

# Two Affine Scaling Methods for Solving Optimization Problems Regularized with an $\mathcal{L}_1$ -norm

Zhirong Li<sup>1</sup>

<sup>1</sup>David R. Cheriton, School of Computer Science

September 16, 2010

# Outline

- 1 Background
- 2 Motivations
- 3 Our Affine Scaling Gradient Based Algorithm
- 4 Convergence Proof for the Trust Region Method [3]
- 5 Concluding Remarks

# Introduction

- Import applications of optimization problems with the  $\mathcal{L}_1$ -norm regularization, e.g., Volatility Surface Calibration in finance
- Problem Formulation as a Least Squares of Price Errors with  $\mathcal{L}_1$  penalty
- Role of  $\mathcal{L}_1$  Regularization: sparsity, stability at solution
- Typical solutions: Steepest Descent, Newton Type

# Introduction

- Import applications of optimization problems with the  $\mathcal{L}_1$ -norm regularization, e.g., Volatility Surface Calibration in finance
- Problem Formulation as a Least Squares of Price Errors with  $\mathcal{L}_1$  penalty
- Role of  $\mathcal{L}_1$  Regularization: sparsity, stability at solution
- Typical solutions: Steepest Descent, Newton Type

# Introduction

- Import applications of optimization problems with the  $\mathcal{L}_1$ -norm regularization, e.g., Volatility Surface Calibration in finance
- Problem Formulation as a Least Squares of Price Errors with  $\mathcal{L}_1$  penalty
- Role of  $\mathcal{L}_1$  Regularization: sparsity, stability at solution
- Typical solutions: Steepest Descent, Newton Type

# Introduction

- Import applications of optimization problems with the  $\mathcal{L}_1$ -norm regularization, e.g., Volatility Surface Calibration in finance
- Problem Formulation as a Least Squares of Price Errors with  $\mathcal{L}_1$  penalty
- Role of  $\mathcal{L}_1$  Regularization: sparsity, stability at solution
- Typical solutions: Steepest Descent, Newton Type

# $\mathcal{L}_1$ regularization

- Challenges posed by  $\mathcal{L}_1$  Regularization: non-differentiability, essentially a constrained problem

## Weakness of typical monotonic decreasing algorithm

- Steepest Descent: slow convergence, exact line search
- Newton Type: computational cost
- Non-monotone algorithm: simplicity in stepsize calculation, e.g., Barzilai-Borwein (BB) method [1] for quadratic minimization
- Challenges for BB-type gradient based algorithm: divergence with  $\mathcal{L}_1$  penalty added

# $\mathcal{L}_1$ regularization

- Challenges posed by  $\mathcal{L}_1$  Regularization: non-differentiability, essentially a constrained problem

## Weakness of typical monotonic decreasing algorithm

- Steepest Descent: slow convergence, exact line search
- Newton Type: computational cost
- Non-monotone algorithm: simplicity in stepsize calculation, e.g., Barzilai-Borwein (BB) method [1] for quadratic minimization
- Challenges for BB-type gradient based algorithm: divergence with  $\mathcal{L}_1$  penalty added



# $\mathcal{L}_1$ regularization

- Challenges posed by  $\mathcal{L}_1$  Regularization: non-differentiability, essentially a constrained problem

## Weakness of typical monotonic decreasing algorithm

- Steepest Descent: slow convergence, exact line search
- Newton Type: computational cost
- Non-monotone algorithm: simplicity in stepsize calculation, e.g., Barzilai-Borwein (BB) method [1] for quadratic minimization
- Challenges for BB-type gradient based algorithm: divergence with  $\mathcal{L}_1$  penalty added

# $\mathcal{L}_1$ regularization

- Challenges posed by  $\mathcal{L}_1$  Regularization: non-differentiability, essentially a constrained problem

## Weakness of typical monotonic decreasing algorithm

- Steepest Descent: slow convergence, exact line search
- Newton Type: computational cost
- Non-monotone algorithm: simplicity in stepsize calculation, e.g., Barzilai-Borwein (BB) method [1] for quadratic minimization
- Challenges for BB-type gradient based algorithm: divergence with  $\mathcal{L}_1$  penalty added

# Motivations

## Deal with divergence of BB-type Method

- Affine Scaling: handle  $\mathcal{L}_1$ -norm, avoid non-differentiable points
- Globalization: line search and trust region to ensure convergence

- KKT conditions for  $\min_{x \in \mathbb{R}^n} f(x) + \|x\|_1$ :

$$D(x) \cdot (\nabla f(x) + \text{sign}(x)) = 0$$

where

$$D(x)_{i,i} := \begin{cases} 1 & |(\nabla f(x))_i| > 1 \\ |x_i| & |(\nabla f(x))_i| \leq 1 \end{cases}$$

# Motivations

## Deal with divergence of BB-type Method

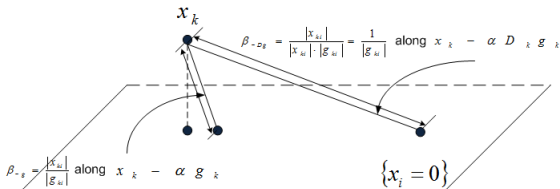
- Affine Scaling: handle  $\mathcal{L}_1$ -norm, avoid non-differentiable points
- Globalization: line search and trust region to ensure convergence
- KKT conditions for  $\min_{x \in \mathbb{R}^n} f(x) + \|x\|_1$ :

$$D(x) \cdot (\nabla f(x) + \text{sign}(x)) = 0$$

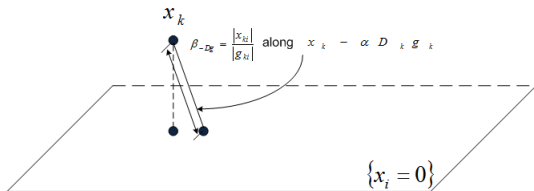
where

$$D(x)_{i,i} := \begin{cases} 1 & |(\nabla f(x))_i| > 1 \\ |x_i| & |(\nabla f(x))_i| \leq 1 \end{cases}$$

# Role of Affine Scaling



Case a:  $(D_k)_{i,i} = |x_{ki}|$  when  $|\nabla f(x_k)_i| \leq 1$



Case b:  $(D_k)_{i,i} = 1$  when  $|\nabla f(x_k)_i| > 1$

# Challenges and Solutions

## Challenges for Non-monotone Gradient Based Algorithm

- What search direction?
- What stepsize along the search direction?
- What type of line search to ensure convergence?

## Solutions

- Search direction: Scaled Steepest Descent Direction
- Stepsize: Safeguarded Barzilai-Borwein Stepsize
- Line Search: Adaptive Line Search with Non-monotone Armijo-Rule check

# Challenges and Solutions

## Challenges for Non-monotone Gradient Based Algorithm

- What search direction?
- What stepsize along the search direction?
- What type of line search to ensure convergence?

## Solutions

- Search direction: Scaled Steepest Descent Direction
- Stepsize: Safeguarded Barzilai-Borwein Stepsize
- Line Search: Adaptive Line Search with Non-monotone Armijo-Rule check

# Search Direction: Scaling Matrix

We use the new scaling matrix

$$D(x) := \text{diag}(v(x))$$

where

$$(v(x))_i := \begin{cases} 1 & |(\nabla f(x))_i| > 1 \\ \min\{|x_i|, 1\} & |(\nabla f(x))_i| \leq 1 \end{cases}$$

By doing so, we avoid adding scaling effect when the iterate is not close to the optimal solution yet. We take  $-D_k g_k$  as the search direction.



# Search Direction: Scaling Matrix

We use the new scaling matrix

$$D(x) := \text{diag}(v(x))$$

where

$$(v(x))_i := \begin{cases} 1 & |(\nabla f(x))_i| > 1 \\ \min\{|x_i|, 1\} & |(\nabla f(x))_i| \leq 1 \end{cases}$$

By doing so, we avoid adding scaling effect when the iterate is not close to the optimal solution yet. We take  $-D_k g_k$  as the search direction.

# Stepsize: BB-type Stepsize

BB-type stepsize

$$\alpha_k^{BB1} := \frac{\langle D_k \Delta x_k, D_k \Delta x_k \rangle}{\langle D_k \Delta x_k, D_k \Delta g_k \rangle}$$

which is derived as the solution to

$$\min_{\alpha \in \mathbb{R}^n} \left\| \frac{1}{\alpha} D_k \Delta x_k - D_k \Delta g_k \right\|^2$$

$\alpha_k^{BB1}$  should be in  $[\alpha_{\min}, \alpha_{\max}]$  where  $0 < \alpha_{\min} < 1 < \alpha_{\max}$  and the iterates are updated by  $x_{k+1} = x_k - \alpha_k^{BB} D_k g_k$

## Stepsize: BB-type Stepsize

BB-type stepsize

$$\alpha_k^{BB1} := \frac{\langle D_k \Delta x_k, D_k \Delta x_k \rangle}{\langle D_k \Delta x_k, D_k \Delta g_k \rangle}$$

which is derived as the solution to

$$\min_{\alpha \in \mathbb{R}^n} \left\| \frac{1}{\alpha} D_k \Delta x_k - D_k \Delta g_k \right\|^2$$

$\alpha_k^{BB1}$  should be in  $[\alpha_{\min}, \alpha_{\max}]$  where  $0 < \alpha_{\min} < 1 < \alpha_{\max}$  and the iterates are updated by  $x_{k+1} = x_k - \alpha_k^{BB} D_k g_k$

# Stepsize: Safeguard Mechanism

## Observations

- Stepsize  $\liminf_{k \rightarrow \infty} \alpha_k^{\overline{BB}} > 0$ : otherwise may stop at non-optimal solution
- Negative Stepsize: Use Rayleigh Quotient type stepsize

$$\frac{\langle D_k \Delta x_k, D_k \Delta x_k \rangle}{\langle D_k \Delta x_k, D_k H \Delta x_k \rangle}$$

- Upper bound on the stepsize:  $1 < \alpha_{\max}$  since unit step is effective when  $\mathcal{L}_1$  term dominates

# Stepsize: Safeguard Mechanism

## Observations

- Stepsize  $\liminf_{k \rightarrow \infty} \alpha_k^{\overline{BB}} > 0$ : otherwise may stop at non-optimal solution
- Negative Stepsize: Use Rayleigh Quotient type stepsize

$$\frac{\langle D_k \Delta x_k, D_k \Delta x_k \rangle}{\langle D_k \Delta x_k, D_k H \Delta x_k \rangle}$$

- Upper bound on the stepsize:  $1 < \alpha_{\max}$  since unit step is effective when  $\mathcal{L}_1$  term dominates

# Stepsize: Safeguard Mechanism

## Observations

- Stepsize  $\liminf_{k \rightarrow \infty} \alpha_k^{\overline{BB}} > 0$ : otherwise may stop at non-optimal solution
- Negative Stepsize: Use Rayleigh Quotient type stepsize

$$\frac{\langle D_k \Delta x_k, D_k \Delta x_k \rangle}{\langle D_k \Delta x_k, D_k H \Delta x_k \rangle}$$

- Upper bound on the stepsize:  $1 < \alpha_{\max}$  since unit step is effective when  $\mathcal{L}_1$  term dominates

# Line Search: Non-monotone Armijo-Rule Condition Check

Shrinking factor  $\theta$  introduced to meet the decrease requirement relative to the reference function value

$$h\left(x_k - \theta \alpha_k^{\overline{BB}} D_k g_k\right) \leq h_{\max} + \gamma \theta \left\langle g_k, -\alpha_k^{\overline{BB}} D_k g_k \right\rangle$$

where

$$h_{\max} := \max \left\{ h\left(x_{k-j}\right) \mid 0 \leq j \leq \min \{k, M-1\} \right\}$$

## Our Affine Scaling Gradient Based Algorithm

---

**Algorithm 3.1** Scaled Gradient Method

---

Given  $x_0, \alpha_0, \alpha_{\min}, \alpha_{\max}, 0 < \tau_1 < \tau_2 < 1, \gamma \in (0, 1)$  and  $M \in \mathbb{Z}_+$

Step 1: If  $\left| h\left(x_k - \theta \alpha_k^{\overline{BB}} D_k g_k\right) - h\left(x_k\right) \right| < tol$  stop

Step 2: Calculate  $h_{\max} \triangleq \max \{h(x_{k-j}) \mid 0 \leq j \leq \min\{k, M-1\}\}$  and

$$\alpha_k^{\overline{BB}} = \frac{\langle D_k(x_k - x_{k-1}), D_k(x_k - x_{k-1}) \rangle}{\langle D_k(x_k - x_{k-1}), D_k([H(x_k - x_{k-1}) + \text{sign}(x_k) - \text{sign}(x_{k-1}))]) \rangle}$$

Step 3: If  $\alpha_k^{\overline{BB}} < 0$  then set  $\alpha_k^{\overline{BB}} = \frac{\langle D_k \Delta x_k, D_k \Delta x_k \rangle}{\langle D_k \Delta x_k, D_k H \Delta x_k \rangle}$ ;

$$\text{set } \alpha_k^{\overline{BB}} = \min \left\{ \alpha_{\max}, \max \left\{ \alpha_{\min}, \alpha_k^{\overline{BB}} \right\} \right\}$$

Step 4: Compute  $\delta \leftarrow \langle g_k, -\alpha_k^{\overline{BB}} D_k g_k \rangle$  and set  $\theta \leftarrow 1$

Step 5: While  $h\left(x_k - \theta \alpha_k^{\overline{BB}} D_k g_k\right) > h_{\max} + \gamma \theta \delta$

$$\{\text{set } \theta_{\text{new}} \in [\tau_1 \theta, \tau_2 \theta], \theta \leftarrow \theta_{\text{new}}\}$$

Step 6:  $x_{k+1} = x_k - \theta \alpha_k^{\overline{BB}} D_k g_k$  and goto step 1

---



## Performance Evaluation

This SG algorithm only requires the first-order information and it converges fast in all 3 scenarios where the quadratic term is negligible ( $\rho \leq 1$ ), comparable ( $\rho \approx 10$ ), dominant ( $\rho \geq 100$ ). 50 test problems for each  $\rho$ .

	$\rho = 0.1$	$\rho = 1$	$\rho = 10$	$\rho = 100$
SD	163/0.0%	158/0.0%	285/0.0%	345/16%
SSD	160/100%	126/100%	160/100%	300/100%

**Table:** Average Number of Iterations / Success Rate for SD and SSD direction

## The SPG Method

---

**Algorithm 3.2** Spectral Projected Gradient Method [9]

---

Given  $x_0, \alpha_0, \alpha_{\min}, \alpha_{\max}, 0 < \tau_1 < \tau_2 < 1, \gamma \in (0, 1)$  and  $M \in \mathbb{Z}_+$ Step 1: If  $\left| f\left(x_k - \theta \alpha_k^{\overline{BB}} g_k\right) - f(x_k) \right| < \text{tol}$  stop

Step 2: Calculate

$$\alpha_k^{\overline{BB}} = \min \left\{ \alpha_{\max}, \max \left\{ \alpha_{\min}, \frac{\langle x_k - x_{k-1}, x_k - x_{k-1} \rangle}{\langle x_k - x_{k-1}, H(x_k - x_{k-1}) \rangle} \right\} \right\}$$

and

$$\text{Proj}(x_k - \alpha_k^{\overline{BB}} g_k)$$

where  $\text{Proj}(x)$  is the orthogonal projection of  $x$  onto the boundary of the feasible region  $\mathcal{F} := \{x : \|x\|_1 \leq t\}$ .

Step 3: Calculate

$$f_{\max} \triangleq \max \{f(x_{k-j}) \mid 0 \leq j \leq \min\{k, M-1\}\}$$

and

$$\delta \leftarrow \langle g_k, \text{Proj}(x_k - \alpha_k^{\overline{BB}} g_k) - x_k \rangle$$

and set  $\theta \leftarrow 1$ Step 4: While  $f\left(x_k - \theta \alpha_k^{\overline{BB}} g_k\right) > f_{\max} + \gamma \theta \delta$  $\{\text{set } \theta_{\text{new}} \in [\tau_1 \theta, \tau_2 \theta], \theta \leftarrow \theta_{\text{new}}\}$ Step 5:  $x_{k+1} = x_k - \theta \alpha_k^{\overline{BB}} g_k$  and goto step 1

---

# Performance Comparison

	1	2	3	4	5	6	7	8	9	Average #
SPG	181	257	212	202	233	169	234	227	294	224
SG	329	258	258	279	449	321	428	284	325	326

**Table:** Number of Iterations for SPG and SG Method, for 9 problems that SPG is successful

We run 50 test cases and 41 out of 50 failed to converge for the SPG method.

## Performance Comparison Cont.

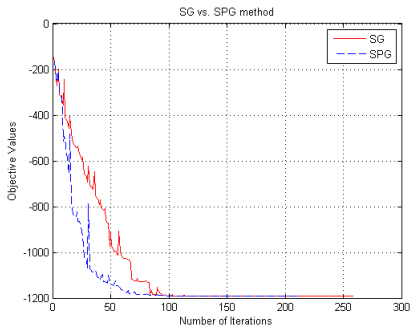


Figure: Trajectory of Objective Values for Test # 20, SG vs. SPG Method

# Definitions

For simplicity, we define our scaling matrix as  $D(x) := \text{diag}(v(x))$  where

$$(v(x))_i := \begin{cases} 1 & i = j^* := \operatorname{argmax}_{j \in \{1, \dots, n\} \wedge |(\nabla f(x))_j| > 1} (|(\nabla f(x))_j|) \\ |x_i| & \text{otherwise} \end{cases}$$

and the piecewise quadratic approximation model as

$$\phi_k(d) = \nabla f(x_k)^T d + \|x_k + d\|_1 - \|x_k\|_1 + \frac{1}{2} d^T M_k d$$

and the optimal decrease along  $d_k$  as

$$\phi_k^*[d_k] := \min \left\{ \phi_k(\alpha d_k) : \left\| \alpha D_k^{-\frac{1}{2}} d_k \right\|_2 \leq \Delta_k, 0 \leq \alpha \leq \beta_k^2 \right\}$$

## Trust Region Method: Part I - Step Calculation

---

**Algorithm 4.1** Affine Scaling Trust Region Algorithm

---

Let  $0 < \mu < \eta < 1$  and  $x_0$ . For  $k = 0, 1, \dots$ Step 1. Compute  $f(x_k), g_k, D_k, M_k$  and  $C_k$ ; define the quadratic model as

$$\psi_k(d) := g_k^T d + \frac{1}{2} d^T M_k d.$$

Step 2. Compute a step  $d_k$  such that  $x_k + d_k \in \text{dif}\{\mathcal{F}\}$ ,  
based on the sub-problem

$$\min_{d \in \mathbb{R}^n} \left\{ \psi_k(d) : \|D_k^{-\frac{1}{2}} d\|_2 \leq \Delta_k \right\}$$

where  $\text{dif}\{\mathcal{F}\}$  is defined to be the union of the region in  $\mathcal{F}$  that is  
differentiable.

Step 3. Compute

$$\rho_k^f := \frac{f(x_k + d_k) - f(x_k) + \|x_k + d_k\|_1 - \|x_k\|_1 + \frac{1}{2} d_k^T C_k d_k}{\phi_k(d_k)}.$$

Step 4. If  $\rho_k^f > \mu$ , then set  $x_{k+1} = x_k + d_k$ . Otherwise set  $x_{k+1} = x_k$ .Step 5. Update the trust region size  $\Delta_k$  as specified below.

# Trust Region Algorithm: Part II - Region Size Update

---

**Algorithm 4.1** Affine Scaling Trust Region Algorithm

---

Updating trust region size  $\Delta_k$

Let  $0 < \gamma_1 < 1 < \gamma_2$  and  $\Lambda_l > 0$  be given.

Step 1. If  $\rho_k^f < \mu$ , then set  $\Delta_{k+1} \in (0, \gamma_1 \Delta_k]$ .

Step 2. If  $\mu < \rho_k^f < \eta$ , then set  $\Delta_{k+1} \in [\gamma_1 \Delta_k, \Delta_k]$ .

Step 3. If  $\rho_k^f \geq \eta$  then

    If  $\Delta_k > \Lambda_l$  then

        set  $\Delta_{k+1} \in$  either  $[\gamma_1 \Delta_k, \Delta_k]$  or  $[\Delta_k, \gamma_2 \Delta_k]$

    Otherwise

        set  $\Delta_{k+1} \in [\Delta_k, \gamma_2 \Delta_k]$ .

---

# Proof Highlights

“The basic idea about the proof is that if the necessary optimality condition is violated, we can show that our algorithm will asymptotically achieve a sufficient decrease, which is bounded away from zero. Thus the objective value will go down to  $-\infty$ . This will lead to the contradiction of our Assumption 1 that the level set of  $f(x) + \|x\|_1$  is compact over  $\mathcal{F}$ .”



# Assumptions

**Assumption 1:** Given an initial point  $x_0 \in \mathbb{R}^n$  and assume that  $(x_0)_i \neq 0, \forall i \in \{1, \dots, n\}$ , the level set  $\mathcal{F} := \{x : h(x) \leq h(x_0)\}$  is compact.

**Assumption 2:**  $\{B_k = \nabla^2 f(x_k)\}$  is bounded. That is, there exists a positive scalar  $\chi_B$  such that  $\|B_k\| \leq \chi_B, \forall k$ .

**Assumption 3:** There exists a positive scalar  $\chi_f$  such that  $\|\nabla f(x)\|_\infty < \chi_f, \forall x \in \mathcal{F}$ .

**Assumption 4:** Assume that

$\phi(d_k) < \beta_g \phi_k^*[-D_k g_k], \left\| D_k^{-\frac{1}{2}} d_k \right\|_2 \leq \Delta_k, x_k + d_k \in \text{dif}(\mathcal{F})$  where  $\beta_g > 0$ .

# Assumptions

**Assumption 1:** Given an initial point  $x_0 \in \mathbb{R}^n$  and assume that  $(x_0)_i \neq 0, \forall i \in \{1, \dots, n\}$ , the level set  $\mathcal{F} := \{x : h(x) \leq h(x_0)\}$  is compact.

**Assumption 2:**  $\{B_k = \nabla^2 f(x_k)\}$  is bounded. That is, there exists a positive scalar  $\chi_B$  such that  $\|B_k\| \leq \chi_B, \forall k$ .

**Assumption 3:** There exists a positive scalar  $\chi_f$  such that  $\|\nabla f(x)\|_\infty < \chi_f, \forall x \in \mathcal{F}$ .

**Assumption 4:** Assume that

$\phi(d_k) < \beta_g \phi_k^*[-D_k g_k], \left\| D_k^{-\frac{1}{2}} d_k \right\|_2 \leq \Delta_k, x_k + d_k \in \text{dif}(\mathcal{F})$  where  $\beta_g > 0$ .

# Assumptions

**Assumption 1:** Given an initial point  $x_0 \in \mathbb{R}^n$  and assume that  $(x_0)_i \neq 0, \forall i \in \{1, \dots, n\}$ , the level set  $\mathcal{F} := \{x : h(x) \leq h(x_0)\}$  is compact.

**Assumption 2:**  $\{B_k = \nabla^2 f(x_k)\}$  is bounded. That is, there exists a positive scalar  $\chi_B$  such that  $\|B_k\| \leq \chi_B, \forall k$ .

**Assumption 3:** There exists a positive scalar  $\chi_f$  such that  $\|\nabla f(x)\|_\infty < \chi_f, \forall x \in \mathcal{F}$ .

**Assumption 4:** Assume that

$\phi(d_k) < \beta_g \phi_k^*[-D_k g_k], \left\| D_k^{-\frac{1}{2}} d_k \right\|_2 \leq \Delta_k, x_k + d_k \in \text{dif}(\mathcal{F})$  where  $\beta_g > 0$ .

# Assumptions

**Assumption 1:** Given an initial point  $x_0 \in \mathbb{R}^n$  and assume that  $(x_0)_i \neq 0, \forall i \in \{1, \dots, n\}$ , the level set  $\mathcal{F} := \{x : h(x) \leq h(x_0)\}$  is compact.

**Assumption 2:**  $\{B_k = \nabla^2 f(x_k)\}$  is bounded. That is, there exists a positive scalar  $\chi_B$  such that  $\|B_k\| \leq \chi_B, \forall k$ .

**Assumption 3:** There exists a positive scalar  $\chi_f$  such that  $\|\nabla f(x)\|_\infty < \chi_f, \forall x \in \mathcal{F}$ .

**Assumption 4:** Assume that

$\phi(d_k) < \beta_g \phi_k^*[-D_k g_k], \left\| D_k^{-\frac{1}{2}} d_k \right\|_2 \leq \Delta_k, x_k + d_k \in \text{dif}(\mathcal{F})$  where  $\beta_g > 0$ .

## Lemma 4.1

## Corollary

Assume assumption 1 – 3 hold and  $d_k$  satisfies assumption 4, let  $\chi$  be the minimum

$$\chi(\mu_k, \Delta_k, \beta_k^{i^*}) = \min \left\{ \frac{\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \rangle^2}{\mu_k \|\hat{g}_k^0\|^2}, (\Delta_k - \beta_k^{i^*}) \frac{\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \rangle}{\|\hat{g}_k^0\|}, (\beta_k^{i^*+1} - \beta_k^{i^*}) \frac{\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \rangle}{\|\hat{g}_k^0\|} \right\}$$

then

$$-\phi(d_k) \geq -\beta_g \phi_k^* [-D_k g_k^0] \geq \frac{\beta_g}{2} \left\{ \sum_{j=1}^{i^*} (\beta_k^j - \beta_k^{j-1}) \frac{\langle \hat{g}_k^{j-1}, \hat{g}_k^0 \rangle}{\|\hat{g}_k^0\|} + \chi(\mu_k, \Delta_k, \beta_k^{i^*}) \right\}$$

where  $i^*$  is the last break point along direction  $-D_k^{\frac{1}{2}} \frac{\hat{g}_k^0}{\|\hat{g}_k^0\|}$  that is crossed,  $i^* \in \{0, 1\}$ ,  
i.e.,

$$\beta_k^{i^*} := \max \{ \beta_k^i : \beta_k^i < \alpha^* \}$$

## Lemma 4.3

## Corollary

*Assume that  $\{\Delta_k\}$  is updated by above trust region update algorithm. If  $\rho_k^f \geq \eta$  for sufficient large  $k$ , then  $\{\Delta_k\}$  is bounded away from zero.*

This lemma provides a property of the trust region size when asymptotically the step is always successful. This is not surprising since we always expand the trust region when the approximation is good by the trust region size update rule.

## Lemma 4.4

## Corollary

Assume that  $f : \mathbb{R}^n \mapsto \mathbb{R}$  is continuously differentiable on  $\text{dif}(\mathcal{F})$  and assumptions 1 ~ 3 hold. Assume  $\liminf_{k \rightarrow \infty} \|\hat{g}_k^0\| > 0$  and strict complementarity condition holds, then there exists an  $\bar{\epsilon} > 0$  such that

$$\liminf_{k \rightarrow \infty} \beta_k^{i^*} = 0$$

and

$$\liminf_{k \rightarrow \infty} \frac{\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \rangle}{\|\hat{g}_k^0\|} \geq \bar{\epsilon} > 0.$$

This lemma will be used to show that, if optimality condition  $\liminf_{k \rightarrow \infty} \|\hat{g}_k^0\| = 0$  is violated, the algorithm will asymptotically achieve a sufficient decrease which is bounded away from zero.

## Theorem 4.5

## Theorem

Assume that  $f : \mathbb{R}^n \mapsto \mathbb{R}$  is continuously differentiable on  $\mathcal{F}$  and assumptions 1 ~ 3 hold. If  $\{d_k\}$  generated by Algorithm 4.1 satisfies assumption 4 and at every limit point of  $\{x_k\}_{k=1}^{\infty}$ , strict complementarity holds, then

$$\liminf_{k \rightarrow \infty} \|\hat{g}_k^0\| = 0$$

This theorem proves that there exists a subsequence of  $\{x_k\}$  such that the norm of the scaled gradient is approaching zero.



## Theorem 4.6

## Theorem

Assume assumption 1 ~ 3 and strict complementarity condition hold, and  $\nabla f(x)$  is uniformly continuous on  $\mathcal{F}$ . If  $\{x_k\}$  is generated by algorithm 4.1 and assumption 4 also holds for  $d_k$ , then

$$\lim_{k \rightarrow \infty} \|\hat{g}_k^0\| = \lim_{k \rightarrow \infty} \left\| D_k^{\frac{1}{2}} g_k^0 \right\| = 0$$

This theorem proves that the norm of the scaled gradient for every subsequence of  $\{x_k\}$  is approaching zero, i.e., the affine scaling trust region algorithm terminates with first -order necessary optimality condition being satisfied.

# Conclusions

- The scaled steepest descent method in Chapter 3 is simple, but it may be ineffective for nonconvex and very nonlinear problems.
- If a sufficient decrease is obtained along a scaled descent direction, it is theorized that the first order necessary optimality condition holds.
- If a sufficient decrease along the global solution to the trust region sub-problem is derived, asymptotically speaking, the second-order necessary optimality condition and superlinear convergence are expected to be achieved.

# Conclusions

- The scaled steepest descent method in Chapter 3 is simple, but it may be ineffective for nonconvex and very nonlinear problems.
- If a sufficient decrease is obtained along a scaled descent direction, it is theorized that the first order necessary optimality condition holds.
- If a sufficient decrease along the global solution to the trust region sub-problem is derived, asymptotically speaking, the second-order necessary optimality condition and superlinear convergence are expected to be achieved.

# Conclusions

- The scaled steepest descent method in Chapter 3 is simple, but it may be ineffective for nonconvex and very nonlinear problems.
- If a sufficient decrease is obtained along a scaled descent direction, it is theorized that the first order necessary optimality condition holds.
- If a sufficient decrease along the global solution to the trust region sub-problem is derived, asymptotically speaking, the second-order necessary optimality condition and superlinear convergence are expected to be achieved.

# Contributions and Outlook

## Contributions are two-fold

- investigate an efficient non-monotone method using gradient and appropriate affine scaling for minimizing a nonlinear function with the  $\mathcal{L}_1$ -norm regularization
- analyze and establish convergence properties of Coleman, Li and Wang's trust region method [3]

## Outlook

- solve real life volatility surface calibration problem
- establish the proof of the convergence for the proposed scaled gradient method

# Contributions and Outlook




## Contributions are two-fold

- investigate an efficient non-monotone method using gradient and appropriate affine scaling for minimizing a nonlinear function with the  $\mathcal{L}_1$ -norm regularization
- analyze and establish convergence properties of Coleman, Li and Wang's trust region method [3]

## Outlook

- solve real life volatility surface calibration problem
- establish the proof of the convergence for the proposed scaled gradient method

# For Further Reading I

-  Barzilai, J. and Borwein, J.M. , “Two Point Step Size Gradient Methods,” in IMA Journal of Numerical Analysis Vol.8, pp.141 – 148, 1988
-  T.F. Coleman and Y. Li, ”An Interior, Trust Region Approach for Nonlinear Minimization Subject to Bounds”, SIAM Journal on Optimization, Vol. 6. No 2, May 1996, pp. 418 – 445
-  Thomas F. Coleman, Yuying Li and Cheng Wang, “Stable Local Volatility Function Calibration Using Kernel Splines,” in print, 2010