Nonsmooth, Nonconvex Minimization Problems with Applications

Xiaojun Chen

Department of Applied Mathematics

Hong Kong Polytechnic University

ICCOPT, Santiago, Chile, 26-29 July 2010

Nonsmooth, nonconvex minimization

$$\min_{x \in X} f(x),$$

where $X \subseteq R^n$ is convex but $f : R^n \to R$ is

- not convex
- not differentiable
- not locally Lipschitz in some applications

Outline

- Mathematical models and applications:
  - Traffic assignment under uncertainty
  - Distribution of points on the sphere
  - Variable selection, signal reconstruction
- Smoothing algorithms

Part I: Mathematical models and applications

- **I. Stochastic complementarity problems**

  — Traffic assignment under uncertainty

  M. Fukushima (Kyoto Univ.)

  A. Sumalee (PolyU, Transportation engineering)

  C. Zhang (Beijing Jiaotong Univ.), et al.

- **II. Optimization on the sphere**

  — Distribution of points on the sphere

  I. Sloan, R. Womersley (Univ. New South Wales)

  A. Frommer, B. Lang (Wuppertal Univ.)

  J. Ye (Victoria Univ.), et al.

- **III. The ℓ_2-ℓ_p $(0 < p < 1)$ minimization**

  — Variable selection, signal reconstruction

  Y. Ye (Stanford Univ.)

  F. Xu (Xi'an Jiaotong Univ.), W. Zhou (PolyU), et al.

I. Stochastic complementarity problems

Nonlinear complementarity problem (NCP): Given $F : R^n \to R^n$,

$$x \geq 0, \quad F(x) \geq 0, \quad x^T F(x) = 0.$$

The NCP can be reformulated as a system of nonlinear equations

$$\Phi(x, F(x)) = \begin{pmatrix} \phi(x_1, F_1(x)) \\ \vdots \\ \phi(x_n, F_n(x)) \end{pmatrix} = 0$$

or a minimization problem

$$\min_{x \in R^n} \|\Phi(x, F(x))\|^2$$

by using an NCP function ϕ.

NCP functions

A function $\phi : R^2 \to R$ is called an NCP-function if

$$\phi(a,b) = 0 \quad \Leftrightarrow \quad ab = 0, a \geq 0, b \geq 0.$$

Example of NCP functions

$\phi_{NR}(a,b) = \min(a,b)$	natural residual
$\phi_{FB}(a,b) = a + b - \sqrt{a^2 + b^2}$	Fischer-Burmeister function
$\phi_{CCK}(a,b) = \lambda \phi_{FB}(a,b) + (1-\lambda) a_+ b_+$	penalized FB function

Smoothing Newton methods and semismooth Newton methods are efficient to solve the NCP via the nonsmooth equations $\Phi(x, F(x)) = 0$ or minimization problem $\min \|\Phi(x, F(x))\|^2$.

Cottle-Pang-Stone (1992), Facchinei-Pang (2000), Ferris-Pang (1997), B.Chen-Harker (1997), C.Chen-Mangasarian (1996), Chen-Qi-Sun (1998), Chen-Ye (1999), Chen-Chen-Kanzwo (2000), Luo-Tseng(1997), Yamashita-Fukushima (1997), Qi-Sun (1993), Ralph (1994) et al.

Stochastic NCP using an NCP function

Stochastic NCP: Given $F : R^n \times \Omega \to R^n$,

$$x \geq 0, \quad F(x,\omega) \geq 0, \quad x^T F(x,\omega) = 0, \quad \text{for} \quad \omega \in \Omega.$$

- Expected value (EV) formulation

$$x \geq 0, \quad E[F(x,\omega)] \geq 0, \quad x^T E[F(x,\omega)] = 0$$

$$\Leftrightarrow \qquad \min_{x \in R^n} \|\Phi(x, E[F(x,\omega)])\|^2$$

- Expected residual minimization (ERM) formulation

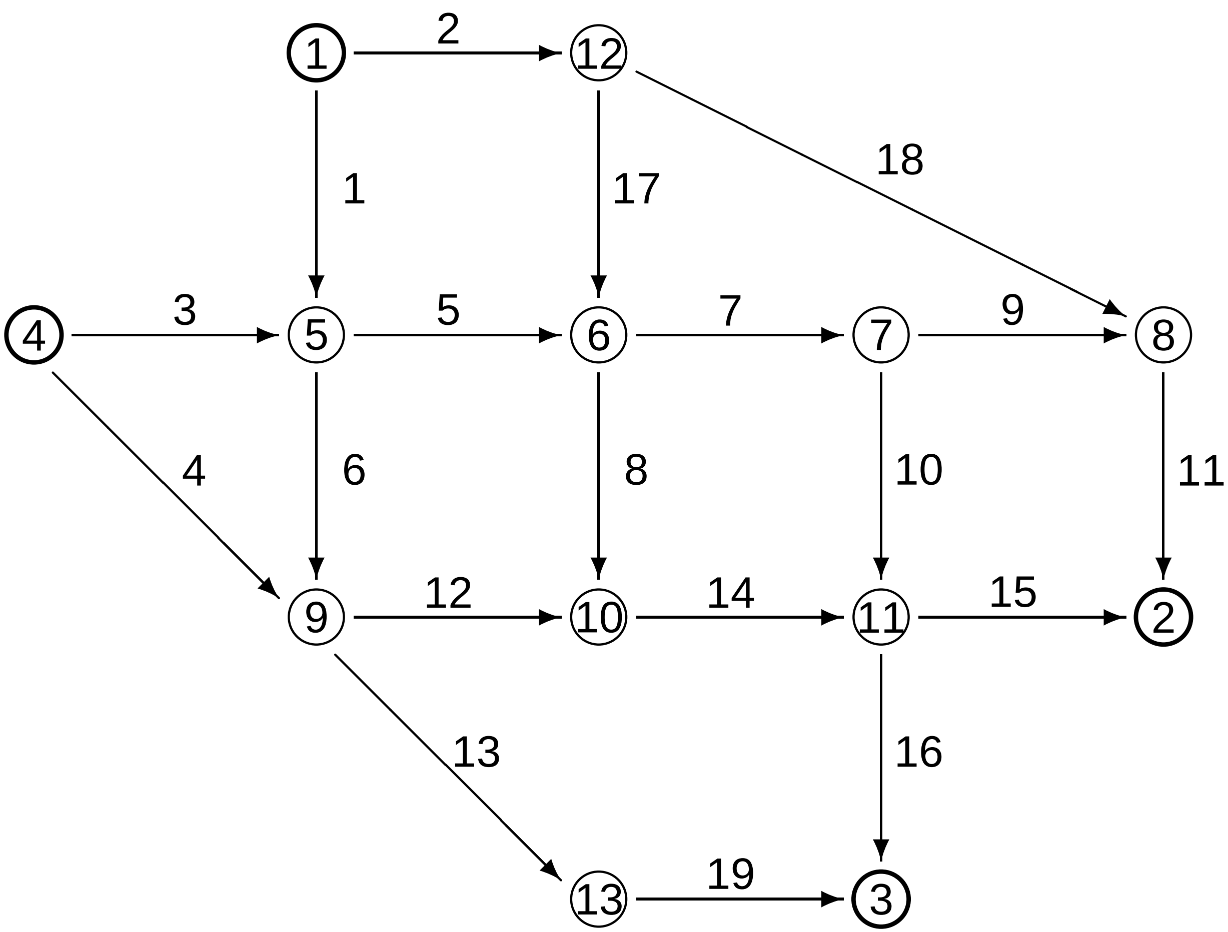
$$\min_{x \geq 0} E[\|\Phi(x, F(x,\omega))\|^2]$$

- Best worst case(BWC) formulation

$$\min_{x \geq 0} \sup_{\omega \in \Omega} \|\Phi(x, F(x,\omega))\|^2$$

Traffic assignment

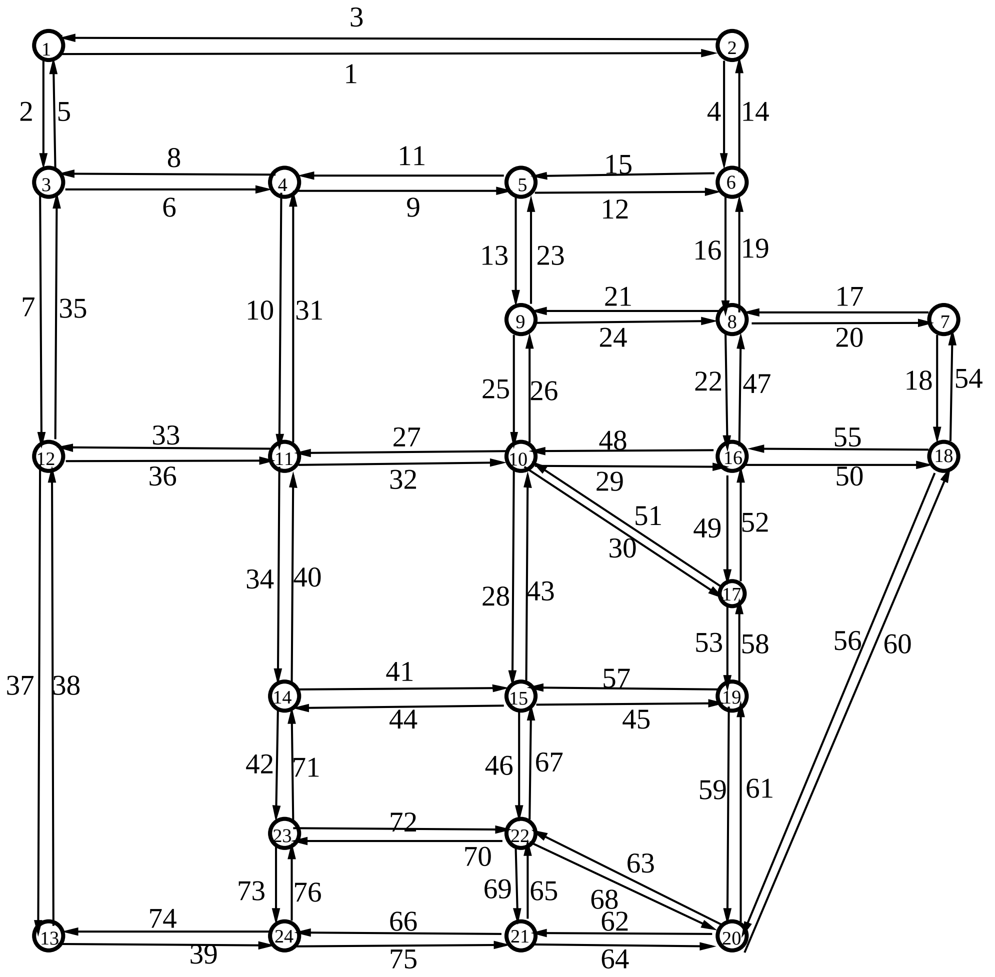
Nguyen and Dupuis Network



13 nodes, 19 links, 25 paths connecting 4 origin-destination (OD) pairs

$1 \rightarrow 2$, $4 \rightarrow 2$, $1 \rightarrow 3$ and $4 \rightarrow 3$.

Sioux Falls network



24 nodes, 76 links, 528 OD pairs, 1179 paths

Wardrop's user equilibrium

- Wardrop's user equilibrium At the equilibrium point no traveler can change his route to reduce his travel cost.
- For one scenario $\omega \in \Omega$, the static Wardrop's user equilibrium is equivalent to NCP

$$x \geq 0, \quad F(x,\omega) \geq 0, \quad x^T F(x,\omega) = 0,$$

where

$$x = \begin{pmatrix} y \\ u \end{pmatrix}, \quad F(x,\omega) = \begin{pmatrix} G(y,\omega) - \Gamma^T u \\ \Gamma y - \mathbf{Q}(\omega) \end{pmatrix}.$$

y : a path flow pattern, u : a travel cost vector.

G : path travel cost function
Γ : Origin-Destination(OD) route incidence matrix
Q : demand on each OD-pair

ERM formulation

- Expected residual minimization (ERM) formulation

$$\min_{x\geq 0} f(x) := E[\|\min(x, F(x,\omega))\|^2] \qquad \text{(ERM)}$$

  Chen-Fukushima(MOR2005),
  Fang-Chen-Fukushima(SIOPT2007).

- Error bounds

$$E[\|x - x^*_\omega\|] \leq kE[\|\min(x, F(x,\omega))\|^2]$$

$$E[\text{dist}(x - X^*_\omega)] \leq kE[\|\min(x, F(x,\omega))\|^2]$$

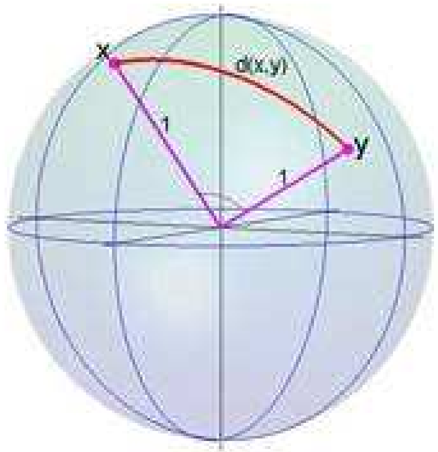
  Chen-Xiang (MP 2006, 2009, SIOPT 2007).

- Smoothing algorithms for solving ERM
  Chen-Zhang-Fukushima(MP2009), Zhang-Chen(SIOPT2009).

- Applications in traffic assignment
  Zhang-Chen-Sumalee (2009)

II. Optimization on the sphere

$\mathbb{S}^2 = \{\mathbf{z} \in R^3 \ : \ \|\mathbf{z}\|_2 = 1\}$, Area $|\mathbb{S}^2| = 4\pi$



$\mathbb{P}_t$: the linear space of restrictions of polynomials of degree $\leq t$ in 3 variables to $\mathbb{S}^2$.

dim $\mathbb{P}_t = (t+1)^2$

Distribution of points on the sphere

$\mathbb{P}_t$ can be spanned by an orthonormal set of real spherical harmonics with degree r and order k,

$$\{ \quad Y_{rk} \quad | \quad k = 1, \ldots, 2r+1, \quad r = 0, 1, \ldots, t\}.$$

Let $X_N = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\} \subset S^2$ be a set of N-points on the sphere.

The Gram matrix

$$G_t(X_N) = Y(X_N)^T Y(X_N),$$

where $Y(X_N) \in R^{(t+1)^2 \times N}$ and the j-th column of $Y(X_N)$ is given by

$$Y_{rk}(\mathbf{z}_j), \qquad k = 1, \ldots, 2r+1, \quad r = 0, 1, \ldots, t.$$

Four sets of points on the sphere

Set of points $X_N = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\} \subset S^2$

minimum energy system $\quad \text{argmin} \sum_{i \neq j}^{N} \frac{1}{\|\mathbf{z}_i - \mathbf{z}_j\|}$

extremal system $\quad \text{argmax} \det(Y(X_N)Y(X_N)^T)$

minimum cond points $\quad \text{argmin} \frac{\lambda_{\max}(Y(X_N)Y(X_N)^T)}{\lambda_{\min}(Y(X_N)Y(X_N)^T)}$

(Chen-Womersley-Ye 2010)

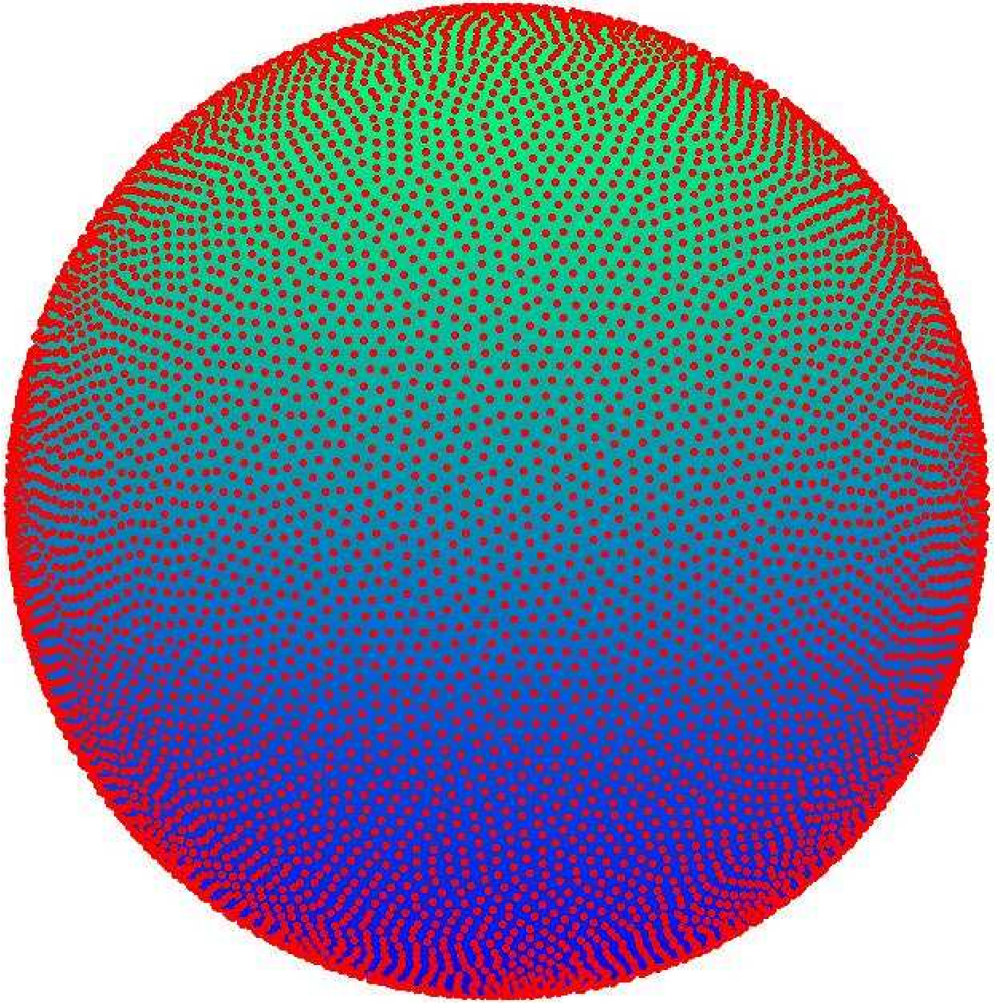
spherical $t-$design $\quad \int_{\mathbb{S}^2} p(\mathbf{z})d\mathbf{z} = \frac{4\pi}{N} \sum_{i=1}^{N} p(\mathbf{z}_i), \quad \forall p \in \mathbb{P}_t$

$\Leftrightarrow \quad F(X_N) = 0$ $\quad$ (Chen-Womersley, SINUM2006)

Well conditioned spherical t-design

(Chen-Frommer-Lang, 2009, An-Chen-Sloan-Womersley 2010)

Spherical 100-design with $N = 10201$ points



III. The ℓ_2-ℓ_p $(0 < p < 1)$ minimization

Given a matrix $A \in R^{m\times n}$, a vector $b \in R^m$, a number $\lambda > 0$,

$$\min_{x\in R^n} \|Ax - b\|_2^2 + \lambda\|x\|_p^p \qquad (\ell_2\text{-}\ell_p)$$

- Nonsmooth, nonconvex, non-Lipschitz minimization
- Compressive sensing, sparse solutions of systems
- Signal reconstruction, variable selection, image processing.

$$\|x\|_0 = \sum_{\substack{i=1\\x_i\neq 0}}^{n} |x_i|^0 \quad \longleftarrow \quad \|x\|_p^p = \sum_{i=1}^{n} |x_i|^p \quad \longrightarrow \quad \|x\|_1 = \sum_{i=1}^{n} |x_i|$$

Bruckstein-Donoho-Elad (2009), Candén-Wakin-Boyd (2008), Chartrand-Staneva (2008), Chartrand-Yin (2009), Foucart-Lai (2009), Ge-Jiang-Ye (2010), Lai-Wang (2009), Nikolova et al (2008), Xu et al (2010).

The lower bound theory I

Chen-Xu-Ye, 2009

Let a_i be the ith column of A. Let

$$L_i = \left(\frac{\lambda p(1-p)}{2\|a_i\|^2} \right)^{\frac{1}{2-p}}, \quad i = 1, \cdots, n.$$

Theorem 1 For any local solution x^* of (ℓ_2-ℓ_p), the following statements hold.

- $x_i^* \in (-L_i, L_i) \quad \Rightarrow \quad x_i^* = 0, \quad i \in \{1, \cdots, n\}.$
- The columns of the sub-matrix $B := A_\Lambda \in R^{m \times |\Lambda|}$ of A are linearly independent, where $\Lambda =$ support $\{x^*\}$.
- (ℓ_2-ℓ_p) has a finite number of local minimizers.

The lower bound theory II

Chen-Xu-Ye 2009

For an arbitrarily given point x^0, let

$$L = \left(\frac{\lambda p}{2\|A\|\sqrt{f(x^0)}} \right)^{\frac{1}{1-p}}.$$

Theorem 2 Let x^* be any local minimizer of (ℓ_2-ℓ_p) satisfying $f(x^*) \leq f(x^0)$. Then we have

- $x_i^* \in (-L, L) \quad \Rightarrow \quad x_i^* = 0, \quad i \in \{1, \cdots, n\}.$
- The number of nonzero entries in x^* is bounded by

$$\|x^*\|_0 \leq \min\left(m, \frac{f(x^0)}{\lambda L^p} \right).$$

Reweighted ℓ_1 minimization algorithm (RL1)

Chen-Zhou 2010

Given $\varepsilon > 0$. The Iterative RL1 (IRL1) for (ℓ_2-ℓ_1):

$$x^{k+1} = \arg\min_{x \in R^n} \|Ax - b\|_2^2 + \lambda \|W^k x\|_1$$

$$W^k = \operatorname{diag}(w_i^k), \qquad w_i^k = \frac{p}{(|x_i^k| + \varepsilon)^{1-p}}, \quad i = 1, \ldots, n.$$

Extensive numerical experiments (Candén-Wakin-Boyd(2008), Chartrand-Staneva(2008), et al) have shown that the IRL1 is very efficient. However no convergence results have been given.

Theorem 3 Let $\{x^k\}$ be a sequence generated by the IRL1. Then the sequence $\{x^k\}$ converges to a stationary point x^* of

$$\min_{x \in R^n} \|Ax - b\|_2^2 + \lambda \sum_{i=1}^{n} (|x_i| + \varepsilon)^p, \qquad 0 < p < 1.$$

Part II: Smoothing algorithms

- Definition 1: Let $f : R^n \to R$ be locally Lipschitz. We call $\tilde{f} : R^n \times R_+ \to R$ a smoothing function of f, if $\tilde{f}(\cdot, \mu)$ is continuously differentiable in R^n for any fixed $\mu > 0$, and

$$\lim_{\mu \downarrow 0} \tilde{f}(x, \mu) = f(x), \quad \text{for any } x \in R^n.$$

- Subdifferential associated with $\tilde{f}$

$$G_{\tilde{f}}(x) = \{V \; : \; \exists N \in \mathcal{N}^\sharp_\infty, \; x^\nu \underset{N}{\to} x, \; \mu_\nu \downarrow 0 \quad \text{with} \quad \nabla_x \tilde{f}(x^\nu, \mu_\nu) \underset{N}{\to} V\}.$$

Rockafellar and Wets (1998): $G_{\tilde{f}}(x)$ is nonempty and bounded,

$$\partial f(x) = \text{co}\{\lim_{\substack{x_i \to x \\ x_i \in D_f}} \nabla f(x_i)\} \subseteq \text{co}G_{\tilde{f}}(x).$$

In many cases: $\partial f(x) = \text{co}G_{\tilde{f}}$

Smoothing algorithms

- Choose a smoothing function $\tilde{f}(x, \mu)$ and an algorithm for smooth problems
- Use $\tilde{f}(x_k, \mu_k)$ and its gradient $\nabla \tilde{f}(x_k, \mu_k)$ at each step of the algorithm
- Update the smoothing parameter μ_k at each step. The updating scheme plays a key role in convergence analysis of the smoothing method.

Challeges:

1 How to choose a smoothing function and an algorithm for the problem ?

2 How to update the smoothing parameter μ_k ?

We develop efficient smoothing projected gradient method and smoothing conjugate gradient method.
We prove global convergence of these methods to a stationary point.

Smoothing gradient method

Step 1. Choose constants $\sigma,\ \rho \in (0,1)$, and an initial point x^0. Set $k = 0$.

Step 2. Compute the gradient

$$g_k = \nabla \tilde{f}(x^k, \mu_k).$$

Step 3. Compute the step size ν_k by the Armijo line search, where $\nu_k = \max\{\rho^0, \rho^1, \cdots\}$ and ρ^i satisfies

$$\tilde{f}(x^k - \rho^i g_k, \mu_k) \le \tilde{f}(x^k, \mu_k) - \sigma \rho^i g_k^T g_k.$$

Set $x^{k+1} = x^k - \nu_k g_k$.

Step 4. If $\|\nabla \tilde{f}(x^{k+1}, \mu_k)\| \ge n\mu_k$, then set $\mu_{k+1} = \mu_k$; otherwise, choose $\mu_{k+1} = \sigma \mu_k$.

Smoothing conjugate gradient method Chen-Zhou (SIIMS 2010).

Smoothing model of the ℓ_2-ℓ_p minimization

- Let $\psi_\mu(x) = (s_\mu(x_1), \cdots, s_\mu(x_n))^T$, and

$$s_\mu(t) = \begin{cases} |t| & |t| > \mu \\ \dfrac{t^2}{2\mu} + \dfrac{\mu}{2} & |t| \le \mu. \end{cases}$$

$$\min_{x \in R^n} f(x, \mu) := \|Ax - b\|^2 + \lambda \|\psi_\mu(x)\|_p^p.$$

- For any $\mu > 0$, the set of local minimizers $\mathcal{X}_\mu^*$ of the smoothing model is nonempty and bounded, and f_μ is continuously differentiable.

Smooth version of the lower bound theory

Let

$$L = \left(\frac{\lambda p}{2\|A\|\sqrt{f(x^0)}} \right)^{\frac{1}{1-p}} \quad \text{and} \quad L_i = \left(\frac{\lambda p(1-p)}{2\|a_i\|^2} \right)^{\frac{1}{2-p}}, \quad i = 1, \cdots, n.$$

Theorem 4

- For any local minimizer x_μ^* of the smoothing (ℓ_2-ℓ_p),

$$(x_\mu)_i^* \in (-L_i, L_i) \quad \Rightarrow \quad |(x_\mu^*)_i| \le \mu, \quad i \in \{1, \cdots, n\}.$$

- For any local minimizer x_μ^* of the smoothing (ℓ_2-ℓ_p) satisfying $f(x_\mu^*) \le f(x^0)$,

$$(x_\mu)_i^* \in (-L,\ L) \quad \Rightarrow \quad |(x_\mu^*)_i| \le \mu, \quad i \in \{1, \cdots, n\}.$$

Stationary Point

For $x \in R^n$, let $X = \mathrm{diag}(x)$.

(1) x is said to satisfy the first order necessary condition (KKT condition) of the ℓ_2-ℓ_p problem if

$$2XA^T(Ax-b)+\lambda p|x|^p=0.$$

(2) x is said to satisfy the second order necessary condition of the ℓ_2-ℓ_p problem if

$$2XA^TAX+\lambda p(p-1)\mathrm{diag}(|x|^p)$$

is positive semidefinite.

Let $\mathcal{X}$ be the set of KKT points of the ℓ_2-ℓ_p problem and $\mathcal{X}_\mu$ be the set of KKT points of its smoothing problem.

Theorem 5 Let $x_\mu \in \mathcal{X}_\mu$. We have

$$\lim_{\mu\downarrow 0} \mathrm{dist}(x_\mu, \mathcal{X}) = 0.$$

Properties of smoothing function

- $f_(x, \mu)$ is continuously differentiable and

$$| f(x, \mu) - f(x) | \leq \lambda n \left(\frac{\mu}{2}\right)^p .$$

- For any $\hat{x} \in R^n$, the level set

$$S_\mu(\hat{x}) = \{x \in R^n \big| f(x, \mu) \leq f(\hat{x}, \mu)\}$$

  is bounded;
- The gradient of $f(\cdot, \mu)$ is Lipschitz continuous.

Theorem 6 From any initial point x^0, the sequence $\{x^k\}$ generated by the SG method satisfies

$$\mu_k \equiv \varepsilon, \text{ for all large k} \quad \text{and} \quad \lim_{k\to\infty} \inf \|\nabla f(x^k, \mu_k)\| = 0.$$

Error bound

Theorem 7 There is $\bar{\mu} > 0$, such that for any $\mu \in (0, \bar{\mu}]$ and any $x_\mu \in \mathcal{X}_\mu$, there is $x^* \in \mathcal{X}$ such that

$$\Gamma_\mu := \{i \big| |(x_\mu^*)_i| \le \mu, \ \ i \in \mathcal{N}\} = \{i \big| |x_i^*| = 0, \ \ i \in \mathcal{N}\} =: \Gamma.$$

Define

$$(\bar{x}_\mu^*)_i = \begin{cases} 0 & i \in \Gamma \\ (x_\mu)_i & i \in \mathcal{N} \backslash \Gamma. \end{cases}$$

Let B be the submatrix of A whose columns are indexed by $\mathcal{N} \backslash \Gamma$.

Suppose $\lambda_{min}(B^T B) > \dfrac{\lambda p(1-p)}{2} L^{p-2}$, then

$$\big\|\bar{x}_\mu^* - x^*\big\| \le \big\|G^{-1}\big\| \big\|\nabla f(\bar{x}_\mu^*, \mu)\big\|.$$

where $G = 2B^T B + \lambda p(p-1)L^{p-2} I$, and $\lambda_{min}(B^T B)$ denotes the smallest eigenvalue of $B^T B$.

Orthogonal Matching Pursuit (OMP)

Mallat-Zhang (1993), Chen-Donoho-Saunders (1998), Mruckstein-Donoho-Elad (2009)

Parameters: Given the error threshold β.

Initialization: Set the initial point $x^0 = 0$, the initial residual $r^0 = b - Ax^0 = b$, the initial solution support $\Lambda_0 = \emptyset$.

Main Iteration: Increment k by 1 and perform the following steps:

- Find the index j_k that solves the optimization problem

$$j_k = \arg\max \frac{\|(A^{k-1}x^{k-1} - b)^T a_j\|_2^2}{\|a_j\|} \text{ for } j \in \mathcal{N} \setminus \Lambda_{k-1}.$$

- Let $\Lambda_k = \Lambda_{k-1} \bigcup [j_k]$.
- Find $x^k = \arg\min\{ \ \|Ax - b\|_2^2 \quad | \quad \text{support}\{x\} = \Lambda^k \ \}$.
- Calculate the new residual $r^k = Ax^k - b$.
- If $\|A^T r^k\| < \beta$, stop.

Output: A point $x_{omp} := x^k$, a set Λ=support(x_{omp}) and a matrix $B = A_\Lambda \in R^{m \times |\Lambda|}$.

OMP-SCG hybrid method

Step 1. Using the OMP to get x_{omp}, $\Lambda =$support(x_{omp}), and $B = A_\Lambda \in R^{m\times|\Lambda|}$.

Step 2. Using the SCG algorithm with the initial point $x^0 = x_{omp}$ to find

$$y^* = \arg\min \|By - b\|_2^2 + \lambda \|y\|_p^p.$$

Step 3. Output an numerical solution x^*, where

$$x_j^* = \begin{cases} y_j^* & |y_j| \geq L \quad \text{and} \quad j \in \Lambda, \\ 0 & j \notin \Lambda. \end{cases}$$

Other smoothing functions

$$\min_{x \in R^n} \qquad f(x) := \|Ax - b\|_2^2 + \lambda \sum_{i=1}^{n} \varphi(x_i),$$

φ is a potential function ($\alpha \in (0,1)$ is a parameter)

	Convex	Non Lipschitz
(f1)	$\varphi(t) = \|t\|$	$\varphi(t) = \|t\|^p$
	Non convex	Non Lipschitz
(f2)	$\varphi(t) = \|t\|^\alpha$	$\varphi(t) = (\|t\|^\alpha)^p$
(f3)	$\varphi(t) = \dfrac{\alpha\|t\|}{1+\alpha\|t\|}$	$\varphi(t) = \dfrac{\alpha\|t\|^p}{1+\alpha\|t\|^p}$
(f4)	$\varphi(t) = \log(\alpha\|t\|+1)$	$\varphi(t) = \log(\alpha\|t\|^p+1)$

$$|t| \;\Rightarrow\; s_\mu(t) = \begin{cases} |t| & |t| > \mu \\ \frac{t^2}{2\mu} + \frac{\mu}{2} & |t| \le \mu. \end{cases}$$

References

I. Stochastic complementarity problems

1. X. Chen and M. Fukushima, Expected Residual Minimization Method for Stochastic Linear Complementarity Problems, Mathematics of Operations Research 30(2005) 1022-1038.
2. H. Fang, X. Chen and M. Fukushima, Stochastic R0 matrix linear complementarity problems , SIAM J. Optimization 18(2007) 482-506.
3. X. Chen, C. Zhang and M. Fukushima, Robust solution of monotone stochastic linear complementarity problems, Mathematical Programming 117(2009) 51-80
4. C. Zhang and X. Chen, Smoothing projected gradient method and its application to stochastic linear complementarity problems, SIAM J. Optimization 20(2009) 627-649.
5. C. Zhang, X. Chen and A Sumalee, Robust Wardrop's User Equilibrium Assignment under Stochastic Demand and Supply: Expected Residual Minimization Approach, Submitted to Transportation Research Part B (second revised version, minor changes, 2010)

References

II Optimization on the sphere

1. X. Chen and R. Womersley, Existence of solutions to systems of underdetermined equations and spherical designs , SIAM J. Numerical Analysis 44(2006) 2326-2341
2. X. Chen, A. Frommer and B. Lang, Computational Existence Proofs for Spherical t-Designs , to appear in Numerische Mathematik.
3. C. An, X. Chen, I. H. Sloan, R. S. Womersley, Well conditioned spherical designs for integration and interpolation on the two-Sphere, to appear in SIAM J. Numerical Analysis.
4. X. Chen, R. S. Womersley and J. Ye, Minimizing the Condition Number of a Gram Matrix, 2010, accepted for publication in SIAM J. Optimization subject to a revised version.

References

III Nonsmooth, nonconvex regularization

1. X.Chen, F. Xu and Y. Ye, Lower bound theory of nonzero entries in solutions of ℓ_2-ℓ_p minimization, to appear in SIAM J. Scientific Computing
2. X. Chen and W. Zhou, Smoothing Nonlinear Conjugate Gradient Method for Image Restoration using Nonsmooth Nonconvex Minimization, to appear in SIAM J. Imaging Sciences
3. X. Chen and W. Zhou, Convergence of Reweighted ℓ_1 Minimization Algorithms and Unique Solution of Truncated ℓ_p Minimization, April, 2010, http://www.polyu.edu.hk/ama/staff/xjchen/ChenXJ.htm

Thank you

Note: The remainder of slides is for questions.

Computational results

LASSO: Solve the ℓ_2-ℓ_1 problem by the least squares algorithm (2004).

IRL1: At kth iteration, use LASSO to solve the following ℓ_2-ℓ_1 problem

$$\min \|Ax - b\|_2^2 + \lambda \sum_{i=1}^{n} \frac{|x_i|}{\sqrt{|x_i|^k + \varepsilon}},$$

where $\varepsilon > 0$ is a parameter.

OMP-SCG

Example 1: Variable selection

- This example is artificially generated and is used firstly in Tibshirani (1996).

- True optimal solution $x^* = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$.
  We simulated 100 data sets consisting of m observations from the model
  $$Ax = b + \sigma\epsilon,$$
  where ϵ is standard normal.

- MSE: The mean squared errors over the test set;
  ANZ: The average number of correctly identified zero coefficient;
  NANZ: The average number of the coefficients erroneously set to zero over test set.

Results for variable selection

m	σ	Approach	MSE	ANZ	$NANZ$
		LASSO	0.4730	4.77	0.23
40	3	IRL1	0.4688	4.83	0.17
		OMP-SCG	0.4755	4.88	0.12
		LASSO	0.1595	4.77	0.23
40	1	IRL1	0.1541	4.86	0.14
		OMP-SCG	0.1511	4.91	0.09
		LASSO	0.3582	4.92	0.08
60	1	IRL1	0.3503	4.93	0.07
		OMP-SCG	0.3464	4.95	0.05

The OMP-SCG performs the best, followed by LASSO and IRL1.

Example 2: Prostate cancer

- This data sets are from the UCI Standard database.
- The data set consists of the medical records of 97 patients who were about to receive a radical prostatectomy. The predictors are eight clinical measures: lcavol, lweight, age, lbph, svi,lcp, gleason and pgg45.
- One of the main aims here is to identify which predictors are more important in predicting the response.

Results for prostate cancer

Parameter	LASSO	IRL1	OMP-SCG
x_1(lcavol)	0.545	0.6187	0.6436
x_2(lweight)	0.237	0.2362	0.2804
x_3(lage)	0	0	0
x_4(lbph)	0.098	0.1003	0
x_5(svi)	0.165	0.1858	0.1857
x_6(lcp)	0	0	0
x_7(gleason)	0	0	0
x_8(pgg45)	0.059	0	0
Number of nonzreo	5	4	3
Prediction error	0.478	0.468	0.4419

- SCG and OMP-SCG succeed in finding three main factors and have better prediction accuracy than IRL1 and LASSO.

Error bound

- For given ε,

$$\begin{aligned}\left\|\bar{x}_\mu^* - x^*\right\| &\leq \left\|G^{-1}\right\|\left\|\nabla f(\bar{x}_\mu^*, \mu)\right\| \\ &=: \text{error bound}\end{aligned}$$

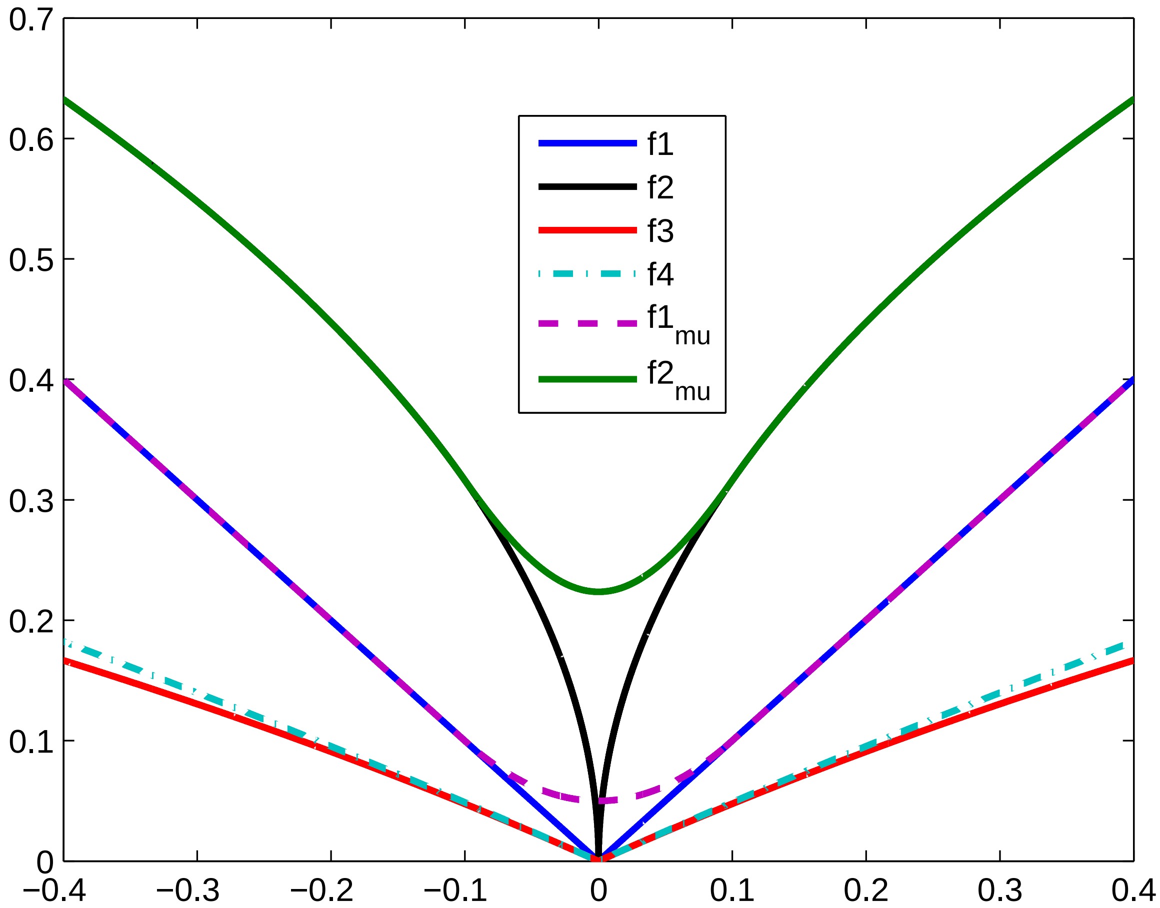
μ	L	λ	error bound
0.001	0.015	0.1304	1.5793×10^{-5}
0.0001	0.0119	0.1164	5.7310×10^{-6}
0.00001	0.0119	0.1164	5.5721×10^{-6}

Example 3: signal reconstruction

- A real-valued, finite-length signal $x \in R^n$ and x is T-sparse;
- $A \in R^{m \times n}$ is a Gaussian random matrix.

Problem	LASSO	IRL1	OMP-SCG			
	(Error,Time)	(Error,Time)	L	λ	Error	Time
$n = 512$ $T = 60$ $m = 184$	$(5.33 \times 10^{-4}$, 0.653)	$(1.29 \times 10^{-5}$, 6.82)	0.8	0.002	1.12×10^{-16}	1.02
$n = 512$ $T = 60$ $m = 182$	(38.64, 0.43)	$(2.41 \times 10^{-5}$, 7.84)	0.7	0.001	1.03×10^{-16}	1.34
$n = 512$ $T = 130$ $m = 225$	(122.25, 0.69)	(119.43, 19.99)	0.00001	0.00006	0.41	4.03

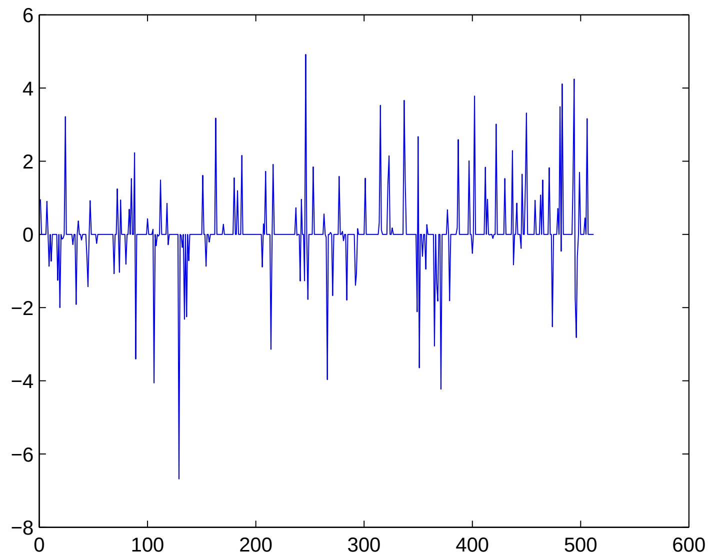
$$f1(t) = |t|, f2(t) = |t|^{\frac{1}{2}}, f3(t) = \frac{\alpha|t|}{1+\alpha|t|}, f4(t) = \log(\alpha|t|+1)$$



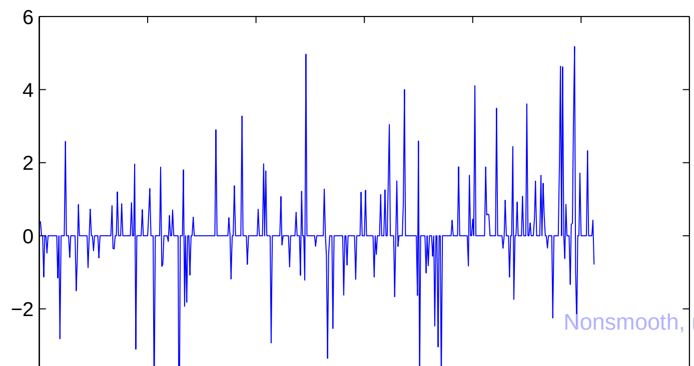
Results of prostate cancer by all the PFs

p	*(L, Number of nonzero, Prediction error)*			
	f_1	f_2	f_3	f_4
0.9	(0.0001, 4, 0.4754)	(0.011, 4, 0.473)	(2.500, 4, 0.475)	(2.040, 4, 0.474)
0.8	(0.0015, 4, 0.4740)	(0.013, 4, 0.468)	(1.990, 4, 0.474)	(1.851, 4, 0.474)
0.7	(0.0050, 4, 0.4741)	(0.012, 4, 0.465)	(1.755, 4, 0.474)	(1.550, 4, 0.474)
0.6	(0.0084, 4, 0.4661)	(0.015, 3, 0.446)	(1.545, 4, 0.475)	(1.344, 4, 0.475)
0.5	(0.0119, 3, 0.4419)	(0.016, 3, 0.445)	(1.420, 3, 0.477)	(1.200, 3, 0.483)
0.4	(0.0148, 3, 0.4456)	(0.014, 3, 0.445)	(1.480, 3, 0.477)	(1.114, 3, 0.484)
0.3	(0.0176, 3, 0.4429)	(0.012, 3, 0.443)	(1.590, 3, 0.484)	(1.190, 3, 0.483)
0.2	(0.0196, 3, 0.4359)	(0.018, 3, 0.443)	(1.955, 3, 0.483)	(1.240, 3, 0.482)

signal reconstruction



(a) Original signal



Remarks on lower bound theory

- The theory establishes a theoretical justification for “zeroing” some small entries in an approximate solution.
- The theory gives a theoretical explanation why using $\|x\|_p^p$ can generate more sparse solutions.
- The theory shows clearly the relationship between the sparsity of the solution and the choice of the regularization parameter and norm.
- It provides a systematic mechanism for selecting the regularization parameter.

Notations

- $Q^r(\omega)$: demand on each OD-pair
- $C_a(\omega)$: capacity on each link
- K: link-route incidence matrix
- Γ: OD-route incidence matrix
- The generalized Bureau of public road (BPR) link cost function

$$T_a(v,\omega) = t_a^0\Big(1 + b_a\Big(\frac{v_a}{C_a(\omega)}\Big)^{n_a}\Big),$$

  where t_a^0, b_a and n_a are given parameters and v_a is the link flow.
- The nonadditive path travel cost function

$$G(y,\omega) = \eta_1 K^T T(Ky,\omega) + \Psi(K^T T(Ky,\omega)) + \Lambda(y,\omega),$$

  where y is the path flow, $\eta_1 > 0$ is the time-based operating costs factor, Ψ is the translation function converting time T to money, and Λ is the perturbed financial cost function.

Main Contribution for the ℓ_2-ℓ_p minimization

joint work with F. Xu, Y. Ye, W. Zhou

- We derive a lower bound theory for nonzero entries in every local minimizer of the ℓ_2-ℓ_p minimization problems. This theory shows clearly the relationship between the sparsity of the solution and the choice of parameters in the model.
- We develop a hybrid orthogonal matching pursuit-smoothing conjugate gradient method.
- We prove global convergence of the ℓ_1 reweighted minimization algorithm.
- We prove uniqueness of solution under the truncated null space property which is weaker than the restricted isometry property introduced by Candés and Tao (2005).

Uniqueness

$$\min_{x \in R^n} \|x_T\|_p^p, \qquad \text{s.t.} \quad Ax = b, \tag{1}$$

where $\|x_T\|_p^p = \sum_{i \in T} |x_i|^p$ and T is a subset of $\{1, \ldots, n\}$.

$$\mathcal{F} = \{ x \,|\, Ax = b\}, \qquad S(x) = \{ i \quad | \quad x_i \neq 0 \}$$

Theorem 3 Let $x^* \in \mathcal{F}$ and T be a subset of $\{1, \ldots, n\}$. Let $S = T \cap S(x^*)$. If $S = \emptyset$, then x^* is a solution of (1). If $S \neq \emptyset$ and

$$\|\eta_S\|_p \leq \gamma \|\eta_{(T \cap S^C)}\|_p, \qquad \gamma < 1$$

for all $\eta \in N(A)$, then x^* is the unique solution of (1).

Ge-Jiang-Ye (2010) showed that for $T = \{1, \ldots, n\}$, (1) is NP-hard. The set of basic feasible solutions is the set of local minimizers.

Truncated null space property

A satisfies the t-null space property(NSP) of order K for $\gamma > 0$, $0 < t \leq n$ if

$$\|\eta_S\|_p \leq \gamma \|\eta_{(T \cap S^C)}\|_p$$

$|T| = t$, all subsets $S \subset T$ with $|S| \leq K$, and all $\eta \in N(A)$.

Theorem 4 If A satisfies the restricted isometry property

$$\alpha_s \|x\|_2 \leq \|Ax\|_2 \leq \beta_s \|x\|_2, \qquad \forall \quad \|x\|_0 \leq s$$

$$\frac{\beta_{2t_1}^2}{\alpha_{2t_1}^2} - 1 < 4(\sqrt{2} - 1)\left(\frac{t_1}{K}\right)^{\frac{1}{p} - \frac{1}{2}},$$

for some $t_1 \geq K$ then A satisfies the t-NSP of order K for $\gamma < 1$ and $|T| = n$.