\ \mathbf{D}_1^{t+1} & \mathbf{0} \end{bmatrix} [\mathbf{V}^t \quad \mathbf{V}_\perp^t]^\top.$$

Connection to Riemannian Optimization

Recall $f(\mathbf{X}) := \frac{1}{2}\|\mathbf{y} - \mathcal{A}(\mathbf{X})\|_2^2$.

- Riemannian Gradient: $\operatorname{grad} f(\mathbf{X}) = P_{T_{\mathbf{X}}}(\mathcal{A}^*(\mathcal{A}(\mathbf{X}) - \mathbf{y}))$.
  $P_{T_{\mathbf{X}}}(\cdot)$ is the orthogonal projector onto the tangent space at $\mathbf{X}$.
- Riemannian Hessian: $\operatorname{Hess} f(\mathbf{X})$
  $[\eta] = P_{T_{\mathbf{X}}}(\mathcal{A}^*(\mathcal{A}(\eta))) + h(\mathbf{y} - \mathcal{A}(\mathbf{X}))$.
  $h(\cdot)$ depends on $\mathbf{X}, \eta$.
- Riemannian Newton direction η_{Newton}

$$-\operatorname{grad} f(\mathbf{X}) = \operatorname{Hess} f(\mathbf{X})[\eta_{\text{Newton}}]$$

$$\Longleftrightarrow -\operatorname{grad} f(\mathbf{X}) = P_{T_{\mathbf{X}}}(\mathcal{A}^*(\mathcal{A}(\eta_{\text{Newton}}))) + h(\mathbf{y} - \mathcal{A}(\mathbf{X}))$$

- Update in RISRO:

$$\mathbf{X}^t + \eta^t = [\mathbf{U}^t \quad \mathbf{U}_\perp^t] \begin{bmatrix} \mathbf{B}^{t+1} & \mathbf{D}_2^{t+1\top} \\ \mathbf{D}_1^{t+1} & \mathbf{0} \end{bmatrix} [\mathbf{V}^t \quad \mathbf{V}_\perp^t]^\top.$$

Theorem 2: η^t solves

$$-\operatorname{grad} f(\mathbf{X}^t) = P_{T_{\mathbf{X}^t}}(\mathcal{A}^*(\mathcal{A}(\eta))).$$

Connection of RISRO and Riemannian optimization

Suppose $\mathbf{y} = \mathcal{A}(\mathbf{X}) + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is a fixed matrix and $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. Then for any η,

$$\{\mathbb{E}(\mathrm{Hess} f(\mathbf{X})[\eta])\}\,|_{\mathbf{X}=\mathbf{X}^t} = P_{T_{\mathbf{X}^t}}\left(\mathcal{A}^*(\mathcal{A}(\eta))\right).$$

Connection of RISRO and Riemannian optimization

Suppose $\mathbf{y} = \mathcal{A}(\mathbf{X}) + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is a fixed matrix and $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. Then for any η,

$$\{\mathbb{E}(\mathrm{Hess} f(\mathbf{X})[\eta])\}\,|_{\mathbf{X}=\mathbf{X}^t} = P_{T_{\mathbf{X}^t}}\left(\mathcal{A}^*(\mathcal{A}(\eta))\right).$$

By Theorem 2, η^t solves

$$-\mathrm{grad} f(\mathbf{X}^t) = \{\mathbb{E}(\mathrm{Hess} f(\mathbf{X})[\eta])\}\,|_{\mathbf{X}=\mathbf{X}^t}.$$

Connection of RISRO and Riemannian optimization

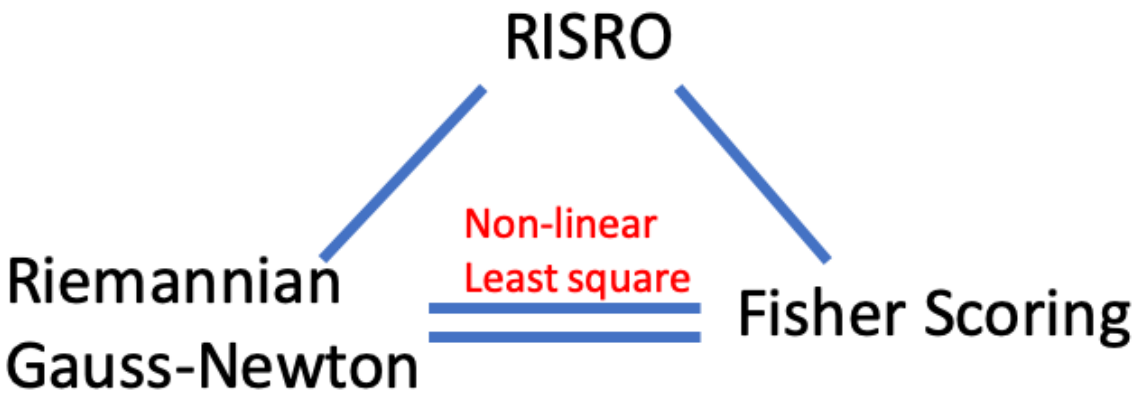
Suppose $\mathbf{y} = \mathcal{A}(\mathbf{X}) + \epsilon$, where $\mathbf{X}$ is a fixed matrix and $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. Then for any η,

$$\{\mathbb{E}(\mathrm{Hess} f(\mathbf{X})[\eta])\}\,|_{\mathbf{X}=\mathbf{X}^t} = P_{T_{\mathbf{X}^t}}\left(\mathcal{A}^*(\mathcal{A}(\eta))\right).$$

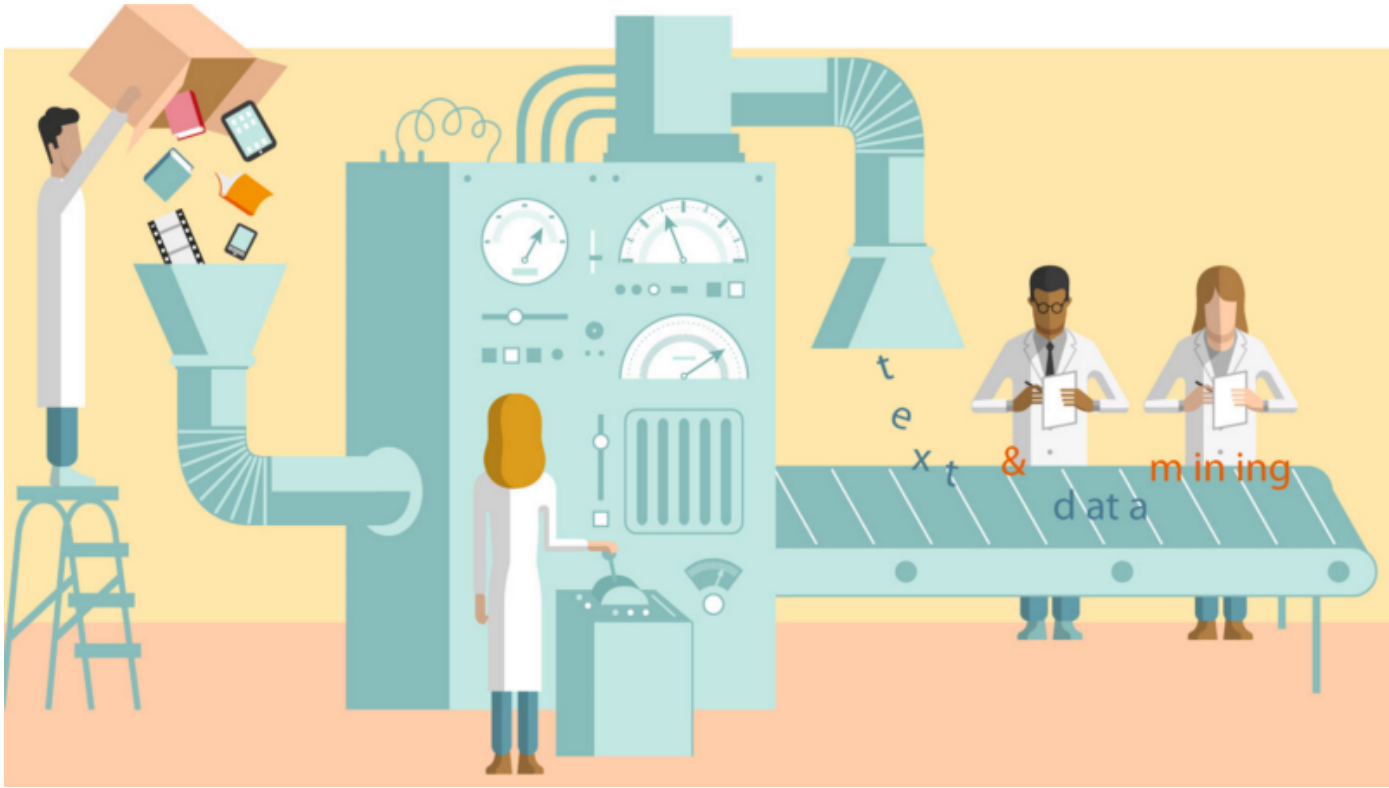
By Theorem 2, η^t solves

$$-\mathrm{grad} f(\mathbf{X}^t) = \{\mathbb{E}(\mathrm{Hess} f(\mathbf{X})[\eta])\}\,|_{\mathbf{X}=\mathbf{X}^t}.$$

This algorithm is called Fisher Scoring in literature [Lange, 2010].



Applications to Statistics and Machine Learning



Applications to Statistics and Machine Learning

- Low-rank matrix trace regression model:

$$\mathbf{y}_i = \langle \mathbf{A}_i, \mathbf{X}^* \rangle + \epsilon_i, \qquad \text{for } 1 \le i \le n,$$

  $\mathbf{X}^* \in \mathbb{R}^{p_1 \times p_2}$ is the true model parameter and $\operatorname{rank}(\mathbf{X}^*) = r$.

- Phase retrieval

$$\mathbf{y}_i = |\langle \mathbf{a}_i, \mathbf{x}^* \rangle|^2 \quad \text{for} \quad 1 \le i \le n,$$

  $\mathbf{x}^* \in \mathbb{R}^p$.

Goal: estimate or recovery $\mathbf{X}^*$ (or $\mathbf{x}^*$).

Low-rank matrix trace regression

Theorem: Suppose $\mathcal{A}$ satisfies the 3r-RIP with RIP constant δ and

- $\|\mathbf{X}^0 - \mathbf{X}^*\|_{\mathbf{F}} \leq C(\delta) \cdot \sigma_r(\mathbf{X}^*)$
- $\sigma_r(\mathbf{X}^*) \geq C'(\delta) \cdot \sqrt{r}\|\mathcal{A}^*(\epsilon)\|$.

Then iterations generated by RISRO satisfy

$$\|\mathbf{X}^{t+1} - \mathbf{X}^*\|_{\mathbf{F}}^2 \leq c_1(\delta)\|\mathbf{X}^t - \mathbf{X}^*\|^2 \left(\frac{\|\mathbf{X}^t - \mathbf{X}^*\|_{\mathbf{F}}^2}{\sigma_r^2(\mathbf{X}^*)} + \frac{\sqrt{r}\|\mathcal{A}^*(\epsilon)\|}{\sigma_r(\mathbf{X}^*)} \right) + c_2(\delta) r \|\mathcal{A}^*(\epsilon)\|^2,$$

for all $t \geq 0$.

★ First term: Decreases quadratic-linearly.

★ Second term: Statistical error independent of t.

Summary

- Introduce a new algorithm, RISRO, for rank constrained least squares.
  $\Longrightarrow$ Tuning free, fast and has high-order convergence
- Introduce the recursive importance sketching framework
  $\Longrightarrow$ Provide a platform to compare different algorithms from a sketching perspective
- Connect RISRO with Riemannian optimization

Thank you! Questions?

Luo, Y., Huang, W., Li, X., & Zhang, A. R. (2020). Recursive Importance Sketching for Rank Constrained Least Squares: Algorithms and High-order Convergence. arXiv preprint arXiv:2011.08360.

Bibliography


Absil, P.-A., Mahony, R., and Sepulchre, R. (2009).
Optimization algorithms on matrix manifolds.
Princeton University Press.


Bauch, J. and Nadler, B. (2020).
Rank 2*r* iterative least squares: efficient recovery of ill-conditioned low rank matrices from few entries.
arXiv preprint arXiv:2002.01849.


Bhojanapalli, S., Neyshabur, B., and Srebro, N. (2016).
Global optimality of local search for low rank matrix recovery.
In *Advances in Neural Information Processing Systems*, pages 3873–3881.


Boumal, N. and Absil, P.-a. (2011).
Rtrmc: A Riemannian trust-region method for low-rank matrix completion.
In *Advances in neural information processing systems*