Nonsmooth, Nonconvex Regularized Optimization for Sparse Approximations

Xiaojun Chen

Department of Applied Mathematics
The Hong Kong Polytechnic University

24 August 2012

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

DEPARTMENT OF APPLIED MATHEMATICS
應 用 數 學 系

Regularized minimization problem

$$\min_{x \in R^n} \quad f(x) := \Theta(x) + \lambda \sum_{i=1}^{r} \varphi(d_i^T x), \tag{1}$$

- $\Theta : R^n \to R_+$ is continuously differentiable,
  and $\nabla\Theta$ is globally Lipschitz with a Lipschitz constant $\beta > 0$.
- $\varphi : R \to R_+$ is continuous, $\varphi(t) = \varphi(-t)$, $\varphi(0) = 0$,
  and nondecreasing, concave in $[0, \infty)$,
  continuously differentiable in $(0, \infty)$.
- $d_i \in R^n$, $i = 1, \ldots, r$.
- $\lambda > 0$, regularization parameter.

Nonsmooth, nonconvex, non-Lipschitz minimization

- Compressive sensing, sparse solutions of systems
- Signal reconstruction, variable selection, image processing.

Baraniuk, Plenary Talk, ISMP2012, $\|x\|_p^p = \sum_{i=1}^n |x_i|^p$, $p \in (0, 1]$.

Nonconvex least squares problems

$$\min_{x \in R^n} \|Ax - b\|^2 + \lambda \sum_{i=1}^{r} \varphi(d_i^T x), \qquad \text{(LS)}$$

$A \in R^{m \times n}$, $b \in R^m$ with $\Theta(x) = \|Ax - b\|^2$ and $\beta = \|\nabla^2 \Theta(x)\| = 2\|A^T A\|$.
Widely used penalty functions:

$$\varphi(t) = \frac{\alpha|t|}{1+\alpha|t|}, \qquad \varphi(t) = \log(1+\alpha|t|),$$

$$\varphi(t) = \int_0^{|t|} (1 - s/(\alpha\lambda))_+ ds, \qquad \varphi(t) = \lambda - (\lambda - |t|)_+^2/\lambda$$

$$\varphi(t) = \int_0^{|t|} \min(1, (\alpha - s/\lambda)_+/(\alpha - 1)) ds, \qquad \varphi(t) = |t|^p,$$

where $\alpha > 0$ and $p \in (0, 1)$.
For $|t|^p$, the smoothness and convexity are dependent on the value of p,

Nonconvex least squares problems

$$\min_{x\in R^n} \|Ax-b\|^2 + \lambda \sum_{i=1}^{r} \varphi(d_i^T x), \qquad \text{(LS)}$$

$A \in R^{m\times n}$, $b \in R^m$ with $\Theta(x) = \|Ax-b\|^2$ and $\beta = \|\nabla^2\Theta(x)\| = 2\|A^TA\|$.
Widely used penalty functions:

$$\varphi(t) = \frac{\alpha|t|}{1+\alpha|t|}, \qquad \varphi(t) = \log(1+\alpha|t|),$$

$$\varphi(t) = \int_0^{|t|} (1 - s/(\alpha\lambda))_+ ds, \qquad \varphi(t) = \lambda - (\lambda - |t|)_+^2/\lambda$$

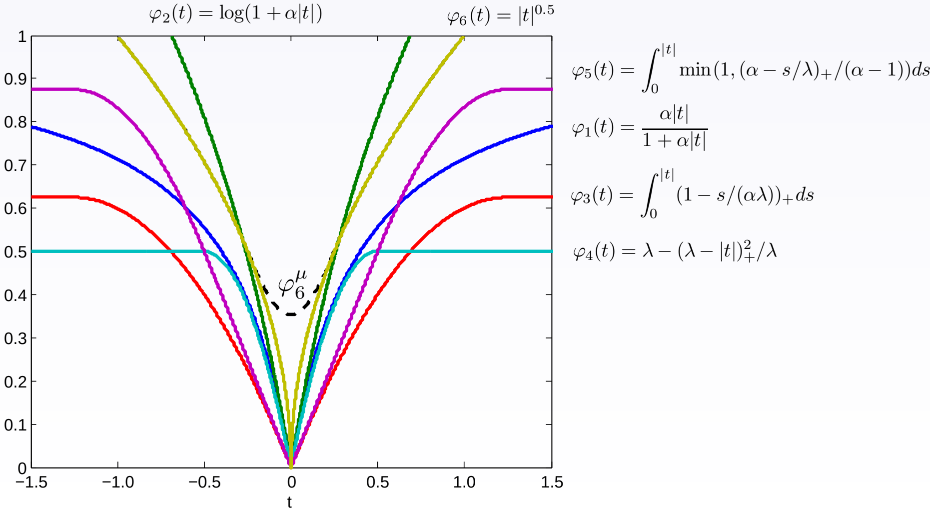
$$\varphi(t) = \int_0^{|t|} \min(1, (\alpha - s/\lambda)_+/(\alpha-1))ds, \qquad \varphi(t) = |t|^p,$$

where $\alpha > 0$ and $p \in (0,1)$.
For $|t|^p$, the smoothness and convexity are dependent on the value of p,

- $|t|^p$ $(p > 1)$ is smooth, convex,
- $|t|^p$ $(p = 1)$ is nonsmooth, convex,
- $|t|^p$ $(0 < p < 1)$ is non-Lipschitz, nonconvex.

Concave penalty functions

$$\alpha = 2.5,\ \ \lambda = 0.5,\ \mu = 0.5,\ \varphi_6^\mu(t) = s(t,\mu)^{0.5},\ \ s(t,\mu) = \begin{cases} |t| & \text{if } |t| \geq \mu \\ \dfrac{t^2}{2\mu} + \dfrac{\mu}{2} & \text{if } |t| < \mu \end{cases}$$



Joint work with

Wei Bian, Dongdong Ge, Lengfeng Niu, Michael Ng, Fengmin Xu, Zizhou Wang, Yinyu Ye, Ya-xiang Yuan, Chao Zhang, Weijun Zhou

1. W. Bian, X. Chen, Y. Ye, Complexity analysis of interior point algorithms for non-Lipshchitz and nonconvex optimization, July, 2012.
2. X. Chen, L. Niu and Y. Yuan, Optimality conditions and smoothing trust region Newton method for non-Lipschitz optimization, March, 2012.
3. W. Bian and X. Chen, Smoothing SQP algorithm for non-Lipschitz optimization with complexity analysis, February 2012.
4. X. Chen, D. Ge, Z. Wang and Y. Ye, Complexity of the unconstrained L_2-L_p minimization, May 2011.
5. X. Chen, M. Ng and C. Zhang, Non-Lipschitz ℓ_p-regularization and box constrained model for image restoration, IEEE Trans. Imaging P 2012.
6. X. Chen, F. Xu and Y. Ye, Lower bound theory of nonzero entries in solutions of ℓ_2-ℓ_p minimization, SIAM Sci. Comput., 2010.
7. X. Chen and W. Zhou, Smoothing nonlinear conjugate gradient method for image restoration using nonsmooth nonconvex minimization, SIAM Imaging Sci., 2010.

Smoothing / Interior Point Algorithms $(0 < p \leq 1)$

Wei Bian, X. Chen, Smoothing SQP algorithm, complexity $O(\epsilon^{-2})$

$$\min_{x \in R^n} \quad f(x) := \Theta(x) + \lambda \sum_{i=1}^{n} \varphi(|x_i|^p) \tag{2}$$

Bian, Chen, Yinyu Ye, Interior point algorithms, complexity $O(\epsilon^{-2})$, $O(\epsilon^{-\frac{3}{2}})$

$$\min_{x \in [0,u]} \quad f(x) := \Theta(x) + \lambda \sum_{i=1}^{n} \varphi(x_i^p) \tag{3}$$

Chen, Lingfeng Niu, Ya-xiang Yuan, Smoothing trust region Newton method

$$\min_{x \in R^n} \quad f(x) := \Theta(x) + \lambda \sum_{i=1}^{r} \varphi(|d_i^T x|^p) \tag{4}$$

Smoothing / Interior Point Algorithms ($0 < p \le 1$)

Nonconvex penalty function

$$\Phi(x) = \lambda \sum_{i=1}^{r} \varphi(|d_i^T x|^p), \qquad 0 < p \le 1$$

Let $t = |d_i^T x|^p$.

$$\varphi(t) = \frac{\alpha t}{1+\alpha t}, \qquad \varphi(t) = \log(1+\alpha t),$$

$$\varphi(t) = \int_0^t (1 - s/(\alpha\lambda))_+ ds, \qquad \varphi(t) = \lambda - (\lambda - t)_+^2/\lambda$$

$$\varphi(t) = \int_0^t \min(1, (\alpha - s/\lambda)_+/(\alpha - 1)) ds, \qquad \varphi(t) = t,$$

$D = [d_1, \ldots, d_r]^T$ is the first order difference matrix, Total Variation (TV)
$D = [d_1, \ldots, d_n]^T$ is the identity matrix.

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 $\lambda > 0$,

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

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 $\lambda > 0$,

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

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

Bruckstein-Donoho-Elad (2009), Candén-Wakin-Boyd (2008), Chartrand-Staneva (2008), Chartrand-Yin (2009), Fan-Li (2001), Foucart-Lai (2009), Knight-Fu (2000), Ge-Jiang-Ye (2010), Hintermueller-Wu (2012), Huang-Horowitz-Ma (2008), Lai-Wang (2009), Men-Yang (2010), Nikolova et al (2008, 2011), Shi et al(2012), Xu et al (2010, 2012), Zhang (2010), Lu (2012), Sun (2012), Yin (2012), Ito-Kunisch (2012), So et al (2012), Lewis-Wright (2012), Wen et al (2012).

Extension

Low rank matrix: $\sum \sigma_i^p(X)$ and Group selection: $\sum(\sum_{i\in I_j} |x_i|)^p$

The lower bound theory I

Chen-Xu-Ye 2010 SIAM Sci. Comput.

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 minimizer x^* of the ℓ_2-ℓ_p problem, 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 $\{x^*\}$.
- The ℓ_2-ℓ_p problem has a finite number of local minimizers.

$\| \cdot \| := \| \cdot \|_2$.

The lower bound theory II

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 the ℓ_2-ℓ_p problem satisfying $f(x^*) \le f(x^0)$. 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

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

Sparsity of minimizers of the ℓ_2-ℓ_p problem

Chen-Ge-Wang-Ye 2011

Theorem 3 Let

$$\beta(k) = k^{\frac{p}{2}-1}\left(\frac{2\alpha}{p(1-p)}\right)^{\frac{p}{2}}\|b\|^{2-p}, \quad \alpha = \max_{1\le i\le n}\|a_i\|^2, \quad 1 \le k \le n.$$

- If $\lambda \ge \beta(k)$, any minimizer x^* of the ℓ_2-ℓ_p problem satisfies $\|x^*\|_0 < k$ for $k \ge 2$.
- If $\lambda \ge \beta(1)$, $x^* = 0$ is the unique minimizer of the ℓ_2-ℓ_p problem.
- Suppose that set $C := \{\, x \mid Ax = b \,\}$ is non-empty. Then, if $\lambda \le \frac{\|b\|^2}{\|x_c\|_p^p}$ for some $x_c \in C$, any minimizer x^* of the ℓ_2-ℓ_p problem satisfies $\|x^*\|_0 \ge 1$.

Theorem 4 Let

$$\gamma(k) = k^{p-1}\left(\frac{2\|A\|}{p}\right)^{p}\|b\|^{2-p}.$$

If $\lambda \ge \gamma(k)$, then any local minimizer x^* of the ℓ_2-ℓ_p problem, with $f(x^*) \le f(0) = \|b\|^2$, satisfies $\|x^*\|_0 < k$ for $k \ge 2$.

The complexity of the ℓ_q-ℓ_p minimization

Chen-Ge-Wang-Ye 2011

Given $A \in R^{m\times n}$, $b \in R^m$, $\lambda > 0$, $q \geq 1, 0 \leq p < 1$, consider

$$\min_{x\in R^n} \|Ax - b\|_q^q + \lambda \|x\|_p^p \qquad (\ell_q\text{-}\ell_p).$$

Theorem 5 The ℓ_q-ℓ_p minimization is strongly NP-hard.

Consider

$$\min_{x\in R^n} \|Ax - b\|_q^q + \lambda \sum_{i=1}^{n} (|x_i| + \epsilon)^p \qquad (\ell_q\text{-}\ell_p\text{-}\epsilon),$$

where $\epsilon > 0$.

Theorem 6 The ℓ_q-ℓ_p-ϵ minimization is strongly NP-hard.

The complexity of constrained problems

Ge-Jiang-Ye (Math. Program. 2011) show that the following two problems are strongly NP hard

$$\begin{aligned} \min_{x \in R^n} \quad & \|x\|_p^p \\ \text{s.t.} \quad & Ax = b \end{aligned}$$

and

$$\begin{aligned} \min_{x \in R^n} \quad & \||x| + \epsilon\|_p^p \\ \text{s.t.} \quad & Ax = b. \end{aligned}$$

Natarajan (SIAM J. Computing, 1995) show that the following problem is NP-hard

$$\begin{aligned} \min_{x \in R^n} \quad & \|x\|_0 \\ \text{s.t.} \quad & \|Ax - b\|_2 \le \epsilon \end{aligned}$$

$\epsilon > 0$.

Smoothing algorithms

- **Definition 1:** Let $f : R^n \to R$ be continuous. 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

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

- **Subdifferential associated with $\tilde{f}$ if f is locally Lipschitz**

$$G_{\tilde{f}}(x) = \{ v \ : \ \nabla_x \tilde{f}(x^k, \mu_k) \to v, \ \text{for } x^k \to x, \ \mu_k \downarrow 0 \ \}.$$

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

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

Smoothing algorithms

- **Definition 1:** Let $f : R^n \to R$ be continuous. 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

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

- **Subdifferential associated with $\tilde{f}$ if f is locally Lipschitz**

$$G_{\tilde{f}}(x) = \{ v \ : \ \nabla_x \tilde{f}(x^k, \mu_k) \to v, \ \text{for } x^k \to x, \ \mu_k \downarrow 0 \ \}.$$

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

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

Gradient Consistency

$$\partial f(x) = \mathrm{co} G_{\tilde{f}}$$

Chen: Composite $(x)_+$, ISMP2012 special issue

Burke, Hoheisel, Kanzow: smoothing functions, draft 2012

Smoothing algorithms

Main steps

- Choose a smoothing function $\tilde{f}(x,\mu)$
  and an algorithm for the smooth problems.
- Use $\tilde{f}(x^k,\mu_k)$ and its gradient $\nabla_x \tilde{f}(x^k,\mu_k)$ 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.

Smoothing algorithms

Main steps

- Choose a smoothing function $\tilde{f}(x,\mu)$
  and an algorithm for the smooth problems.
- Use $\tilde{f}(x^k,\mu_k)$ and its gradient $\nabla_x \tilde{f}(x^k,\mu_k)$ 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.

Challenges:

- How to choose smoothing functions and algorithms for the problem ?
- How to update the smoothing parameter μ_k ?

Smoothing algorithms

Main steps

- Choose a smoothing function $\tilde{f}(x,\mu)$
  and an algorithm for the smooth problems.
- Use $\tilde{f}(x^k,\mu_k)$ and its gradient $\nabla_x \tilde{f}(x^k,\mu_k)$ 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.

Challenges:

- How to choose smoothing functions and algorithms for the problem ?
- How to update the smoothing parameter μ_k ?

Smoothing projected gradient method Zhang-Chen, SIAM Optim. 2009
Smoothing conjugate gradient method Chen-Zhou, SIAM Imaging Sci. 2010
Smoothing direct-search methods Garmanjani-Vicente, IMA NA, to appear
global convergence of these methods to a stationary point.

Smoothing / Interior Point Algorithms $(0 < p \le 1)$

Wei Bian, X. Chen, Smoothing SQP algorithm, complexity $O(\epsilon^{-2})$

$$\min_{x \in R^n} \quad f(x) := \Theta(x) + \lambda \sum_{i=1}^{n} \varphi(|x_i|^p) \tag{2}$$

Bian, Chen, Yinyu Ye, Interior point algorithms, complexity $O(\epsilon^{-2})$, $O(\epsilon^{-\frac{3}{2}})$

$$\min_{x \in [0,u]} \quad f(x) := \Theta(x) + \lambda \sum_{i=1}^{n} \varphi(x_i^p) \tag{3}$$

Chen, Lingfeng Niu, Ya-xiang Yuan, Smoothing trust region Newton method

$$\min_{x \in R^n} \quad f(x) := \Theta(x) + \lambda \sum_{i=1}^{r} \varphi(|d_i^T x|^p) \tag{4}$$

Remark: Quadratic approximation at each step of the three algorithms. Cubic regularization with $O(\epsilon^{-\frac{3}{2}})$ for smooth nonconvex optimization: Nesterov-Polyak(2006) and Cartis-Gould-Toint(2012).

Smoothing function Approximations

$$s(t,\mu) = \begin{cases} |t| & \text{if } |t| \geq \mu \\ \dfrac{t^2}{2\mu} + \dfrac{\mu}{2} & \text{if } |t| < \mu \end{cases}$$

A Smoothing Function of f

$$\tilde{f}(x,\mu) = \Theta(x) + \sum_{i=1}^{n} \varphi(s^p(x_i,\mu)), \qquad \tilde{g}(x,\mu) := \nabla_x \tilde{f}(x,\mu)$$

Strictly Convex Quadratic function approximation around y

$$\tilde{f}(x,\mu) \leq Q(x,y,\mu) = \tilde{f}(y,\mu) + \langle \tilde{g}(y,\mu), x-y \rangle + \frac{1}{2}\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{d_i(y,\mu)}$$

$$\frac{1}{d_i(y,\mu)} = \begin{cases} \max\{\beta + 8\lambda p|\frac{y_i}{2}|^{p-2}, \dfrac{|\tilde{g}_i(y,\mu)|}{|\frac{y_i}{2}|^{1-\frac{p}{2}}\mu^{\frac{p}{2}}}\} & \text{if } |y_i| > 2\mu \\ \max\{\beta + 8\lambda p \mu^{p-2}, \dfrac{|\tilde{g}_i(y,\mu)|}{\mu}\} & \text{if } |y_i| \leq 2\mu. \end{cases}$$

SSQP Algorithm

Choose $x^0 \in R^n$, $\mu_0 > 0$ and $\sigma \in (0,1)$. Set $k = 0$ and $z^0 = x^0$.
For $k \geq 0$, set

$$x_i^{k+1} = x_i^k - d_i(x^k, \mu_k)\tilde{g}_i(x^k, \mu_k), \quad i = 1, \dots, n \tag{5a}$$

$$\mu_{k+1} = \begin{cases} \mu_k & \text{if } \tilde{f}(x^{k+1}, \mu_k) - \tilde{f}(x^k, \mu_k) \leq -4\alpha p \mu_k^p \\ \sigma\mu_k & \text{otherwise,} \end{cases} \tag{5b}$$

$$z^{k+1} = \begin{cases} x^k & \text{if } \mu_{k+1} = \sigma\mu_k \\ z^k & \text{otherwise.} \end{cases} \tag{5c}$$

In (5a) $$x^{k+1} = \arg\min_{x \in R^n} Q(x, x^k, \mu_k)$$

$$\nabla_x Q(x, x^k, \mu_k) = \tilde{g}(x^k, \mu_k) + \nabla_x^2 Q(x, x^k, \mu_k)(x - x^k),$$

$$\nabla_x^2 Q(x, x^k, \mu_k) = \mathrm{diag}\left(\frac{1}{d_1(x^k, \mu_k)}, \dots, \frac{1}{d_n(x^k, \mu_k)}\right) \succ 0$$

$$f(x) \leq \tilde{f}(x, \mu_k) \leq Q(x, x^k, \mu_k), \qquad x \in R^n$$

Worst-case complexity for SSQP

Definition 2 Let $G : R^n \to R^n$ be defined by

$$G(x) = X\nabla\Theta(x) + p|X|^p[\nabla\varphi(t)_{t=|x_i|^p}]_{i=1}^n,$$

where X =diag$(x_1, \ldots, x_n)$ and $|X|^p$ =diag$(|x_1|^p, \ldots, |x_n|^p)$. For a given $\epsilon \geq 0$, we call $x^* \in R^n$ an ϵ scaled first order stationary point of (2) if

$$\|G(x^*)\|_\infty \leq \epsilon.$$

And x^* is called a scaled first order stationary point of (2) if $\epsilon = 0$.

Worst-case complexity for SSQP

Definition 2 Let $G : R^n \to R^n$ be defined by

$$G(x) = X\nabla\Theta(x) + p|X|^p[\nabla\varphi(t)_{t=|x_i|^p}]_{i=1}^n,$$

where X =diag$(x_1, \dots, x_n)$ and $|X|^p$ =diag$(|x_1|^p, \dots, |x_n|^p)$. For a given $\epsilon \geq 0$, we call $x^* \in R^n$ an ϵ scaled first order stationary point of (2) if

$$\|G(x^*)\|_\infty \leq \epsilon.$$

And x^* is called a scaled first order stationary point of (2) if $\epsilon = 0$.

Definition 3 For $\epsilon \geq 0$, x^* is an ϵ global minimizer of (2) if

$$f(x^*) - \min_{x \in R^n} f(x) \leq \epsilon.$$

Theorem 7 For any $\epsilon \in (0, 1]$, the SSQP Algorithm obtains an ϵ scaled first order stationary point or ϵ global minimizer of (2) in no more than $O(\epsilon^{-2})$ steps.

Interior Point Algorithm $(0 < p \leq 1)$

$$\min_{x\in[0,u]} \quad f(x) := \Theta(x) + \lambda \sum_{i=1}^{n} \varphi(x_i^p), \qquad (3)$$

$\|\nabla^2\Theta(x)\| \leq \beta$. For any x, $x^+ \in (0, u]$, we obtain

$$\begin{aligned} f(x^+) &\leq f(x) + \langle \nabla f(x), x^+ - x\rangle + \frac{\beta}{2}\|x^+ - x\|^2 \\ &= f(x) + \langle X\nabla f(x), d_x\rangle + \frac{\beta}{2}\|Xd_x\|^2, \qquad (Xd_x = x^+ - x). \end{aligned}$$

Box constrained quadratic program

$$\begin{aligned} \min \quad & \langle X\nabla f(x), d_x\rangle + \frac{\beta}{2} d_x^T X^2 d_x \\ \text{s.t.} \quad & d_x^2 \leq \frac{1}{4}e_n, \quad d_x \leq X^{-1}(u - x). \end{aligned}$$

$$d_x = \mathrm{Proj}_{\mathcal{D}_x}[-\frac{1}{\beta}X^{-1}\nabla f(x)], \qquad \mathcal{D}_x = [-\frac{1}{2}e_n, \min\{\frac{1}{2}e_n, X^{-1}(u-x)\}]$$

Interior Point Algorithm ($0 < p \leq 1$)

Interior Point Algorithm

Choose $x^0 \in (0, u]$. For $k \geq 0$, set

$$d_k = \text{Proj}_{\mathcal{D}_k}[-\frac{1}{\beta} X_k^{-1} \nabla f(x^k)], \qquad x^{k+1} = x^k + X_k d_k$$

Definition 3a For $\epsilon \geq 0$, x^* is an ϵ global minimizer of (3) if

$$x^* \in [0, u] \quad \text{and} \quad f(x^*) - \min_{0 \leq x \leq u} f(x) \leq \epsilon.$$

Definition 4 For $\epsilon \geq 0$, x is an ϵ scaled first order stationary point of (3), if $x \in (0, u]$ and

1. $|[X \nabla f(x)]_i| \leq \epsilon$ if $x_i < u_i - \delta\epsilon$;
2. $[\nabla f(x)]_i \leq \epsilon$ if $x_i \geq u_i - \delta\epsilon$, where $\delta > 0$ is a small constant.

Theorem 8 For any $\epsilon \in (0, 1]$, the Interior Point Algorithm obtains an ϵ scaled first order stationary point or ϵ global minimizer of (3) in no more than $O(\epsilon^{-2})$ steps.

A special case of Problem (3)

$$\min_{x\geq 0} \quad f(x) := \Theta(x) + \lambda\|x\|_p^p, \qquad (3')$$

Second Order Interior Point Algorithm

For given $\epsilon \in (0, 1]$, choose $x^0 > 0$. For $k \geq 0$,

$$d_k \in \arg\min_{\|d\|^2 \leq \epsilon/\Gamma} d^T X_k \nabla f(x^k) + \frac{1}{2} d^T X_k \nabla^2 f(x^k) X_k d$$

$$x^{k+1} = x^k + X_k d_k$$

$\Gamma = (2\gamma\eta^3 + \lambda\eta^p)^2$, $\eta \geq \sup\{\|x\|_\infty : f(x) \leq f(x^0), x \geq 0\}$, $\|\nabla^3\Theta(x)\| \leq \gamma$.

Definition 4a For $\epsilon \geq 0$, $x > 0$ is an ϵ scaled second order stationary point of (3′), if

$$\|X\nabla f(x)\|_\infty \leq \epsilon \quad \text{and} \quad X\nabla^2 f(x)X \succeq -\sqrt{\epsilon} I.$$

Theorem 9 For any $\epsilon \in (0, 1]$, the Second Order Interior Point Algorithm obtains an ϵ scaled second order stationary point or ϵ global minimizer of (3′) in no more than $O(\epsilon^{-\frac{3}{2}})$ steps.

Smoothing trust region algorithm $(0 < p \leq 1)$

$$\min_{x \in R^n} \quad f(x) := \Theta(x) + \lambda \sum_{i=1}^{r} \varphi(|d_i^T x|^p), \tag{4}$$

For $\bar{x} \neq 0$, let $J_{\bar{x}} = \{i \mid d_i^T \bar{x} \neq 0, i = 1, \cdots, r\}$. Let $Z_{\bar{x}}$ be an $n \times \ell$ matrix whose columns are an orthonormal basis for the null space of $\{d_i \mid i \notin J_{\bar{x}}\}$. Let

$$w(x) := \Theta(x) + \lambda \sum_{i \in J_{\bar{x}}} \varphi(|d_i^T x|^p), \qquad (\, f(\bar{x}) = w(\bar{x})\,)$$

Theorem 9 (Second order necessary condition)
If $\bar{x}$ is an nonzero local minimizer of problem (4), then we have

$$Z_{\bar{x}}^T \nabla w(\bar{x}) = 0 \tag{6}$$

$$\forall\, v \in R^{\ell}, \text{ there is an } H \in \partial_C^2 w(\bar{x}), \text{ such that } v^T Z_{\bar{x}}^T H Z_{\bar{x}} v \geq 0. \tag{7}$$

Smoothing trust region algorithm $(0 < p \leq 1)$

$$\min_{x \in R^n} \quad f(x) := \Theta(x) + \lambda \sum_{i=1}^{r} \varphi(|d_i^T x|^p), \qquad (4)$$

For $\bar{x} \neq 0$, let $J_{\bar{x}} = \{i \mid d_i^T \bar{x} \neq 0, i = 1, \cdots, r\}$. Let $Z_{\bar{x}}$ be an $n \times \ell$ matrix whose columns are an orthonormal basis for the null space of $\{d_i \mid i \notin J_{\bar{x}}\}$. Let

$$w(x) := \Theta(x) + \lambda \sum_{i \in J_{\bar{x}}} \varphi(|d_i^T x|^p), \qquad (\, f(\bar{x}) = w(\bar{x}) \,)$$

Theorem 9 (Second order necessary condition)
If $\bar{x}$ is an nonzero local minimizer of problem (4), then we have

$$Z_{\bar{x}}^T \nabla w(\bar{x}) = 0 \qquad (6)$$

$$\forall\, v \in R^{\ell}, \text{ there is an } H \in \partial_C^2 w(\bar{x}), \text{ such that } v^T Z_{\bar{x}}^T H Z_{\bar{x}} v \geq 0. \qquad (7)$$

Theorem 10 (Second order sufficient condition)
If (6) holds and $Z_{\bar{x}}^T H Z_{\bar{x}} \succ 0, \forall H \in \partial_C^2 w(\bar{x})$, then $\bar{x}$ is a strictly local minimizer of (4).

Smoothing trust region algorithm $(0 < p \leq 1)$

$$s(t,\mu) = \sqrt{t^2 + 4\mu^2} \qquad \text{a smoothing function of } |t|$$

A Smoothing Function of f

$$\tilde{f}(x,\mu) = \Theta(x) + \sum_{i=1}^{r} \varphi(s^p(d_i^T x, \mu))$$

Smoothing Trust Region Algorithm

Choose $x^0 \in R^n$, $\mu_0 > 0, \Delta_0, \underline{\Delta}, \zeta > 0$, $\nu \in (0,1)$. For $k \geq 0$,

Step 1. $\min \; d^T \nabla \tilde{f}(x^k, \mu_k) + \frac{1}{2} d^T \nabla^2 \tilde{f}(x^k, \mu_k) d$

s.t. $\|d\| \leq \Delta_k$

Step 2 Update x^k and Δ_k to get x^{k+1} and Δ_{k+1}

Step 3 If $\|\nabla \tilde{f}(x^k, \mu_k)\| \leq \zeta \mu_k$ and $\Delta_k \geq \underline{\Delta}$, choose $\mu_{k+1} = \nu \mu_k$; otherwise, set $\mu_{k+1} = \mu_k$.

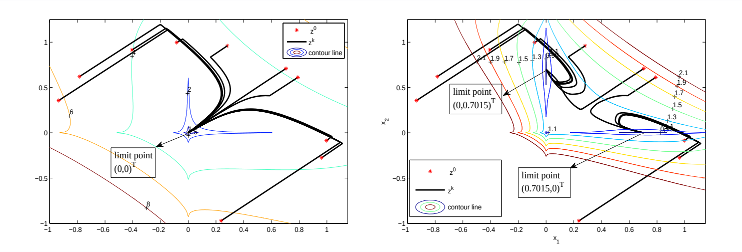
Theorem 11 Any accumulation point of $\{x^k\}$ generated by the Smoothing Trust Region Algorithm satisfies the second order necessary conditions (6)-(7).

Example 1: SSQP Algorithm

$$\min_{x \in R^2} \quad f(x) := (x_1 + x_2 - 1)^2 + \lambda(\sqrt{|x_1|} + \sqrt{|x_2|}). \tag{8}$$

λ	global minimizer	global minimum
$\frac{8}{3\sqrt{3}}$	$(0, 0)$	1
1	$(0, 0.7015)$ and $(0.7015, 0)$	0.927

When $\lambda = \frac{8}{3\sqrt{3}}$, $(1/3, 0)$ and $(0, 1/3)$ are two nonzero vectors satisfying the first and second order necessary conditions.



(a) $\lambda = \frac{8}{3\sqrt{3}}$ (b) $\lambda = 1$

Example 2: Prostate cancer

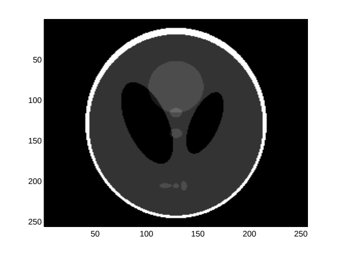
- This data sets are from the UCI Standard database.
- The data set consists of the medical records of 97 patients who were about to receive a radical prostatectomy. The predictors are 8 clinical measures: lcavol, lweight, age, lbph, svi,lcp, gleason and pgg45.
- The aim is to find few main factors with small prediction error.
- A training set with 67 observations, and a test set with 30 observations, $A \in R^{67\times 8}$, $b \in R^{67}$.

Results for prostate cancer

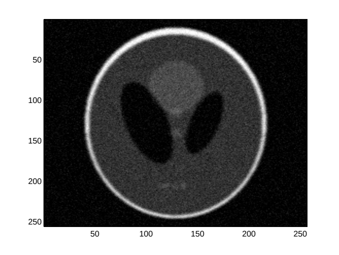
Parameter	LASSO	IRL1	OMP-SCG	SSQP	STR	FIP
x_1(lcavol)	0.545	0.6187	0.6436	0.6437	0.646	0.6433
x_2(lweight)	0.237	0.2362	0.2804	0.2765	0.275	0.2767
x_3(lage)	0	0	0	0	0	0
x_4(lbph)	0.098	0.1003	0	0	0	0
x_5(svi)	0.165	0.1858	0.1857	0.1327	0.128	0.1337
x_6(lcp)	0	0	0	0	0	0
x_7(gleason)	0	0	0	0	0	0
x_8(pgg45)	0.059	0	0	0	0	0
$\|x\|_0$	5	4	3	3	3	3
Prediction error	0.478	0.468	0.4419	0.4264	0.428	0.426

IRL1: Iterative reweighted ℓ_1 norm, OMP-SCG: Orthogonal matching pursuit
STR: Smoothing trust region,   FIP: First order interior point

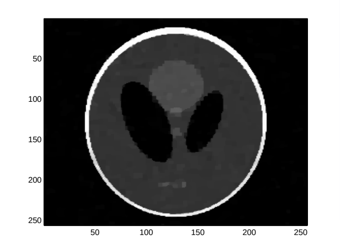
Example 3 Image restoration: $\sum_{i=1}^{r} \|d_i^T x\|_p^p$, $0 \le x \le e$



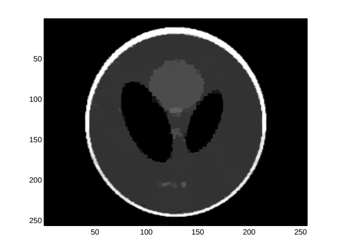
(c) Original image



(d) Observed image

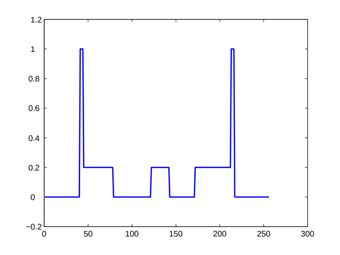


(e) restored $\varphi(t) = \frac{0.5|t|}{1+0.5|t|}$

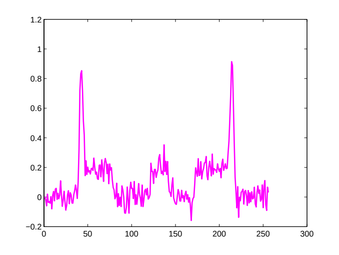


(f) restored $\varphi(t) = |t|^{0.5}$

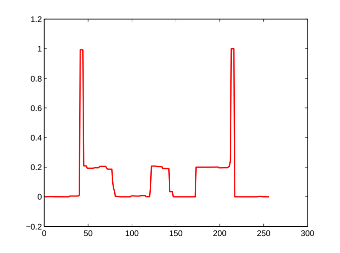
The restored 126th line for the Shepp-Logan image of size 256×256



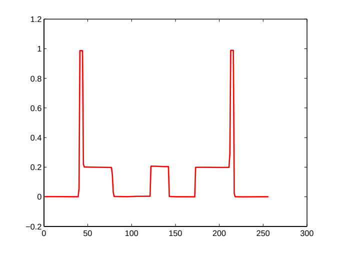
(g) original



(h) observed



(i) restored $\varphi(t) = \frac{0.5|t|}{1+0.5|t|}$



(j) restored $\varphi(t) = |t|^{0.5}$

Thank You

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學
DEPARTMENT OF APPLIED MATHEMATICS
應 用 數 學 系