Convex optimization algorithms for sparse and low-rank representations

Lieven Vandenberghe, Hsiao-Han Chao (UCLA)

ECC 2013 Tutorial Session

Sparse and low-rank representation methods in control, estimation, and system identification

July 17, 2013

Convex penalty functions for non-convex structure

1-norm promotes sparsity (Claerbout & Muir 1973; Tibshirani 1996, . . .)

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

Trace norm (nuclear norm) promotes low rank (Fazel, Boyd, Hindi 2001, . . .)

$$\|X\|_* = \sum_{i=1}^{n} \sigma_i(X)$$

Extensions: sums of norms, atomic norms, . . .

(Yuan & Lin 2006; Bach 2008, Chandrasekaran *et al.* 2010, Shah *et al.* 2012, . . .)

useful in convex optimization heuristics; supported by recent theory

Example: subspace system identification

$$\text{minimize} \quad \sum_{t=1}^{N} \|y(t) - \hat{y}(t)\|_2^2 + \gamma \|W_1 Y W_2\|_*$$

- variables are $y(t)$ (model outputs); Y is block-Hankel matrix from $y(t)$
- $\hat{y}(t)$ is given, measured output sequence
- different subspace methods use different W_1, W_2

Motivation

- first term penalizes deviation of model outputs from measured outputs
- 2nd term promotes low $\mathbf{rank}(W_1 Y W_2)$, preserving Hankel structure
- can add constraints on $y(t)$, use other penalties ($e.g.$, ℓ_1, Huber, . . .)

more examples and applications in the other talks of the session

Interior-point methods

Trace norm minimization (with $\mathcal{A} : \mathbf{R}^n \to \mathbf{R}^{p\times q}$ a linear mapping)

$$\text{minimize} \quad \|\mathcal{A}(x) - B\|_*$$

Equivalent semidefinite program

$$\begin{array}{ll} \text{minimize} & (\mathbf{tr}\, U + \mathbf{tr}\, V)/2 \\ \text{subject to} & \left[\begin{array}{cc} U & (\mathcal{A}(x) - B)^T \\ \mathcal{A}(x) - B & V \end{array}\right] \succeq 0 \end{array}$$

- expensive to solve via general-purpose solvers
- customized solvers have complexity $O(pqn^2)$ if $n \geq \max\{p, q\}$

  (*cf.*, complexity of dense least-squares problem of same size)

Outline

Algorithms for problems

$$\text{minimize} \quad f(x) + \gamma \|\mathcal{A}(x) - B\|_*$$

- f convex, not necessarily differentiable or strictly convex
- $\mathcal{A}(x)$ is linear matrix valued function of x

Proximal algorithms

- proximal-point algorithm: augmented Lagrangian methods
- Douglas-Rachford splitting: primal, dual (ADMM), primal-dual
- forward-backward methods: dual proximal gradient, Chambolle-Pock

Convex optimization with composite structure

$$\text{minimize} \quad f(x) + g(Ax)$$

- f and g are 'simple' convex functions
- dual has a similar structure:

$$\text{maximize } -g^*(z) - f^*(-A^T z)$$

$g^*(z) = \sup_y \left(z^T y - g(y)\right)$ and f^* are the conjugates of g and f

Example ($\|\cdot\|$ is arbitrary norm with dual norm $\|\cdot\|_{\mathrm{d}}$)

$$g(y) = \gamma\|y - b\|, \qquad g^*(z) = \begin{cases} b^T z & \|z\|_{\mathrm{d}} \le \gamma \\ +\infty & \text{otherwise} \end{cases}$$

Optimality conditions

Primal optimality conditions

$$0 \in \partial f(x) + A^T \partial g(Ax)$$

∂ denotes subdifferential (set of subgradients)

Dual optimality conditions

$$0 \in \partial g^*(z) - A \partial f^*(-A^T z)$$

Primal-dual optimality conditions

$$0 \in \begin{bmatrix} 0 & A^T \\ -A & 0 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} + \begin{bmatrix} \partial f(x) \\ \partial g^*(z) \end{bmatrix}$$

Outline

1. Duality and optimality conditions
2. **Proximal-point algorithm**
3. Douglas-Rachford splitting
4. Forward-backward and semi-implicit methods

Proximal operator

$$\operatorname{prox}_h(x) = \underset{u}{\operatorname{argmin}} \left(h(u) + \frac{1}{2}\|u - x\|_2^2\right)$$

- uniquely defined for all x (if h is closed convex)
- Moreau decomposition: $x = \operatorname{prox}_h(x) + \operatorname{prox}_{h^*}(x)$

Examples

- h is indicator function δ_C of closed convex set: Euclidean projection P_C
- $h(x) = \|x - b\|$: generalized soft-thresholding operation

$$\operatorname{prox}_{th}(x) = x - P_{tC}(x - b), \qquad tC = \{x \mid \|x\|_{\mathrm{d}} \leq t\}$$

(Moreau 1965, surveys in Bauschke & Combettes 2011, Parikh & Boyd 2013)

Proximal point algorithm

to minimize $h(x)$, apply fixed-point iteration to prox_{th}

$$x^+ = \mathrm{prox}_{th}(x)$$

- minimizers of h are fixed points of prox_{th}
- implementable if inexact prox-evaluations are used

Convergence

- $O(1/\epsilon)$ iterations to reach $h(x) - h(x^\star) \leq \epsilon$
- $O(1/\sqrt{\epsilon})$ iterations with accelerated algorithm (Güler 1992)

Monotone operator

Monotone (set-valued) operator. $F : \mathbf{R}^n \to \mathbf{R}^n$ with

$$(y - \hat{y})^T(x - \hat{x}) \geq 0 \qquad \forall x,\ \hat{x},\ y \in F(x),\ \hat{y} \in F(\hat{x})$$

Examples

- subdifferential of closed convex function
- linear function $F(x) = Bx$ with $B + B^T$ positive semidefinite
- r.h.s. of primal-dual optimality condition for composite problem

$$0 \in \begin{bmatrix} 0 & A^T \\ -A & 0 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} + \begin{bmatrix} \partial f(x) \\ \partial g^*(z) \end{bmatrix}$$

Proximal point algorithm for monotone inclusion

to solve $0 \in F(x)$, run fixed-point iteration

$$x^+ = (I + tF)^{-1}(x)$$

the mapping $(I + tF)^{-1}$ is called the **resolvent** of F

- $x = (I + tF)^{-1}(\hat{x})$ is (unique) solution of $\hat{x} \in x + tF(x)$
- resolvent of subdifferential $F(x) = \partial h(x)$ is prox-operator:

$$(I + t\partial h)^{-1}(x) = \mathrm{prox}_{th}(x)$$

- PPA converges if F has a zero and is maximal monotone

Augmented Lagrangian method

proximal-point algorithm applied to the dual in

P:	minimize	$f(x) + g(y)$	D: maximize	$-g^*(z) - f^*(-A^T z)$
	subject to	$Ax = y$		

1. minimize augmented Lagrangian

$$(x^+, y^+) = \operatorname*{argmin}_{\tilde{x},\tilde{y}} \left(f(\tilde{x}) + g(\tilde{y}) + \frac{t}{2}\|A\tilde{x} - \tilde{y} + z/t\|_2^2\right)$$

2. dual update: $z^+ = z + t(Ax^+ - y^+)$

- known in image processing as Bregman iteration (Yin *et al.* 2008)
- practical with inexact minimization (Rockafellar 1976, Liu *et al.* 2012, . . .)

Proximal method of multipliers

apply proximal point algorithm to

$$0 \in \begin{bmatrix} 0 & A^T \\ -A & 0 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} + \begin{bmatrix} \partial f(x) \\ \partial g^*(z) \end{bmatrix}$$

Algorithm (Rockafellar 1976)

1. minimize generalized augmented Lagrangian

$$(x^+, y^+) = \underset{\tilde{x},\tilde{y}}{\operatorname{argmin}} \left(f(\tilde{x}) + g(\tilde{y}) + \frac{t}{2}\|A\tilde{x} - \tilde{y} + z/t\|_2^2 + \frac{1}{2t}\|\tilde{x} - x\|_2^2 \right)$$

2. dual update: $z^+ = z + t(Ax^+ - y^+)$

Outline

1. Introduction
2. Proximal-point algorithm
3. **Douglas-Rachford splitting**
4. Forward-backward and semi-implicit methods

Douglas-Rachford splitting algorithm

$$0 \in F(x) = F_1(x) + F_2(x)$$

with F_1 and F_2 maximal monotone

Algorithm (Lions and Mercier 1979)

$$\begin{aligned} x^+ &= (I + tF_1)^{-1}(z) \\ y^+ &= (I + tF_2)^{-1})(2x^+ - z) \\ z^+ &= z + y^+ - x^+ \end{aligned}$$

- useful when resolvents of F_1 and F_2 are inexpensive, but not $(I + tF)^{-1}$
- under weak conditions (existence of solution), x converges to solution

Alternating direction method of multipliers (ADMM)

Douglas-Rachford splitting applied to optimality condition for dual

$$\text{maximize} \quad -g^*(z) - f^*(-A^T z)$$

1. alternating minimization of augmented Lagrangian

$$\begin{aligned} x^+ &= \operatorname*{argmin}_{\tilde{x}} \left(f(\tilde{x}) + \frac{t}{2}\|A\tilde{x} - y + z/t\|_2^2 \right) \\ y^+ &= \operatorname*{argmin}_{\tilde{y}} \left(g(\tilde{y}) + \frac{t}{2}\|Ax^+ - \tilde{y} + z/t\|_2^2 \right) \end{aligned}$$

2. dual update $z^+ = z + t(Ax^+ - y)$

also known as split Bregman method (Goldstein and Osher 2009)

Gabay & Mercier 1976; recent survey in Boyd, Parikh, Chu, Peleato, Eckstein 2011

Primal application of Douglas-Rachford method

D-R splitting algorithm applied to optimality condition for primal

$$\text{minimize} \quad \underbrace{f(x) + g(y)}_{h_1(x,y)} + \underbrace{\delta_{\{0\}}(Ax - y)}_{h_2(x,y)}$$

Main steps

- prox-operator of h_1: separate evaluations of prox_f and prox_g
- prox-operator of h_2: projection on subspace $H = \{(x, y) \mid Ax = y\}$

$$P_H(x, y) = \begin{bmatrix} I \\ A \end{bmatrix} (I + A^T A)^{-1}(x + A^T y)$$

also known as *method of partial inverses* (Spingarn 1983, 1985)

Primal-dual application

$$0 \in \underbrace{\begin{bmatrix} 0 & A^T \\ -A & 0 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix}}_{F_2(x,z)} + \underbrace{\begin{bmatrix} \partial f(x) \\ \partial g^*(z) \end{bmatrix}}_{F_1(x,z)}$$

Main steps

- resolvent of F_1: prox-operator of f, g
- resolvent of F_2:

$$\begin{bmatrix} I & tA^T \\ -tA & I \end{bmatrix}^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} I \\ tA \end{bmatrix} (I + t^2 A^T A)^{-1} \begin{bmatrix} I \\ -tA \end{bmatrix}^T$$

Summary: Douglas-Rachford splitting methods

$$\text{minimize } f(x) + g(Ax)$$

Most expensive steps

- **Dual** (ADMM)

$$\text{minimize (over } x\text{)} \quad f(x) + \frac{t}{2}\|Ax - y + z/t\|_2^2$$

  a linear equation with coefficient $\nabla^2 f(x) + tA^T A$ if f is quadratic

- **Primal** (Spingarn): equation with coefficient $I + A^T A$
- **Primal-dual**: equation with coefficient $I + t^2 A^T A$

Outline

1. Introduction

2. Proximal-point algorithm

3. Douglas-Rachford splitting

4. **Forward-backward and semi-implicit methods**

Forward-backward method

$$0 \in F(x) = F_1(x) + F_2(x)$$

Forward-backward iteration (for single-valued F_1)

$$x^+ = (I + tF_2)^{-1}(I - tF_1(x))$$

- converges if F_1 is co-coercive with parameter L and $t = 1/L$

$$(F_1(x) - F_1(\hat{x}))^T(x - \hat{x}) \geq \frac{1}{L}\|F_1(x) - F_1(\hat{x})\|_2^2 \quad \forall x, \hat{x}$$

- Tseng's modified method (1991) only requires Lipschitz continuous F_1

Dual proximal gradient method

$$0 \in \underbrace{\partial g^*(z)}_{F_2(z)} \underbrace{- A\nabla f^*(-A^T z)}_{F_1(z)}$$

Proximal gradient iteration

$$\begin{aligned} x &= \operatorname*{argmin}_{\tilde{x}} \left(f(\tilde{x}) + z^T A\tilde{x} \right) = \nabla f^*(-A^T z) \\ z^+ &= \operatorname{prox}_{tg^*}(z + tAx) \end{aligned}$$

- does not involve linear equation
- requires Lipschitz continuous ∇f^* (strongly convex f)
- accelerated methods: FISTA (Beck & Teboulle 2009), Nesterov's methods

for a comparison with ADMM, see Fazel, Pong, Sun, Tseng 2013

Primal-dual (Chambolle-Pock) method

$$0 \in \begin{bmatrix} 0 & A^T \\ -A & 0 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} + \begin{bmatrix} \partial f(x) \\ \partial g^*(z) \end{bmatrix}$$

Algorithm (with parameter $\theta \in [0,1]$) (Chambolle & Pock 2011)

$$\begin{aligned} z^+ &= \operatorname{prox}_{tg^*}(z + tA\bar{x}) \\ x^+ &= \operatorname{prox}_{tf}(x - tA^T z^+) \\ \bar{x}^+ &= x^+ + \theta(x^+ - x) \end{aligned}$$

- step size fixed ($t \leq 1/\|A\|_2$) or adapted by line search
- can be interpreted as semi-implicit forward-backward iteration
- can be interpreted as pre-conditioned proximal-point algorithm

Subspace system identification example

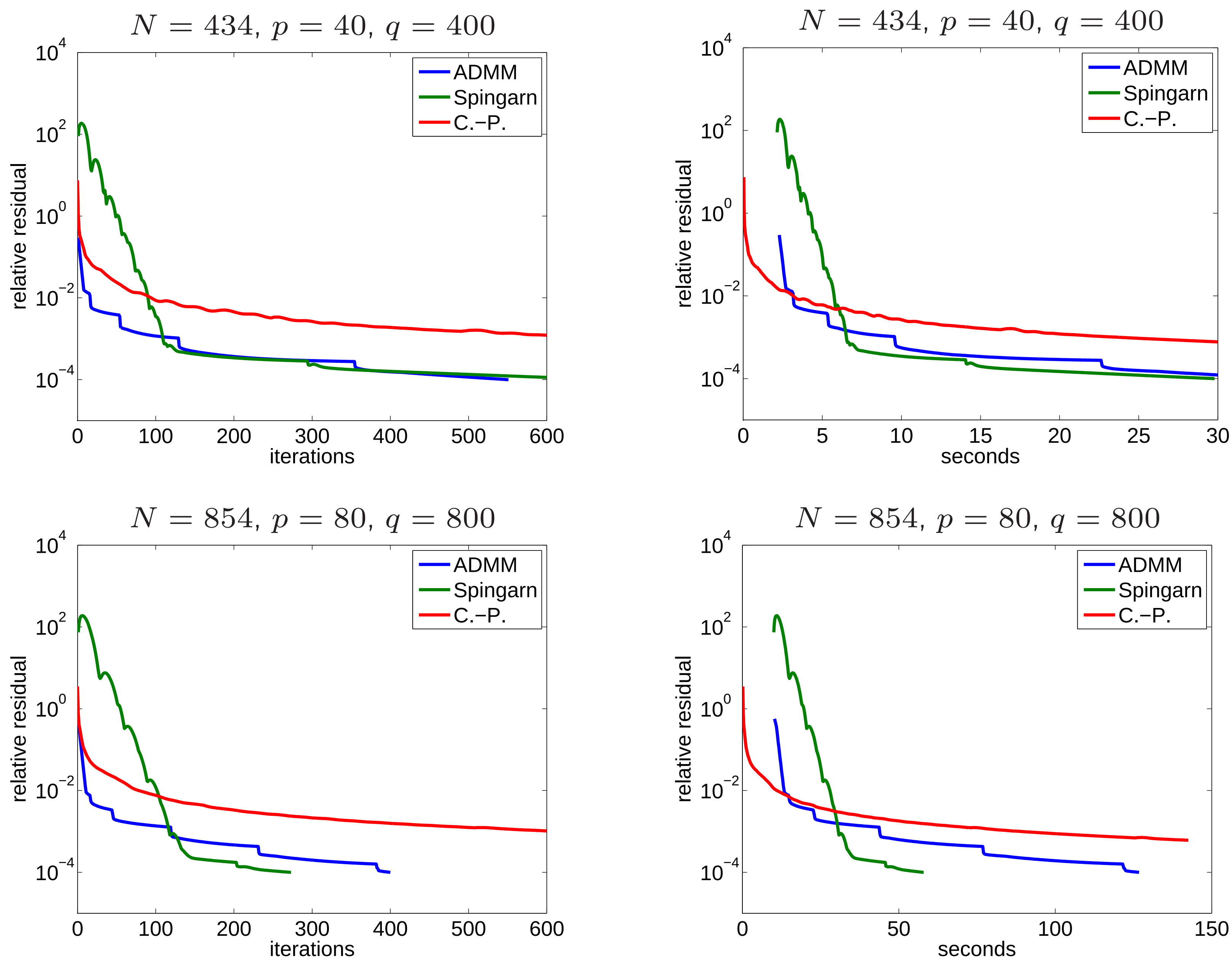
$$\text{minimize} \quad \sum_{t=1}^{N} \|y(t) - \hat{y}(t)\|_2^2 + \gamma \|Y\Pi\|_*$$

- one input, two outputs (Daisy continuous stirring tank data)
- Y is Hankel matrix from $y(t)$, $Y\Pi$ has rank n for an nth order model
- $2N$ optimization variables; $Y\Pi$ has size $p \times q$

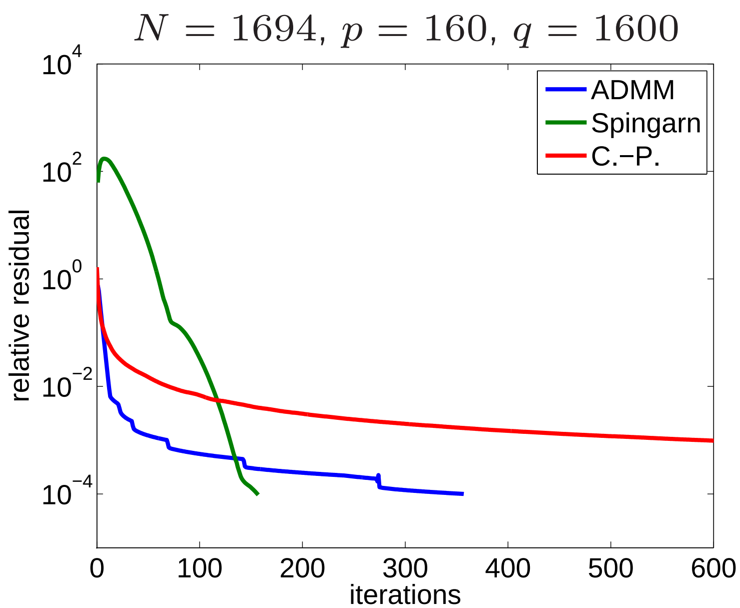
Algorithms

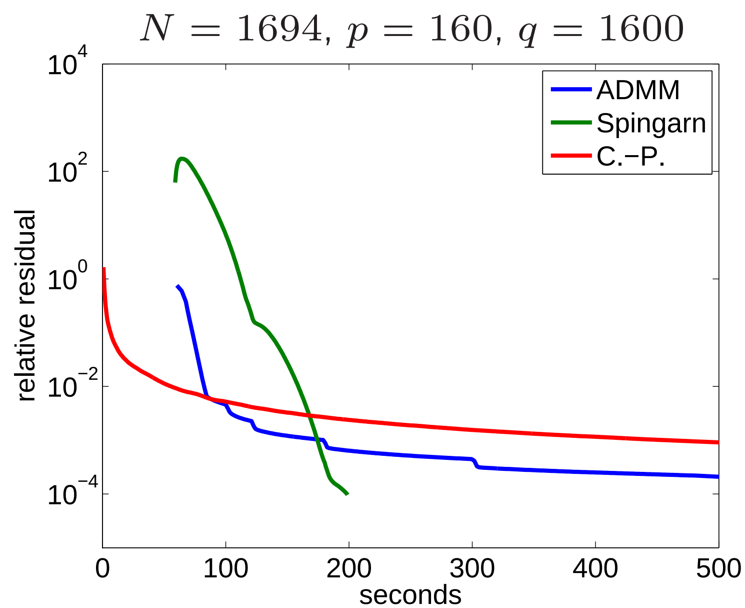
- ADMM with adaptive step size (code from Liu, Hansson, Vandenberghe 2013)
- primal Douglas-Rachord (Spingarn) with fixed step size
- Chambolle-Pock with backtracking line search

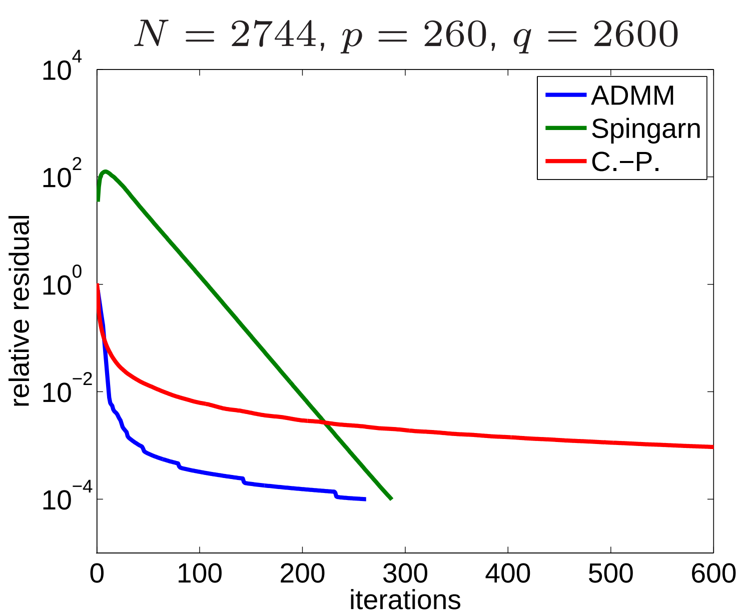
Convergence

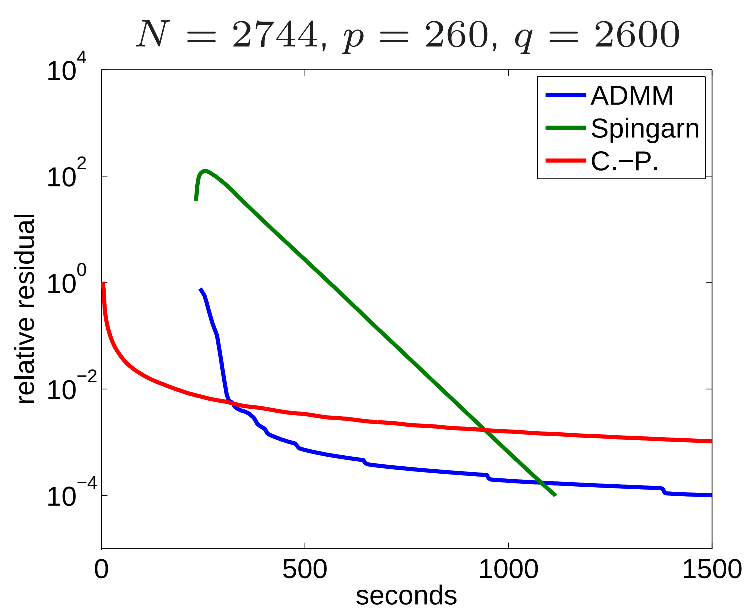


Convergence









Proximal algorithms for trace norm optimization

$$\text{minimize} \quad f(x) + \|\mathcal{A}(x) - B\|_*$$

Douglas-Rachford splitting methods (primal, dual, primal-dual)

subproblems include quadratic term $\|\mathcal{A}(x)\|_F^2$ in cost function

Forward-backward methods (dual or primal-dual)

only require application of $\mathcal{A}$ and its adjoint $\mathcal{A}^{\text{adj}}$

Proximal mapping of trace norm

- requires an SVD (for projection on max. singular value norm ball)
- avoided in methods based on nonconvex low-rank parametrizations (Recht *et al.* 2010, Burer & Monteiro 2003, . . .)