Dealing with Constraints via Random Permutation

Ruoyu Sun  
UIUC

Joint work with Zhi-Quan Luo (U of Minnesota and CUHK (SZ)) and Yinyu Ye (Stanford)

Simons Institute Workshop on “Fast Iterative Methods in Optimization”  
October 3, 2017
Motivation
Optimization for Large-scale Problems

- How to solve large-scale constrained problems?
- Popular idea: solve small subproblems
  - CD (Coordinate Descent)-type: $\min f(x_1, \ldots, x_N)$.
    $x_1 \rightarrow x_2 \rightarrow \cdots \rightarrow x_N$
- SGD (Stochastic Gradient Descent): $\min \sum_i f_i(x_i)$.
  $f_1 \rightarrow f_2 \rightarrow \cdots \rightarrow f_N$
- Widely (and wildly) used in practice: deep learning, glmnet for LASSO, libsvm for SVM, recommendation systems, EM
- Compared to other ideas, e.g., first-order methods and sketching:
  - Similar cheap iteration idea
  - “Orthogonal” to other ideas, so can combine
Optimization for Large-scale Problems

- How to solve large-scale constrained problems?
- Popular idea: **solve small subproblems**
  - **CD** (Coordinate Descent)-type: \( \min f(x_1, \ldots, x_N) \).
    \[ x_1 \rightarrow x_2 \rightarrow \cdots \rightarrow x_N \]
  - **SGD** (Stochastic Gradient Descent): \( \min \sum_i f_i(x) \).
    \[ f_1 \rightarrow f_2 \rightarrow \cdots \rightarrow f_N \]
- Widely (and wildly) used in practice: deep learning, glmnet for LASSO, libsvm for SVM, recommendation systems, EM
Optimization for Large-scale Problems

- How to solve large-scale constrained problems?
- Popular idea: solve small subproblems
  - CD (Coordinate Descent)-type: \( \min f(x_1, \ldots, x_N) \).
    \( x_1 \rightarrow x_2 \rightarrow \cdots \rightarrow x_N \)
  - SGD (Stochastic Gradient Descent): \( \min \sum_i f_i(x) \).
    \( f_1 \rightarrow f_2 \rightarrow \cdots \rightarrow f_N \)
- Widely (and wildly) used in practice: deep learning, glmnet for LASSO, libsvm for SVM, recommendation systems, EM
- Compared to other ideas, e.g., first-order methods and sketching:
  - Similar cheap iteration idea
  - “Orthogonal” to other ideas, so can combine
Go Beyond Unconstrained Optimization

- Many problems have (linear) constraints

- **Classical convex optimization**, e.g., linear programming.
  - Combinatorial optimization (this workshop)
  - Operations research problems

- **Machine learning** applications, e.g., structured sparsity and deep learning

- Can we apply the *decomposition* idea? *Turn out to be tricky!*

- Algorithm: CD + multiplier $\rightarrow$ **ADMM** (Alternating Direction Method of Multipliers)
Multi-block ADMM

Consider a linearly constrained problem

$$\min_{x \in \mathbb{R}^N} f(x_1, x_2, \ldots, x_n)$$

subject to

$$Ax \triangleq A_1 x_1 + \cdots + A_n x_n = b,$$

$$x_j \in \mathcal{X}_j \subseteq \mathbb{R}^{d_j}, \quad j = 1, \ldots, n.$$
Multi-block ADMM

- Consider a linearly constrained problem

\[
\min_{x \in \mathbb{R}^N} \ f(x_1, x_2, \ldots, x_n) \\
\text{s.t.} \quad Ax \triangleq A_1 x_1 + \cdots + A_n x_n = b, \\
x_j \in \mathcal{X}_j \subseteq \mathbb{R}^{d_j}, \ j = 1, \ldots, n.
\]  

- Augmented Lagrangian function:

\[
L_\gamma(x_1, \ldots, x_n; \lambda) = f(x) - \langle \lambda, \sum_i A_i x_i - b \rangle + \frac{\gamma}{2} \| \sum_i A_i x_i - b \|^2.
\]

- **Multi-block ADMM** (primal CD, dual ascent)

\[
\begin{aligned}
x_1 & \leftarrow \arg \min_{x_1 \in \mathcal{X}_1} L_\gamma(x_1, \ldots, x_n; \lambda), \\
& \vdots \\
x_n & \leftarrow \arg \min_{x_n \in \mathcal{X}_n} L_\gamma(x_1, \ldots, x_n; \lambda), \\
\lambda & \leftarrow \lambda - \gamma(Ax - b),
\end{aligned}
\]
Divergence of 3-block ADMM


• 3-block ADMM may **diverge** [Chen-He-Ye-Yuan-13].

• Example: solve $3 \times 3$ linear system

\[
\begin{align*}
\min_{x_1, x_2, x_3} & \quad 0, \\
\text{s.t.} & \quad \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0, 
\end{align*}
\]
Random Permutation Helps

- **RP-ADMM**: Randomly permute update order (312), (123), (213), ...

- **New outlet?**
Outline

Motivation

Background

Convergence Analysis of RP-ADMM
  Main Results
  Proof Sketch

Variants of ADMM

Convergence Rate: Related Result and Discussion
Background
ADMM usually refers to 2-block ADMM

\[ \min_{x,y} f(x) + g(y) \]
\[ \text{s.t. } Ax + By = c. \]
Two-block ADMM

• ADMM usually refers to 2-block ADMM
  [Glowinski-Marroco-75], [Gabay-Mercier-76],
  [Boyd-Parikh-Chu-Peleato-Eckstein-11] (5800 citations)

\[
\min_{x, y} f(x) + g(y) \\
\text{s.t. } Ax + By = c.
\]  \hspace{1cm} (4)

• Augmented Lagrangian function:
  \[
  L(x, y; \lambda) = f(x) + g(y) - \langle \lambda, Ax + By - c \rangle + \frac{\gamma}{2} \|Ax + By - c\|^2.
  \]

• Two-block ADMM:
  \[
  \begin{cases}
  x \leftarrow \arg \min_x L(x, y; \lambda), \\
  y \leftarrow \arg \min_y L(x, y; \lambda), \\
  \lambda \leftarrow \lambda - \gamma (Ax + By - c).
  \end{cases}
  \]  \hspace{1cm} (5)
Variants of multi-block ADMM

- Multi-block cyclic ADMM may diverge

- **Question**: How to make multi-block ADMM converge?

- Approach 1: Change algorithm.
  - **Gaussian substitution** [He-Tao-Yuan-11].

  - Strong convexity + small stepsize $\gamma = O(\sigma/N)$ [Han-Yuan-12].

- And many other related works [Deng-Lai-Peng-Yin-13], [Lin-Ma-Zhang-14], [Lin-Ma-Zhang-15], [Sun-Toh-Yang-14], [Li-Sun-Toh-15], etc.

- What is a minimal modification + stepsize 1?
Variants of multi-block ADMM

● Multi-block cyclic ADMM may diverge

● **Question**: How to make multi-block ADMM converge?

● **Approach 1**: Change algorithm.
  ● Gaussian substitution [He-Tao-Yuan-11].

● **Approach 2**: Change algorithm + problem.
  ● Strong convexity + small stepsize $\gamma = O(\sigma/N)$ [Han-Yuan-12].

● **And many other related works** [Deng-Lai-Peng-Yin-13], [Lin-Ma-Zhang-14], [Lin-Ma-Zhang-15], [Sun-Toh-Yang-14], [Li-Sun-Toh-15], etc.
Variants of multi-block ADMM

- Multi-block cyclic ADMM may diverge

- **Question**: How to make multi-block ADMM converge?

- **Approach 1**: Change algorithm.
  - Gaussian substitution [He-Tao-Yuan-11].

- **Approach 2**: Change algorithm + problem.
  - Strong convexity + small stepsize $\gamma = O(\sigma/N)$ [Han-Yuan-12].

- And many other related works [Deng-Lai-Peng-Yin-13], [Lin-Ma-Zhang-14], [Lin-Ma-Zhang-15], [Sun-Toh-Yang-14], [Li-Sun-Toh-15], etc.

- What is a minimal modification + stepsize 1?
We know:
1) ADMM may diverge;
2) Randomization helps CD/SGD [Strohmer-Vershynin-08], [Leventhal-Lewis-10], [Nesterov-11], [Roux et al-12], [Blatt et al-07]

First idea: (independently) randomized ADMM
$$(x_3 x_1 x_1 \lambda), (x_1 x_3 x_2 \lambda), \ldots$$

Bad news: can diverge!
- Diverge for Gaussian data
- Converge for the counter-example in [Chen-He-Ye-Yuan-13]

Second idea: random permutation
$$(x_3 x_1 x_2 \lambda), (x_2 x_1 x_3 \lambda), \ldots$$
It always converges in the simulation.
Summarize ADMM Variants

- **Cyclic**: \((x_1 x_2 x_3 \lambda), (x_1 x_2 x_3 \lambda), \ldots\)
- **Random permutation** (RP): \((x_3 x_1 x_2 \lambda), (x_2 x_1 x_3 \lambda), \ldots\)
- **Independently random** (IR): \((x_3 x_1 x_1 \lambda), (x_2 x_1 x_2 \lambda), \ldots\)

Simulation: RP always converges, other two can diverge. RP > IR, Cyclic.

Wait... practitioners may not care? (divergence of cyclic ADMM is just worst-case?)
Summarize ADMM Variants

- **Cyclic**: \((x_1 x_2 x_3 \lambda), (x_1 x_2 x_3 \lambda), \ldots\)

- **Random permutation (RP)**: \((x_3 x_1 x_2 \lambda), (x_2 x_1 x_3 \lambda), \ldots\)

- **Independently random (IR)**: \((x_3 x_1 x_1 \lambda), (x_2 x_1 x_2 \lambda), \ldots\)

- Simulation: **RP always converges**, other two can diverge.
  
  \(\text{RP} \succ \text{IR, Cyclic.}\)

- **Wait...practitioners may not care?** (divergence of cyclic ADMM is just worst-case?)
Numerical Experiments: Cyc-ADMM Often Diverges

Table 1: Solve Linear Systems by Cyc-ADMM, RP-ADMM and GD

<table>
<thead>
<tr>
<th>N</th>
<th>Diverg. Ratio for Cyc-ADMM</th>
<th>Iterations for $\epsilon = 0.001$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CycADMM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$i$</td>
</tr>
<tr>
<td>Gaussian $N(0, 1)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.7%</td>
<td>3.2e01</td>
</tr>
<tr>
<td>100</td>
<td>3%</td>
<td>1.0e03</td>
</tr>
<tr>
<td>Uniform $[0, 1]$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Numerical Experiments: Cyc-ADMM Often Diverges

Table 1: Solve Linear Systems by Cyc-ADMM, RP-ADMM and GD

<table>
<thead>
<tr>
<th>N</th>
<th>Diverg. Ratio for Cyc-ADMM</th>
<th>Iterations for $\epsilon = 0.001$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CycADMM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gaussian $N(0, 1)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.7%</td>
<td>3.2e01</td>
</tr>
<tr>
<td>100</td>
<td>3%</td>
<td>1.0e03</td>
</tr>
<tr>
<td>Uniform $[0, 1]$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3.2%</td>
<td>7.0e01</td>
</tr>
<tr>
<td>100</td>
<td>100%</td>
<td>N/A</td>
</tr>
</tbody>
</table>

- Cyc-ADMM can diverge often; sometimes diverges w.p. 100%.
- In fact, easy to diverge if off-diagonal entries are large.
  Cyc-ADMM is somewhat similar to Cyc-BCD.
- RP-ADMM converges faster than GD.
Cyclic ADMM may diverge: a “robust” claim.

- Not worst-case example; happen often.
Remarks on Divergence of Cyclic ADMM

- Cyclic ADMM may diverge: a “robust” claim.
  - Not worst-case example; happen often.
  - Stepsize does not help (at least constant).
  - Strong convexity does not help (at least for stepsize 1).
Remarks on Divergence of Cyclic ADMM

- Cyclic ADMM may diverge: a “robust” claim.
  - Not worst-case example; happen often.
  - Stepsize does not help (at least constant).
  - Strong convexity does not help (at least for stepsize 1).

- Order (123) fails; maybe (231) works?
Remarks on Divergence of Cyclic ADMM

- **Cyclic ADMM may diverge**: a “robust” claim.
  - Not worst-case example; happen often.
  - **Stepsize** does not help (at least constant).
  - **Strong convexity** does not help (at least for stepsize 1).

- Order (123) fails; maybe (231) works?

- **Fact**: Any fixed order diverges.
Summary: Why We Want to Understand RP-ADMM

Theoretical Curiosity + Practical Need.

- **First**, decomposition idea can be useful for solving constrained problems
  - cyclic ADMM may not converge.
  - **RP-ADMM**: a simple solution
Summary: Why We Want to Understand RP-ADMM

Theoretical Curiosity + Practical Need.

- **First**, decomposition idea can be useful for solving constrained problems
  - cyclic ADMM may not converge.
  - **RP-ADMM**: a simple solution

- **Second**, help understand RP-rule, e.g. RP-CD, RP-SGD.
Theoretical Curiosity + Practical Need.

- **First**, decomposition idea can be useful for solving constrained problems
  - cyclic ADMM may not converge.
  - **RP-ADMM**: a simple solution

- **Second**, help understand **RP-rule**, e.g. RP-CD, RP-SGD.
  - Many people write IR papers.
  - Many people run RP experiments (default choice in deep learning package e.g. Torch)
Convergence Analysis of RP-ADMM
Solve Linear System

- Solve a square linear system of equations \((f_i = 0, \ \forall i)\).

\[
\begin{aligned}
\min_{x \in \mathbb{R}^N} & \quad 0, \\
\text{s.t.} & \quad A_1 x_1 + \cdots + A_n x_n = b,
\end{aligned}
\]  

(6)

where \(A = [A_1, \ldots, A_n] \in \mathbb{R}^{N \times N}\) is full-rank, \(x_i \in \mathbb{R}^{d_i}\) and \(\sum_i d_i = N\).
Solve Linear System

• Solve a square linear system of equations \((f_i = 0, \forall i)\).

\[
\begin{align*}
\min_{x \in \mathbb{R}^N} & \quad 0, \\
\text{s.t.} & \quad A_1 x_1 + \cdots + A_n x_n = b,
\end{align*}
\]

(6)

where \(A = [A_1, \ldots, A_n] \in \mathbb{R}^{N \times N}\) is full-rank, \(x_i \in \mathbb{R}^{d_i}\) and \(\sum_i d_i = N\).

• Why linear system?
  
  • Basic constrained problem
  
  • Already difficult to analyze.
Main results

**Theorem 1**

The expected output of RP-ADMM converges to the solution of (6), i.e.

\[
\{ E_{\xi_k} (y^k) \}_{k \to \infty} \longrightarrow y^*.
\]  

(7)

**Remark:** Expected convergence \(\neq\) convergence, but is a strong evidence for convergence.

Denote \(M\) as the expected iteration matrix of RP-ADMM.

**Theorem 2**

\(\rho(M) < 1\), i.e. spectral radius of \(M\) is less than 1.

\[\text{iiS, Luo, Yinyu Ye, “On the Expected Convergence of Randomly Permutated ADMM”,}\]
Why Spectral Analysis?

Meta-proof-frameworks in optimization don’t work (or I don’t know how).

Potential function.

- E.g. GD, C-CD or R-CD for $\min_x x^T Ax$, the potential function is the (expected) objective.

- Our system: $E(y^{k+1}) = ME(y^k)$, but $\|M\| > 2.3$ for the counterexample. $y^T My$ is not a potential function.

- There exists $P$ such that $P - M^T PM$ is PSD, and $y^T Py$ is a potential function. Hard to compute $P$.

Contraction: can prove convergence of 2-block ADMM.

- Again, how to distinguish between cyclic ADMM and PR-ADMM?

- Not a big surprise. 2-block is very special.
RP-ADMM can be viewed as switched linear systems:

\[ y_{k+1} = M_k y_k, \]

where \( M_k \in \{B_1, \ldots, B_m\} \). For RP-ADMM, \( m = n! \).

**Our problem:** each single \( B_i \) is not stable (corresponding to a single order), but randomly picking from \( \{B_1, \ldots, B_m\} \) makes the system stable.

Related to product of random matrices [Furstenberg-Kesten-60]; but hard to apply to our case.

A useful first step is to find a convex combination of \( B_i \)'s that is stable [Wicks et al.-94]
Theorem 2: a Pure Linear Algebra Problem

- Define matrix $L_\sigma$ by **deleting half off-diagonal entries** of $A^T A$

$$
L_\sigma[\sigma(i), \sigma(j)] \triangleq \begin{cases} 
A_{\sigma(i)}^T A_{\sigma(j)} & j \leq i, \\
0 & j > i,
\end{cases}
$$

(8)

- Example:

$$
L_{(231)} = \begin{bmatrix}
1 & A_1^T A_2 & A_1^T A_3 \\
0 & 1 & 0 \\
0 & A_3^T A_2 & 1
\end{bmatrix}.
$$
Theorem 2: a Pure Linear Algebra Problem

• Define matrix $L_{\sigma}$ by deleting half off-diagonal entries of $A^T A$

$$L_{\sigma}[\sigma(i), \sigma(j)] \triangleq \begin{cases} A_{\sigma(i)}^T A_{\sigma(j)} & j \leq i, \\ 0 & j > i, \end{cases}$$

(8)

• Example:

$$L_{(231)} = \begin{bmatrix} 1 & A_1^T A_2 & A_1^T A_3 \\ 0 & 1 & 0 \\ 0 & A_3^T A_2 & 1 \end{bmatrix}.$$ 

• Define $Q = E(L_{\sigma}^{-1})$. Compare: $E(L_{\sigma}) = \frac{1}{2}(I + A^T A)$. 

19
Theorem 2: a Pure Linear Algebra Problem

- Define matrix $L_\sigma$ by deleting half off-diagonal entries of $A^T A$

\[
L_\sigma[\sigma(i), \sigma(j)] \triangleq \begin{cases} 
A_{\sigma(i)}^T A_{\sigma(j)} & j \leq i, \\
0 & j > i,
\end{cases}
\]  

Example:

\[
L_{(231)} = \begin{bmatrix}
1 & A_1^T A_2 & A_1^T A_3 \\
0 & 1 & 0 \\
0 & A_3^T A_2 & 1
\end{bmatrix}.
\]

- Define $Q = E(L_\sigma^{-1})$. Compare: $E(L_\sigma) = \frac{1}{2}(I + A^T A)$.

- Theorem 2 claims $\rho(M) < 1$, with $M$ being a function of $A$:

\[
M = \begin{bmatrix}
I - QA^T A & QA^T \\
-A + AQAT & I - AQAT
\end{bmatrix}.
\]
Two Main Lemmas to Prove Theorem 2: Lemma 1

- **Step 1**: Relate $M$ to a symmetric matrix $AQA^T$.

**Lemma 1**

$$\lambda \in \text{eig}(M) \iff \frac{(1 - \lambda)^2}{1 - 2\lambda} \in \text{eig}(AQA^T). \quad (10)$$

When $Q$ is symmetric, we have

$$\rho(M) < 1 \iff \text{eig}(AQA^T) \subseteq (0, \frac{4}{3}). \quad (11)$$

- This lemma treats $Q$ as a black box.
Lemma 2

- **Step 2**: Bound eigenvalues of $AQA^T$.

Lemma 2

For any non-singular $A$, let $Q = E(L_\sigma^{-1})$ where $L_\sigma$ is given by (8), then

$$\text{eig}(AQA^T) \subseteq (0, \frac{4}{3}).$$

(12)

- **Remark**: $4/3$ should be **tight**: we find examples $> 1.33$. 
What is $AQA^T$

- $AQA^T$ relates to RP-CD (quadratic): $x \leftarrow (I - QA^T A)x$.
  - RP-CD converges $\iff$ $\text{eig}(AQA^T) \in (0, 2)$.

- Remark: spectrum of RP-CD is "nicer" than Cyc-CD.
- "Pre-assigned" space for RP-ADMM.
What is $AQA^T$

- $AQA^T$ relates to RP-CD (quadratic): $x \leftarrow (I - QA^TA)x$.

- RP-ADMM converges $\iff$ eig$(AQA^T) \in (0, 4/3)$.

Remark: spectrum of RP-CD is "nicer" than Cyc-CD.

"Pre-assigned" space for RP-ADMM.
What is $AQA^T$

- $AQA^T$ relates to RP-CD (quadratic): $x \leftarrow (I - QA^T A)x$.
  - RP-ADMM converges $\iff$ $\text{eig}(AQA^T) \in (0, 4/3)$.
- Cyc-CD (quadratic): $x \leftarrow (I - L_{12\ldots n}^{-1} A^T A)x$
  - Cyc-CD converges $\iff$ $\text{eig}(AL_{12\ldots n}^{-1} A^T) \in (0, 2)$.

Remark: spectrum of RP-CD is "nicer" than Cyc-CD.

"Pre-assigned" space for RP-ADMM.
What is $AQAT$?

- $AQAT$ relates to RP-CD (quadratic): $x \leftarrow (I - QA^T A)x$.
  - RP-ADMM converges $\iff$ eig($AQAT$) $\in (0, 4/3)$.

- Cyc-CD (quadratic): $x \leftarrow (I - \sum_{1, 2, \ldots, n} A^T A)x$
  - Cyc-CD converges $\iff$ eig($AL^{-1}_{12 \ldots n} A^T$) $\in (0, 2)$.

- **Remark**: spectrum of RP-CD is “nicer” than Cyc-CD.
  - “Pre-assigned” space for RP-ADMM.
Proof Sketch of Lemma 2

- **Step 2.1**: Symmetrization $\implies$ induction formula of $Q = E(L_{\sigma}^{-1})$.

- **Step 2.2**: Induction inequality of $\rho = \rho(QA^T A)$:

\[
\rho \leq P(\hat{\rho}, \rho) \triangleq \max_{\theta \geq 0} \hat{\rho} + \theta \left( \frac{\rho}{4\rho - 4 + \theta} - 1 \right),
\]

where $\hat{\rho}$ is the $(n - 1)$-block analog of $\rho(QA^T A)$. 

Remark: $\rho = 4/3$ is the fixed point of $\rho = P(\rho, \rho)$. 

$P(4/3, 4/3) = 4/3 + \max_{\theta \geq 0} \theta \left( \frac{4/3}{4(4/3) - 4 + \theta} - 1 \right) = 4/3 - \max_{\theta \geq 0} \theta 2/3 = 4/3$. 

23
Proof Sketch of Lemma 2

• **Step 2.1**: Symmetrization \(\implies\) induction formula of 
  \(Q = E(L_{\sigma}^{-1}).\)

• **Step 2.2**: Induction inequality of \(\rho = \rho(QA^T A):\)

\[
\rho \leq P(\hat{\rho}, \rho) \triangleq \max_{\theta \geq 0} \hat{\rho} + \theta \left( \frac{\rho}{4\rho - 4 + \theta} - 1 \right), \quad (13)
\]

where \(\hat{\rho}\) is the \((n - 1)\)-block analog of \(\rho(QA^T A).\)

• **Remark**: \(\rho = 4/3\) is the fixed point of \(\rho = P(\rho, \rho).\)
  
  • \(P(\frac{4}{3}, \frac{4}{3}) = \frac{4}{3} + \max_{\theta \geq 0} \theta(\frac{\rho}{\rho + \theta} - 1) = \frac{4}{3} - \max_{\theta \geq 0} \frac{\theta^2}{\rho + \theta} = \frac{4}{3}.\)
Variants of ADMM
Interesting Byproduct: New Randomization Rule

- Finding: 2-level symmetrization is enough.

- New algorithm: Bernolli randomization (BR).
  - Phase 1: sweep $1, \ldots, n$; for each block, update w.p. $1/2$;
  - Phase 2: sweep $n, \ldots, 1$; if previously not updated, now update.

- Examples of valid order: (2, 3; 4, 1), (1, 2, 4; 3).
  Non-examples: (3, 4, 1, 2)
Finding: 2-level symmetrization is enough.

New algorithm: Bernolli randomization (BR).

- Phase 1: sweep $1, \ldots, n$; for each block, update w.p. $1/2$;
- Phase 2: sweep $n, \ldots, 1$; if previously not updated, now update.

Examples of valid order: $(2, 3; 4, 1)$, $(1, 2, 4; 3)$. Non-examples: $(3, 4, 1, 2)$

Proposition: BR-ADMM converges in expectation.
The problem is still $\min_x f(x)$, s.t. $Ax = b$.

Original ADMM: each cycle is $(x_1, x_2, x_3, \lambda)$.

Primal-dual ADMM: each cycle is $(x_1, \lambda, x_2, \lambda, x_3, \lambda)$.
- Cyclic version still can diverge for the counter-example.
- Randomized version was proven to converge with high probability (e.g. [Xu-2017])

However, in simulation, randomized PD-ADMM is much slower than other versions (next page).
Comparison of Algorithms

Uniform [0,1] data:

- cyclic ADMM and primal-dual version of Bernolli randomization fail to converge.
- PD-rand-ADMM is much slower than others.
Standard Gaussian data:

- PD-rand-ADMM is significantly slower than all other methods.
- Recall: randomized ADMM is the only method that diverges!

Strange issue: (independent) random rule is bad for Gaussian data.
Comparison of Algorithms (cont’d)

Simple summary of different methods (no stepsize tuning):

<table>
<thead>
<tr>
<th>Update Order</th>
<th>Original Version</th>
<th>Primal-Dual Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>cyclic</td>
<td>Diverge</td>
<td>Diverge</td>
</tr>
<tr>
<td>indep. random</td>
<td>Diverge</td>
<td>Converge but very slow</td>
</tr>
<tr>
<td>Bernolli random</td>
<td>Converge</td>
<td>Diverge</td>
</tr>
<tr>
<td>random permutation</td>
<td>Convergeiii</td>
<td>Converge?iv</td>
</tr>
</tbody>
</table>

**Observation:** random permutation is a universal “stabilizer”.

**Open question:** Any convergence analysis of P-D version of RP-ADMM?

---

iii Only expected convergence for simple problems are proved.
iv Based on extensive simulation
Convergence Rate: Related Result and Discussion
Convergence Rate of Cyclic CD

- Status: many results on (independently) random rule; little understanding of RP/cyclic/whatever rule
  - A few works [Recht-Re-12], [Gurbuzbalaban-Ozdaglar-Parrilo-15], [Wright-Lee-17] studied random permutation, but why RP is better than IR in general is still unknown
  - Mark Schmidt talked about Gauss-Southwell rule this morning.

- Classical literature says: they are “essentially cyclic” rule, all converge for CD

- However, their convergence speed can be quite different
Convergence Rate of Cyclic CD

- **Question**: “true” convergence rate of cyclic CD or Gauss-Seidal method (Gauss 1823, Seidel 1874)?

- Why care cyclic order?
  - Understanding “non-independently-randomized” rule
  - Almost all convergence rate results on cyclic rule immediately apply to RP-rule
  - Randomization not available sometimes

- **Puzzle**: known rates can be sometimes $n^2$ times worse than R-CD for quadratic case [Beck-Tetruashvili-13], [Sun-Hong-15]

- Some claim cyclic order must be bad; an example given by Strohmer and Richtarik (independently) showed this.
  - Only $O(n)$ gap between C-CD and R-CD;
  - **Only fails for some particular orders.** Randomly pick order and fix, then becomes fast.
Rate of Cyclic CD

- **Answer**: up to $n^2$ times worse than R-CD, for equal-diagonal quadratic case.

**Table 2**: Complexity for equal-diagonal case (divided by $n^2 \kappa \log \frac{1}{\epsilon}$ and ignoring constants. $\tau = \lambda_{\text{max}} / \lambda_{\text{avg}} \in [1, n]$ )

<table>
<thead>
<tr>
<th></th>
<th>C-CD</th>
<th>GD</th>
<th>R-CD</th>
<th>SVRG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower bound</td>
<td>$\tau$</td>
<td>1</td>
<td>$\frac{1}{\tau}$</td>
<td>$\frac{1}{\tau}$</td>
</tr>
<tr>
<td>Upper bound</td>
<td>$\min { \tau \log^2 n, n }$</td>
<td>1</td>
<td>$\frac{1}{\tau}$</td>
<td>$\frac{1}{\tau}$</td>
</tr>
</tbody>
</table>

- Lower bound is based on analyzing one example. Steven Wright mentioned the example in two talks starting from 2015 summer. We independently discover the example.

- Analysis: tricky issue on non-symmetric matrix update. (even more tricky than ADMM case)

---

Sun, Ye, “Worst-case Convergence Rate of Cyclic Coordinate Descent Method: $O(n^2)$ Gap with Randomized Versions”, 2016.
• Same gap exists for Kaczmarz method and POCS (Projection onto Convex Sets).

• POCS, dating back to Von Neumann in 1930’s, has been studied extensively. See a survey [Bauschke-Borwein-Lewis-1997]

• Convergence rate given by Smith, Solmon and Wagner in 1977. Still in textbook.

• Translate to CD: a rate dependent on all eigenvalues.
  • Turn out to be $\infty$-times worse than our bound for the example.
  • Always worse than our bound (up to $O(\log^2 n)$ factor)
Random permutation was studied in [Recht-Re’2012], mainly for RP-SGD.

**Conjecture: Matrix AM-GM inequality ([Recht-Re’2012])**

Suppose $A_1, \ldots, A_n \succeq 0$, then

$$\left\| \frac{1}{n!} \sum_{\sigma \text{ is a permutation}} A_{\sigma_1} \cdots A_{\sigma_n} \right\| \leq \left\| \frac{1}{n} (A_1 + \cdots + A_n) \right\|^n.$$  

If this inequality holds, then the convergence rate of RP-CD for quadratic problems is faster than R-CD.

Zhang gave a proof for $n = 3$; Duchi gave a proof for a variant, again for $n = 3$. 


Another variant of matrix AM-GM inequality

**Conjecture** (variant of matrix AM-GM inequality): If $P_i$ is a projection matrix, $i = 1, \ldots, n$, then

$$
\frac{1}{n!} \sum_{\sigma \text{ is a permutation}} P_{\sigma_1} \cdots P_{\sigma_n} \preceq \frac{1}{n}(P_1 + \cdots + P_n). \tag{14}
$$

**Claim:** If matrix AM-GM inequality (14) holds, then combining with our result $\text{eig}(QA^T A) \in (0, 4/3)$, RP-CD has better convergence rate than that of R-CD for convex quadratic problems.

We know $\text{eig}(I - QA^T A) = \text{eig}(M_{RP-CD}) \in (-1, 1)$.

- Our result is about the left end by improving $-1$ to $-1/3$.
- Matrix AM-GM inequality (14) is about the right end near 1

We have some results on the expected convergence rate of RP-CD and RP-ADMM. Skip here.
Summary

- **Main result**: convergence analysis of RP-ADMM.

  - **Implication 1** (problem): solver for constrained problems.

  - **Implication 2** (algorithm): RP better.
  
    Even much better than independently randomized rule.

- Implication for RP-CD: “truncate” one side spectrum.

- Tight analysis of “non-independent-randomization”: worst-case understanding of cyclic order, but more works are needed.

- Lots of open questions:
  - convergence of PD version of RP-ADMM
  - AM-GM inequality
  - Jacobi preconditioning
Thank You!