A Matrix Splitting Method for Composite Function Minimization

Ganzhao Yuan[1,2], Wei-Shi Zheng[2], Bernard Ghanem[1]

1. King Abdullah University of Science and Technology (KAUST), Saudi Arabia
2. School of Data and Computer Science, Sun Yat-sen University (SYSU), China

yuanganzhao@gmail.com, wszheng@ieee.org, bernard.ghanem@kaust.edu.sa

Outline

- Introduction
- Proposed Matrix Splitting Method
  - Convex Quadratic Vector Case
- Extensions
  - Non-Convex Case
  - Matrix Case
  - Non-Quadratic Case
- Experiments
- Future Work

Outline

- Introduction
- Proposed Matrix Splitting Method
  - Convex Quadratic Vector Case
- Extensions
  - Non-Convex Case
  - Matrix Case
  - Non-Quadratic Case
- Experiments
- Future Work

Introduction

- Composite Function Optimization Problem

$$\min_{\mathbf{x}} \; f(\mathbf{x}) \triangleq \underbrace{\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{x}^T\mathbf{b}}_{q(\mathbf{x})} + h(\mathbf{x})$$

- Assumption
  - $\mathbf{A}$ is symmetric positive semi-definite
  - $h(\cdot)$ is separable

Convex $h(\cdot)$

$$h(\mathbf{x}) = \|\mathbf{x}\|_1$$

$$h(\mathbf{x}) = \begin{cases} \mathbf{0}, & \mathbf{x} \geq \mathbf{0}; \\ \infty, & \text{else.} \end{cases}$$

Nonconvex $h(\cdot)$

$$h(\mathbf{x}) = \|\mathbf{x}\|_0$$

$$h(x) = \begin{cases} 0, & \mathbf{x} \in \{0,1\}^n; \\ \infty, & \text{else.} \end{cases}$$

Introduction

- Optimization Problem

$$\min_{\mathbf{x}} \;\; f(\mathbf{x}) \triangleq \underbrace{\tfrac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{x}^T\mathbf{b}}_{q(\mathbf{x})} + h(\mathbf{x})$$

- Existing Solution: Proximal Gradient Method

$$\mathbf{x}^{k+1} \Leftarrow \min_{x} \;\; g(x, x^k) + h(x)$$

$$\forall \mathbf{z}, \mathbf{x}, \;\; q(\mathbf{x}) \leq \underbrace{q(\mathbf{z}) + \langle \nabla q(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle + \tfrac{L}{2}\|\mathbf{x} - \mathbf{z}\|_2^2}_{g(\mathbf{x},\mathbf{z})}$$

$$\mathbf{x}^{k+1} = \mathrm{prox}_{\gamma h}(\mathbf{x}^k - \gamma \nabla q(\mathbf{x}^k))$$

$$\mathrm{prox}_{\tilde{h}}(\mathbf{a}) = \arg\min_{\mathbf{x}} \;\; \tfrac{1}{2}\|\mathbf{x} - \mathbf{a}\|_2^2 + \tilde{h}(\mathbf{x})$$

$$= (\mathbf{I} + \partial \tilde{h})^{-1}(\mathbf{a})$$

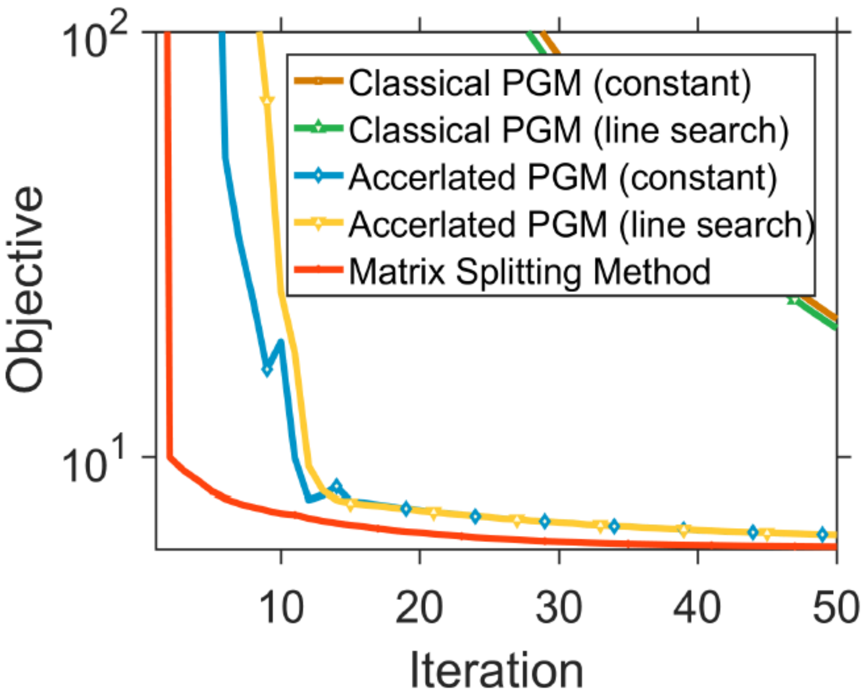
Introduction: Motivation

$$\mathrm{prox}_{\tilde{h}}(\mathbf{a}) = \arg\min_{\mathbf{x}} \ \frac{1}{2}\|\mathbf{x}-\mathbf{a}\|_{\mathbf{B}}^2 + \tilde{h}(\mathbf{x})$$

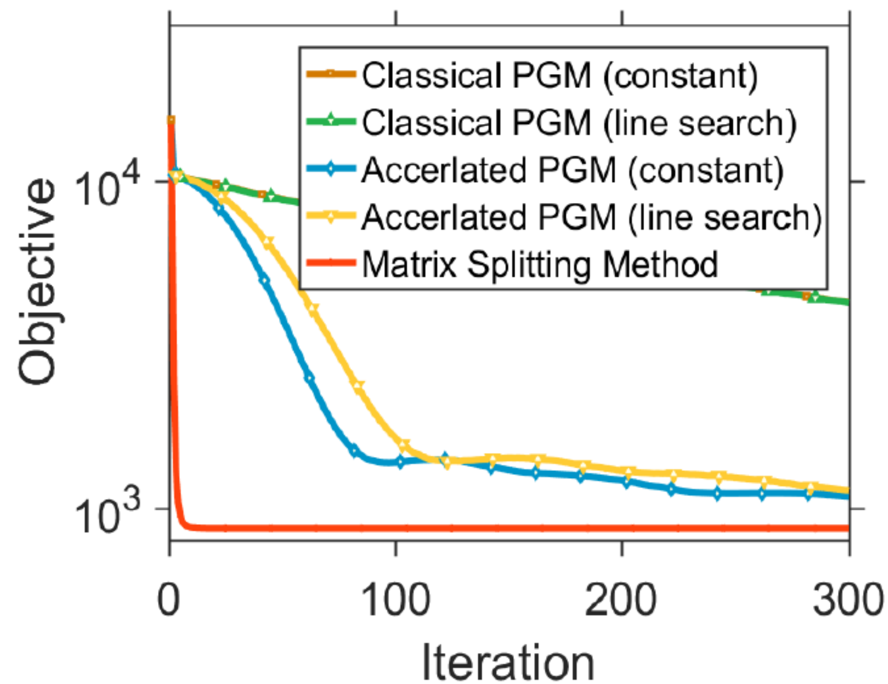
Existing Method	New Method
$\mathbf{B}$ = Scaled Identity Matrix	$\mathbf{B}$ = Triangle Matrix
Closed Form Solution	Closed Form Solution
Proximal Operator	Triangle Proximal Operator
$\mathrm{prox}_{\tilde{h}}(\mathbf{a}) = \boxed{(\mathbf{I}+\partial\tilde{h})^{-1}}(\mathbf{a})$	$\mathrm{prox}_{\tilde{h}}(\mathbf{a}) = (\mathbf{B}+\partial\tilde{h})^{-1}(\mathbf{a})$
resolvent of $\tilde{h}$	triangle resolvent of $\tilde{h}$?

Introduction

- A Toy Problem



(a) $\min_{\mathbf{x}\geq\mathbf{0}} \ \frac{1}{2}\|\mathbf{C}\mathbf{x}-\mathbf{d}\|_2^2$



(b) $\min_{\mathbf{x}} \ \frac{1}{2}\|\mathbf{C}\mathbf{x}-\mathbf{d}\|_2^2 + \|\mathbf{x}\|_0$

Our matrix splitting method significantly outperforms existing popular proximal gradient methods in term of both efficiency and efficacy.

Outline

- Introduction
- Proposed Matrix Splitting Method
  - Convex Quadratic Vector Case
- Extensions
  - Non-Convex Case
  - Matrix Case
  - Non-Quadratic Case
- Experiments
- Future Work

Proposed Matrix Splitting Method

$$\min_{\mathbf{x}} \;\; f(\mathbf{x}) \triangleq \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{x}^T\mathbf{b} + h(\mathbf{x})$$

$$\begin{aligned}\mathbf{A} &\triangleq \mathbf{L} + \mathbf{D} + \mathbf{L}^T \\ &\triangleq \underbrace{\mathbf{L} + \tfrac{1}{\omega}(\mathbf{D} + \theta\mathbf{I})}_{\mathbf{B}} + \underbrace{\mathbf{L}^T + \tfrac{1}{\omega}((\omega - 1)\mathbf{D} - \theta\mathbf{I})}_{\mathbf{C}}\end{aligned} \quad \mathbf{D} = \begin{bmatrix} \mathbf{A}_{1,1} & 0 & 0 \\ 0 & \mathbf{A}_{2,2} & 0 \\ 0 & 0 & \mathbf{A}_{3,3} \end{bmatrix}, \; \mathbf{L} = \begin{bmatrix} 0 & 0 & 0 \\ \mathbf{A}_{2,1} & 0 & 0 \\ \mathbf{A}_{3,1} & \mathbf{A}_{3,2} & 0 \end{bmatrix}$$

- Optimality Condition → Fixed-Point $\mathbf{x} = \mathcal{T}(\mathbf{x})$

$$\mathbf{0} \in (\mathbf{B} + \mathbf{C})\mathbf{x} + \mathbf{b} + \partial h(\mathbf{x})$$
$$-\mathbf{C}\mathbf{x} - \mathbf{b} \in (\mathbf{B} + \partial h)\mathbf{x}$$
$$\mathbf{x} \in -(\mathbf{B} + \partial h)^{-1}(\mathbf{C}\mathbf{x} + \mathbf{b})$$

- Fixed-Point Iterative Scheme

$$\mathbf{x}^{k+1} = \mathcal{T}(\mathbf{x}^k) \triangleq (\mathbf{B} + \partial h)^{-1}(-\mathbf{C}\mathbf{x}^k - \mathbf{b})$$

Proposed Matrix Splitting Method

- How to compute operator $\mathcal{T}(\mathbf{x}^k)$?

$$\text{find } \mathbf{z}^* \text{ that: } \mathbf{0} \in \mathbf{B}\mathbf{z}^* + \mathbf{u} + \partial h(\mathbf{z}^*), \text{ where } \mathbf{u} = \mathbf{b} + \mathbf{C}\mathbf{x}^k$$

- Using forward substitution !

$$\mathbf{0} \in \begin{bmatrix} \mathbf{B}_{1,1} & 0 & 0 & 0 & 0 \\ \mathbf{B}_{2,1} & \mathbf{B}_{2,2} & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & 0 & 0 \\ \mathbf{B}_{n-1,1} & \mathbf{B}_{n-1,2} & \cdots & \mathbf{B}_{n-1,n-1} & 0 \\ \mathbf{B}_{n,1} & \mathbf{B}_{n,2} & \cdots & \mathbf{B}_{n,n-1} & \mathbf{B}_{n,n} \end{bmatrix} \begin{bmatrix} \mathbf{z}_1^* \\ \mathbf{z}_2^* \\ \vdots \\ \mathbf{z}_{n-1}^* \\ \mathbf{z}_n^* \end{bmatrix} + \mathbf{u} + \partial h(\mathbf{z}^*)$$

- It reduces to 1-dimensional sub-problem

$$0 \in \mathbf{B}_{j,j}\mathbf{z}_j^* + \mathbf{w}_j + \partial h(\mathbf{z}_j^*), \text{ where } \mathbf{w}_j = \mathbf{u}_j + \textstyle\sum_{i=1}^{j-1} \mathbf{B}_{j,i}\mathbf{z}_i^*$$

$$\mathbf{z}_j^* = t^* \triangleq \arg\min_t \ \tfrac{1}{2}\mathbf{B}_{j,j}t^2 + \mathbf{w}_j t + h(t)$$

Proposed Matrix Splitting Method

- Condition $\delta \triangleq \frac{2\theta}{\omega} + \frac{2-\omega}{\omega} \min(diag(\mathbf{D})) > 0$
  - Simple Choice: $\omega \in (0, 2), \ \theta = 0.01$
- Convergence Results
  - Monotone Non-increasing and Convergent

$$f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k) \leq -\frac{\delta}{2} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2$$

  - Q-linear Convergence Rate

$$\frac{f(\mathbf{x}^{k+1}) - f(\mathbf{x}^*)}{f(\mathbf{x}^k) - f(\mathbf{x}^*)} \leq \frac{C_1}{1+C_1}$$

  - Iteration Complexity

$$f(\mathbf{x}^k) - f(\mathbf{x}^*) \leq \begin{cases} u^0 \left(\frac{2C_4}{2C_4+1}\right)^k, & \text{if } \sqrt{f^k - f^{k+1}} \geq C_3/C_4, \ \forall k \leq \bar{k} \\ \frac{C_5}{k}, & \text{if } \sqrt{f^k - f^{k+1}} < C_3/C_4, \ \forall k \geq 0 \end{cases}$$

Outline

- Introduction
- Proposed Matrix Splitting Method
  - Convex Quadratic Vector Case
- Extensions
  - Non-Convex Case
  - Matrix Case
  - Non-Quadratic Case
- Experiments
- Future Work

Extensions: Nonconvex Case

$$\min_{\mathbf{x}} \ f(\mathbf{x}) \triangleq \underbrace{\tfrac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{x}^T\mathbf{b}}_{q(\mathbf{x})} + h(\mathbf{x})$$

- Using the same method to compute $\mathcal{T}(\mathbf{x}^k)$
  - It reduces to solve: $t^* \triangleq \arg\min_t \ \frac{1}{2}\mathbf{B}_{j,j}t^2 + \mathbf{w}_j t + h(t)$
- Condition $\delta \triangleq \min\left(\theta/\omega + (1-\omega)/\omega \cdot diag(\mathbf{D})\right) > 0$
  - Simple Choice: $\omega < 1, \ \theta = 0.01$
- Convergence Result
  - Monotonically Nonincreasing and Convergent

$$f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k) \leq -\frac{\delta}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2$$

Extensions: Matrix Case

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times r}} \quad f(\mathbf{X}) \triangleq \underbrace{\tfrac{1}{2}tr(\mathbf{X}^T\mathbf{A}\mathbf{X}) + tr(\mathbf{X}^T\mathbf{R})}_{q(\mathbf{x})} + h(\mathbf{X})$$

- Applications: NMF, Sparse Coding
- Using the same method to decompose **A**

$$\mathbf{A} = \mathbf{B} + \mathbf{C}$$

- Solve the following nonlinear equation w.r.t. **Z***:

$$\mathbf{B}\mathbf{Z}^* + \mathbf{R} + \mathbf{C}\mathbf{X}^k + \partial h(\mathbf{Z}^*) \in \mathbf{0}$$

- It can be decomposed into independent components.

Extensions: Non-quadratic Case

$$\min_{\mathbf{x}} \; f(\mathbf{x}) \triangleq q(\mathbf{x}) + h(\mathbf{x})$$

Majorization Minimization

$$\mathbf{x}^{k+1} \Leftarrow \min_x \; g(x, x^k) + h(x)$$

Quadratic Surrogate (Second Order Upper Bound)

$$q(x) \leq g(x, x^k) \triangleq q(x^k) + \langle \nabla q(x^k), x - x^k \rangle$$
$$+ \tfrac{1}{2}(x - x^k)^T M (x - x^k), \; \text{with } M \succeq \nabla^2 f(x^k)$$

Line Search (as in Damped Newton)

$$\mathbf{x}^{k+1} \Leftarrow \mathbf{x}^k + \beta(\mathbf{x}^{k+1} - \mathbf{x}^k)$$

Outline

- Introduction
- Proposed Matrix Splitting Method
  - Convex Quadratic Vector Case
- Extensions
  - Non-Convex Case
  - Matrix Case
  - Non-Quadratic Case
- Experiments
- Future Work

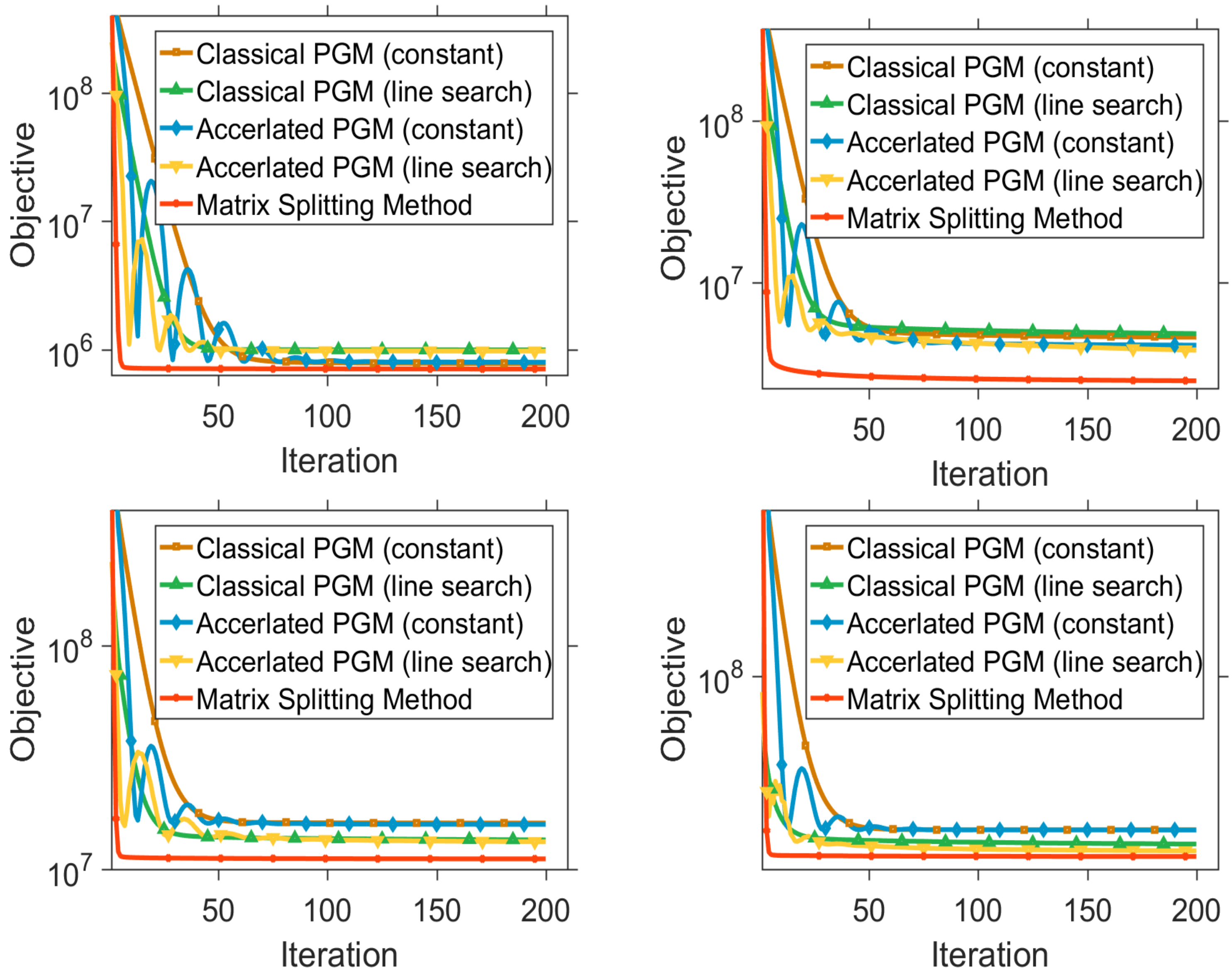
Experiments on NMF

data	n	[17] PG	[14] AS	[14] BPP	[10] APG	[13] CGD	ours MSM
time limit=20							
20news	20	5.001e+06	2.762e+07	8.415e+06	4.528e+06	4.515e+06	4.506e+06
20news	50	5.059e+06	2.762e+07	4.230e+07	3.775e+06	3.544e+06	3.467e+06
20news	100	6.955e+06	5.779e+06	4.453e+07	3.658e+06	3.971e+06	2.902e+06
20news	200	7.675e+06	3.036e+06	1.023e+08	4.431e+06	3.573e+07	2.819e+06
20news	300	1.997e+07	2.762e+07	1.956e+08	4.519e+06	4.621e+07	3.202e+06
COIL	20	2.004e+09	5.480e+09	2.031e+09	1.974e+09	1.976e+09	1.975e+09
COIL	50	1.412e+09	1.516e+10	6.962e+09	1.291e+09	1.256e+09	1.252e+09
COIL	100	2.960e+09	2.834e+10	3.222e+10	9.919e+08	8.745e+08	8.510e+08
COIL	200	3.371e+09	2.834e+10	5.229e+10	8.495e+08	5.959e+08	5.600e+08
COIL	300	3.996e+09	2.834e+10	1.017e+11	8.493e+08	5.002e+08	4.956e+08
TDT2	20	1.597e+06	2.211e+06	1.688e+06	1.591e+06	1.595e+06	1.592e+06
TDT2	50	1.408e+06	2.211e+06	2.895e+06	1.393e+06	1.390e+06	1.385e+06
TDT2	100	1.300e+06	2.211e+06	6.187e+06	1.222e+06	1.224e+06	1.214e+06
TDT2	200	1.628e+06	2.211e+06	1.791e+07	1.119e+06	1.227e+06	1.079e+06
TDT2	300	1.915e+06	1.854e+06	3.412e+07	1.172e+06	7.902e+06	1.066e+06
time limit=30							
20news	20	4.716e+06	2.762e+07	7.471e+06	4.510e+06	4.503e+06	4.500e+06
20news	50	4.569e+06	2.762e+07	5.034e+07	3.628e+06	3.495e+06	3.446e+06
20news	100	6.639e+06	2.762e+07	4.316e+07	3.293e+06	3.223e+06	2.817e+06
20news	200	6.991e+06	2.762e+07	1.015e+08	3.609e+06	7.676e+06	2.507e+06
20news	300	1.354e+07	2.762e+07	1.942e+08	4.519e+06	4.621e+07	3.097e+06
COIL	20	1.992e+09	4.405e+09	2.014e+09	1.974e+09	1.975e+09	1.975e+09
COIL	50	1.335e+09	2.420e+10	5.772e+09	1.272e+09	1.252e+09	1.250e+09
COIL	100	2.936e+09	2.834e+10	1.814e+10	9.422e+08	8.623e+08	8.458e+08
COIL	200	3.362e+09	2.834e+10	4.627e+10	7.614e+08	5.720e+08	5.392e+08
COIL	300	3.946e+09	2.834e+10	7.417e+10	6.734e+08	4.609e+08	4.544e+08
TDT2	20	1.595e+06	2.211e+06	1.667e+06	1.591e+06	1.594e+06	1.592e+06
TDT2	50	1.397e+06	2.211e+06	2.285e+06	1.393e+06	1.389e+06	1.385e+06
TDT2	100	1.241e+06	2.211e+06	5.702e+06	1.216e+06	1.219e+06	1.212e+06
TDT2	200	1.484e+06	1.878e+06	1.753e+07	1.063e+06	1.104e+06	1.049e+06
TDT2	300	1.879e+06	2.211e+06	3.398e+07	1.060e+06	1.669e+06	1.007e+06

data	n	[17] PG	[14] AS	[14] BPP	[10] APG	[13] CGD	ours MSM
time limit=40							
20news	20	4.622e+06	2.762e+07	7.547e+06	4.495e+06	4.500e+06	4.496e+06
20news	50	4.386e+06	2.762e+07	1.562e+07	3.564e+06	3.478e+06	3.438e+06
20news	100	6.486e+06	2.762e+07	4.223e+07	3.128e+06	2.988e+06	2.783e+06
20news	200	6.731e+06	1.934e+07	1.003e+08	3.304e+06	5.744e+06	2.407e+06
20news	300	1.041e+07	2.762e+07	1.932e+08	3.621e+06	4.621e+07	2.543e+06
COIL	20	1.987e+09	5.141e+09	2.010e+09	1.974e+09	1.975e+09	1.975e+09
COIL	50	1.308e+09	2.403e+10	5.032e+09	1.262e+09	1.250e+09	1.248e+09
COIL	100	2.922e+09	2.834e+10	2.086e+10	9.161e+08	8.555e+08	8.430e+08
COIL	200	3.361e+09	2.834e+10	4.116e+10	7.075e+08	5.584e+08	5.289e+08
COIL	300	3.920e+09	2.834e+10	7.040e+10	6.221e+08	4.384e+08	4.294e+08
TDT2	20	1.595e+06	2.211e+06	1.643e+06	1.591e+06	1.594e+06	1.592e+06
TDT2	50	1.394e+06	2.211e+06	1.933e+06	1.392e+06	1.388e+06	1.384e+06
TDT2	100	1.229e+06	2.211e+06	5.259e+06	1.213e+06	1.216e+06	1.211e+06
TDT2	200	1.389e+06	1.547e+06	1.716e+07	1.046e+06	1.070e+06	1.041e+06
TDT2	300	1.949e+06	1.836e+06	3.369e+07	1.008e+06	1.155e+06	9.776e+05
time limit=50							
20news	20	4.565e+06	2.762e+07	6.939e+06	4.488e+06	4.498e+06	4.494e+06
20news	50	4.343e+06	2.762e+07	1.813e+07	3.525e+06	3.469e+06	3.432e+06
20news	100	6.404e+06	2.762e+07	3.955e+07	3.046e+06	2.878e+06	2.765e+06
20news	200	5.939e+06	2.762e+07	9.925e+07	3.121e+06	4.538e+06	2.359e+06
20news	300	9.258e+06	2.762e+07	1.912e+08	3.621e+06	2.323e+07	2.331e+06
COIL	20	1.982e+09	7.136e+09	2.033e+09	1.974e+09	1.975e+09	1.975e+09
COIL	50	1.298e+09	2.834e+10	4.365e+09	1.258e+09	1.248e+09	1.248e+09
COIL	100	1.945e+09	2.834e+10	1.428e+10	9.014e+08	8.516e+08	8.414e+08
COIL	200	3.362e+09	2.834e+10	3.760e+10	6.771e+08	5.491e+08	5.231e+08
COIL	300	3.905e+09	2.834e+10	6.741e+10	5.805e+08	4.226e+08	4.127e+08
TDT2	20	1.595e+06	2.211e+06	1.622e+06	1.591e+06	1.594e+06	1.592e+06
TDT2	50	1.393e+06	2.211e+06	1.875e+06	1.392e+06	1.386e+06	1.384e+06
TDT2	100	1.223e+06	2.211e+06	4.831e+06	1.212e+06	1.214e+06	1.210e+06
TDT2	200	1.267e+06	2.211e+06	1.671e+07	1.040e+06	1.054e+06	1.036e+06
TDT2	300	1.903e+06	2.211e+06	3.328e+07	9.775e+05	1.045e+06	9.606e+05

Table 1: Comparisons of objective values for non-negative matrix factorization for all the compared methods. The 1^{st}, 2^{nd}, and 3^{rd} best results are colored with red, blue and green, respectively.

Experiments on ℓ_0 Sparse Coding



Outline

- Introduction
- Proposed Matrix Splitting Method
  - Convex Quadratic Vector Case
- Extensions
  - Non-Convex Case
  - Matrix Case
  - Non-Quadratic Case
- Experiments
- Future Work

Future Work

- Further Acceleration
  - Nesterov’s Momentum Strategy
  - Richardson’s Extrapolation Strategy
- When A is Sparse → Sparse Gaussian Elimination
- Linearized ADMM → Triangle Proximal ADMM
- Matrix Splitting → Operator Splitting

Thank you!