Two Affine Scaling Methods for Solving Optimization Problems Regularized with an L1-norm

by

Zhirong Li

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2010

© Zhirong Li 2010

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.I hereby declare that I am the sole author of this thesis.

I authorize the University of Waterloo to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the University of Waterloo to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Abstract

In finance, the implied volatility surface is plotted against strike price and time to maturity. The shape of this volatility surface can be identified by fitting the model to what is actually observed in the market. The metric that is used to measure the discrepancy between the model and the market is usually defined by a mean squares of error of the model prices to the market prices. A regularization term can be added to this error metric to make the solution possess some desired properties. The discrepancy that we want to minimize is usually a highly nonlinear function of a set of model parameters with the regularization term. Typically monotonic decreasing algorithm is adopted to solve this minimization problem. Steepest descent or Newton type algorithms are two iterative methods but they are local, i.e., they use derivative information around the current iterate to find the next iterate. In order to ensure convergence, line search and trust region methods are two widely used globalization techniques.

Motivated by the simplicity of Barzilai-Borwein method and the convergence properties brought by globalization techniques, we propose a new Scaled Gradient (SG) method for minimizing a differentiable function plus an $\mathcal{L}_1$-norm. This non-monotone iterative method only requires gradient information and safeguarded Barzilai-Borwein steplength is used in each iteration. An adaptive line search with the Armijo-type condition check is performed in each iteration to ensure convergence. Coleman, Li and Wang proposed another trust region approach in solving the same problem. We give a theoretical proof of the convergence of their algorithm. The objective of this thesis is to numerically investigate the performance of the SG method and establish global and local convergence properties of Coleman, Li and Wang's trust region method proposed in [26]. Some future research directions are also given at the end of this thesis.

Acknowledgments

First and foremost, I am heartily thankful to my supervisor, Professor Yuying Li, whose encouragement, guidance and support from the initial to the final level enabled me to develop an understanding of this research subject. She continually and convincingly conveyed a spirit of adventure in regard to research and scholarship, and an excitement in regard to teaching. Without her guidance and persistent help this thesis would not have been possible.

I would also like to thank my thesis readers, Professor Peter Forsyth and Professor Justin Wan for taking time to review my thesis carefully and providing valuable comments, corrections and advice.

I am indebted to a friendly and cheerful group of fellow students. Wei Zhou has provided good arguments about mathematics theory and helped me get on the road to LATEX. Tiantian Bian has provided experienced suggestions in various aspects and some good laughs. Jun Chen has fascinated me with his ability to write good computer programs. I greatly enjoyed discussing finance problems with Yousef Sohrabi from time to time. In many ways I have learned so much from the members of Scientific Computing Lab at the University of Waterloo.

Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of this thesis.

Dedication

My deepest gratitude goes to my family for their unflagging love and support throughout my life. This thesis is simply impossible without them. I am indebted to my father, Chengyao Li, who inspired, encouraged and fully supported me for every trial that comes my way. In giving me not just financial, but morally and spiritually.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The number of function and derivative evaluations is considered as the main computational cost for solving nonlinear programming problems. For convex problems, interior point method has been proved to be an efficient approach. For nonconvex problems, trust region and affine scaling interior point method [24] was proposed to generate approximate affine scaling Newton steps for the complementarity conditions. These methods are in the family of monotonic decreasing algorithms. Barzilai and Borwein in [1] presented a remarkable result of the stepsize calculation and provided a non-monotone gradient algorithm. In this chapter, we first give a literature survey on the algorithms that have been used to solve least squares problems with $\mathcal{L}_1$ penalty. Then we review some globalization methods to solve nonlinear optimization problems with the $\mathcal{L}_1$-norm regularization. In the end, we summarize the objectives and contributions of this thesis.

1.1 Literature Review of Least Squares with $\mathcal{L}_1$ Penalty

In practice, it can be very important to solve a least squares optimization problem with the $\mathcal{L}_1$-norm regularization. A simple example of such problems is given below

$$\min_{w \in \mathbb{R}^n} \|Xw - y\|_2^2 + \rho \|w\|_1 \tag{1.1}$$

where X is an $m \times n$ design matrix, y is an $m \times 1$ column vector containing the values of the target variables, w is an $n \times 1$ coefficient column vector and ρ is a positive regularization parameter. $\mathcal{L}_1$ regularization has retained many benefits of $\mathcal{L}_2$ regularization. However, $\mathcal{L}_1$ regularization yields sparse and more stable solutions. In compressive sensing, $\mathcal{L}_1$-norm is widely used to formulate the signal recovery algorithm [18, 19, 20]. Problem (1.1) can be formulated either in an unconstrained manner or a constrained manner.

The equivalent form to the objective function (1.1) in a constrained formulation is given by

$$\begin{aligned} & \min\nolimits_{w\in\mathbb{R}^n} \left\|Xw - y\right\|_2^2 \\ st. \qquad & \left\|w\right\|_1 \leq t \end{aligned} \tag{1.2}$$

where $t > 0$ is in relation to the optimal solution to (1.1).

In [14], the constraint in (1.2) can be converted to a linear system of inequalities by adding either plus sign or minus sign in front of each component of w, yielding 2^n combinations. Then the problem is formulated as a standard Quadratic Programming (QP) problem with exponentially many linear inequality constraints. Tibshirani in [14] proposed to sequentially add constraints and seek solutions satisfying Karush-Kuhn-Tucker (KKT) conditions [13]. They claimed that this algorithm on average converges in $0.5n \sim 0.75n$ iterations. The problem of this approach is that it requires to solve a potentially large number (up to 2^n in theory) of QP problems, which is computationally very expensive.

Tibshirani in [14] used a second approach converting the constraints by introducing $2n$ non-negative variables. The design matrix now becomes $\bar{X} = [X, -X]$ and each component w_i is represented as

$$w_i = w_i^+ - w_i^-$$

where $w_i^+ = \begin{cases} |w_i| & w_i > 0 \\ 0 & w_i = 0 \\ 0 & w_i < 0 \end{cases}$ and $w_i^- = \begin{cases} 0 & w_i > 0 \\ 0 & w_i = 0 \\ |w_i| & w_i < 0 \end{cases}$.

Thus the constraint in (1.2) is converted to taking the form as $w_i^+, w_i^- \geq 0, i \in \{1, \cdots, n\}$ and $\sum_{i=1}^n \left[w_i^+ + w_i^-\right] \leq t$. The number of variables has been doubled but one only needs to solve a single QP problem and the number of constraints is $2n + 1$.

Basis Pursuit Denoising method [15] was proposed to solve (1.2) on non-negative variables w^+, w^- in the following form

$$\begin{aligned} &\min\nolimits_{v,\lambda} \|\lambda\|_2^2 + \rho \cdot e^T v \\ st. \qquad &\lambda = y - \bar{X} v \qquad v \geq 0 \end{aligned}$$

where $e^T = \left[\underbrace{1, \cdots, 1}_{\#2n}\right]$ is the vector of all one and $v = \begin{bmatrix} w^+ \\ w^- \end{bmatrix}$. By adopting an Interior Point Method (IPM) [29] with a log barrier function, the problem (1.3) can be iteratively solved

$$\begin{aligned} &\min\nolimits_{v,\lambda} \|\lambda\|_2^2 + \rho e^T \cdot v - \mu \sum\nolimits_{i=1}^{2n} \log v_i \\ st. \qquad &\lambda = y - \bar{X} v \end{aligned} \tag{1.3}$$

where $\mu > 0$. The main idea of this approach is to perform a Newton step based on the KKT conditions on the primal feasibility, the dual feasibility and the duality gap. When $\mu \rightarrow 0$, this algorithm solves the original problem (1.2). It is well known that IPM converges very fast but can be computationally expensive in each iteration.

Active set method was proposed in [16] to solve (1.2) by exploiting linearization of the norm constraint about the current solution estimate w and maintaining an active set acting on the non-zero variables. On each iteration, a quadratic program is solved to find a descent direction. This method converges fast but when the active set is large, partial factorization of a permuted design matrix needs to be performed in order to maintain efficiency. Moreover, this algorithm needs to handle the sign change in the active set during an iteration.

Problem (1.1) can also be solved using Iterated Ridge Regression [25]. This method starts with a solution to the least squares with $\mathcal{L}_2$ penalty and solves the regression formula using approximation of the objective function. Despite of its simplicity, this algorithm may stop at a sub-optimal solution and become very unstable when some components are very close to zero due to approximation.

Grafting [21] is another approach by properly defining the sign of w_i when it is zero. It has been argued that Grafting can be much more efficient than applying a quasi-Newton step. Gauss-Seidel [22] is a similar approach to Grafting except that it is used to optimize the free variables. This method has a very low iteration cost due to its extremely cheap line search. In the meantime, convergence rate of Gauss-Seidel is not comparable to other aforementioned methods.

We prefer to solve (1.1) in its original formulation since

- We want to solve a general nonlinear minimization problem with the $\mathcal{L}_1$-norm regularization. In addition, we do not restrict ourselves to least squares problems only.

- We add proper diagonal scaling in each iteration to handle nondifferentiability caused by the $\mathcal{L}_1$ -norm.

In finance, volatility surface calibration is a very important problem. The volatility surface can be recovered by finding the coefficients for a set of spline basis. We have some observed option prices as the training set and we intend to minimize the discrepancy between the derived model prices and the observed market prices. The objective function we want to minimize is usually least squares of the discrepancy between those two prices. More often than not, it is required to solve a PDE to get the model prices hence the objective function is nonlinear in the model parameters, namely, the spline basis coefficients. $\mathcal{L}_1$ regularization is added to ensure a volatility function is simple to achieve stability; this is achieved in [26] by making the number of nonzero coefficients as small as possible.

Trust region methods and line search methods are two possible globalization approaches for descent algorithm for nonlinear minimization. In Section 1.2, we will give a brief description of the trust region method and the Barzilai-Borwein method and highlight their strength and weakness to motivate the research in this thesis.

1.2 Review of Globalization Methods

In order to ensure the globalization properties, there are two means to achieve this goal. One approach is through line search and typically it provides a monotonic decreasing algorithm. When GLL type line search [8] is performed and Barzilai-Borwein stepsize is used, the algorithm becomes non-monotone. The other approach is trust region method. This method finds the approximate Newton solution within the trust region and it is a monotonic decreasing algorithm.

1.2.1 Trust Region Methods

The pure Newton method is arguably the most complex and the fastest iterative method under certain conditions [13]. However, the Newton step is calculated as the minimizer of the second-order Taylor expansion of the objective function around the current iterate. If the function around the current iterate is highly nonlinear, second-order approximation to the objective function may not be good and result in the next iterate too far away from the current iterate. Therefore it makes sense to consider a quadratic approximation in a small neighborhood, usually within a circle in which we believe the second-order approximation is acceptable. In view of the cost, when the Hessian is not positive definite, pure Newton direction may not be a descent direction. Minimizing the quadratic approximation in a suitably chosen small neighborhood provides a descent direction.

Line search and trust region methods can both use the second-order Hessian information. The difference for trust region method is that the search direction and the step size are determined simultaneously. For the trust region method, it is not necessary to find the exact solution for the sub-problem. Indeed, most practical trust region algorithms seek for an approximate solution within the trust region to give a sufficient decrease. It is relatively easy to compute the Cauchy point for the trust region sub-problem. However, if we always take Cauchy point, it in essence is similar to applying the steepest descent method. Dogleg method and two dimensional subspace minimization method [17] are the two variations improved on the Cauchy point based trust region methods. They both yield the approximate solution to be the Newton solution whenever the Newton solution is within the trust region.

Dogleg method is suitable for positive definite Hessian case. Dogleg method takes a dogleg-like trajectory which is composed of two line segments. If the trust region size is small, the Newton solution may lie outside of the trust region. Dogleg method will first move along the steepest descent direction to the Cauchy point then move towards the Newton solution. The dogleg path can be easily computed.

The two dimensional subspace method enlarges the search space for approximate solution in the span of gradient g and the Newton step $B^{-1}g$. Under slight modifications, this method can be applied to nonconvex problems.

In [26], Coleman, Li and Wang proposed a trust region method which is capable of solving a minimization problem for a twice differentiable function plus an $\mathcal{L}_1$-norm. 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. We will show how their algorithm works in Chapter 4.

1.2.2 Barzilai-Borwein Method and Extensions

1.2.2.1 Original Barzilai-Borwein Method

Barzilai and Borwein proposed in [1] a remarkable choice of the steplength for the gradient method for the unconstrained minimization of a differentiable function. For some convex quadratic functions, this gradient based algorithm achieves better convergence rate than the steepest descent line search method and requires less computational burden as compared to the standard steepest descent method. In their original paper [1], they proved that the generated sequence by Barzilai and Borwein method converges to the global minimizer R-superlinearly for a two dimensional convex quadratic optimization problem. In [2], Raydan established the convergence analysis for convex quadratic minimization in any dimension by spectral decomposition of the residual error norms. For the same problem in dimension ≥ 3, the convergence rate was proved to be R-linear by Y.H. Dai and L.Z.Liao in [3].

The most magnificent part of the Barzilai-Borwein method lies in that convergence is maintained while allowing for some occasionally nondecreasing steps. The calculation of the BB stepsize is also very simple in the form of the Rayleigh Quotient. For the traditional steepest descent method, it is a gradient method with an exact line search. The performance of the steepest descent method in terms of the convergence speed is well known to deteriorate as the condition number of the Hessian increases. For the more ill-conditioned convex quadratic problem, usually it is very difficult to maintain monotonicity at some iterates that in turn result in taking very small stepsizes to converge to the minimum. In contrast, loss of monotonicity sometimes can help take larger stepsizes to jump away from those flat areas and achieve better convergence results as the Barzilai-Borwein method seems to provide.

For a general non-quadratic problem, the Barzilai-Borwein method is not as comparable as the Conjugate Gradient(CG) based algorithm. However, if the objective function is made up of a quadratic plus a small non-quadratic term, improvements are expected to be made relative to CG based method [4].

1.2.2.2 Variations of the Barzilai-Borwein Method

Recently, there have been many variations of the Barzilai-Borwein method in the literature. Barzilai-Borwein method can be classified as a special case in the family of Gradient Methods with Retards(GMR) [5]. In [6], Yuhong Dai et al. investigated the connection between the convergence rate and two versions of the BB stepsize and proposed an Adaptive Steepest Descent (ASD) method, which chooses the appropriate BB stepsize using the metric related to the ratio of the two versions of the BB stepsize. By introducing proper truncation on the BB stepsize, the algorithm is surprisingly monotone.

1.2.2.3 Extension to the Linearly Constrained Optimization Problem

Many researchers use non-monotone Armijo-type line search with acceptability test to ensure global convergence. The way they handle the constraints is to project the trial point back to the feasible region whenever it is out of the boundary. Yu-Hong Dai and Roger Fletcher [7] examined the performance of Adapted Projected Barzilai-Borwein method (PABB) for large scale box-constrained quadratic programming problem. They argued that, in a convex quadratic case, if GLL [8] non-monotone line search is performed on every iteration, the performance is noticeably degraded unless large number of candidate function values that are used to generate the reference function is chosen. They proposed to periodically update the reference values on every 10 iterations to achieve better performance. However, for indefinite Hessian case, the original PBB method outperforms PABB method.

Spectral Projected Gradient (SPG) method was proposed in [9] to find the minimizer of a first-order differentiable general function on a closed convex set. SPG method is a combination of Barzilai-Borwein and line search methods to give convergence. However, they used a simplified adaptive line search method thus only one projection is required for each line search instead of doing projection on every trial point. They used the original Barzilai-Borwein stepsize to find the trial point along the negative gradient direction. If this trial point is out of the feasible region, an orthogonal projection is performed to map the trial point onto the boundary of the feasible region. The new search direction then is determined by connecting the current iterate to the new projected point. Line search is performed on this line segment to find the first trial point at which non-monotone Armijo-type rule is satisfied. They proved that either the algorithm stops at the stationary point or every limit point of the generated sequence is stationary.

In all, both trust region method and Barzilai-Borwein type method can be applied to solve a minimization problem for a twice differentiable function plus an $\mathcal{L}_1$-norm. We are motivated to use affine scaling in both the proposed scaled gradient method and trust region method. Barzilai-Borwein type methods inspire us to investigate the possibility to adopt a non-monotone algorithm along certain properly defined gradient based descent direction. We hope to devise such a new gradient based algorithm so that it possesses the properties of a small computational cost and fast convergence, at least when the smooth function in the objective is not too nonlinear.

1.3 Objectives and Contributions

Nonlinear optimization problem with the $\mathcal{L}_1$-norm regularization has drawn intense research attention due to its wide applications in many areas. A lot of research effort has been dedicated to finding an efficient algorithm to solve this problem. The objectives and contributions of this thesis are two-fold. Motivated by the role of affine scaling in dealing with the $\mathcal{L}_1$-norm and the simplicity of the Barzilai-Borwein type non-monotone algorithms, we want to investigate an efficient non-monotone method using gradient and appropriate affine scaling for minimizing a nonlinear function with the $\mathcal{L}_1$-norm regularization. We also want to analyze and establish convergence properties of Coleman, Li and Wang's trust region method proposed in [26].

The main contributions of the thesis include:

- Propose an affine scaling gradient based algorithm with non-monotone line search for minimizing nonlinear function regularized by an $\mathcal{L}_1$-norm. The novelty of this algorithm comes from taking a scaled descent direction with safeguarded Barzilai-Borwein stepsizes. In order to ensure convergence, an adaptive line search is performed on each iteration. This algorithm does not require exact line search and the overall computational cost is cheap.

- Investigate the computational performance of our proposed scaled steepest descent method and impact of the parameter choice. We also compared our method against the spectral projected gradient method to illustrate the effectiveness of our algorithm.
- Establish the 1-st order convergence properties of the affine scaling trust region method proposed in [26]. The 2-nd order convergence properties are presented but the proof is omitted due to its strong similarity to the proof in [11].

We will discuss the optimality conditions for the problem we consider in Chapter 2. In Chapter 3, we will propose a scaled steepest descent method using safeguarded BB stepsize. In Chapter 4, we will prove the convergence properties of Coleman, Li and Wang's trust region method proposed in [26], which is another approach to solve the same nonlinear minimization optimization problem with the $\mathcal{L}_1$ regularization. Chapter 5 will conclude the thesis and some future research directions will be given.

Chapter 2

Optimality Conditions

The Karush-Kuhn-Tucker Necessary Conditions for smooth and constrained optimization problems [13] can be specified as the following

Proposition 2.1. *Let x^* be a local minimum of the problem*

$$\begin{aligned} &\min f\left(x\right) \\ st. \quad &\mathcal{E}_1\left(x\right)=0,\cdots,\mathcal{E}_m\left(x\right)=0 \\ &\mathcal{I}_1\left(x\right)\leq 0,\cdots,\mathcal{I}_r\left(x\right)\leq 0 \end{aligned}$$

where $f, \mathcal{E}_i, \mathcal{I}_j$ are continuously differentiable function from $\mathbb{R}^n$ to $\mathbb{R}$, and assume that x^* is regular. Then there exist unique Lagrange multiplier vectors $\lambda^* = (\lambda_1^*, \cdots, \lambda_m^*)$, $\mu^* = (\mu_1^*, \cdots, \mu_r^*)$, such that

$$\nabla_x L\left(x^*, \lambda^*, \mu^*\right) = 0$$

$$\mu_j^* \geq 0, j \in \{1, \cdots, r\}$$

$$\mu_j^* = 0, \forall j \notin A\left(x^*\right)$$

where $A(x^*)$ is the set of active constraints at x^* and $L(x,\lambda,\mu)$ is the Lagrangian function defined by $L(x,\lambda,\mu) = f(x) + \sum_{i=1}^{m} \lambda_i \mathcal{E}_i(x) + \sum_{j=1}^{r} \mu_j \mathcal{I}_j(x)$. If in addition f, $\mathcal{E}$ and $\mathcal{I}$ are twice continuously differentiable, there holds

$$y^T \nabla_{xx}^2 L(x^*, \lambda^*, \mu^*)\, y \geq 0$$

for all $y \in \mathbb{R}^n$ such that

$$\nabla \mathcal{E}_i (x^*)^T y = 0, \forall i = 1, \cdots, m$$

and

$$\nabla \mathcal{I}_j (x^*)^T y = 0, \forall j \in A(x^*)$$

where $\nabla_x L(x,\lambda,\mu)$ and $\nabla_{xx} L(x,\lambda,\mu)$ are the first-order partial derivative and second-order partial derivative with respect to x , respectively.

A feasible vector x is said to be regular if the quality constraint gradients $\nabla \mathcal{E}_i(x), i = 1, \cdots, m$, and the active inequality constraint gradients $\nabla \mathcal{I}_j(x), j \in A(x)$, are linearly independent. For any feasible point x, the set of active inequality constraints is denoted by $A(x) = \{j \,|\mathcal{I}_j(x) = 0\}$.

Without loss of generality, we assume in this thesis, that we want to minimize a nonlinear function plus an $\mathcal{L}_1$-norm regularization. Specially, we want to solve

$$\min_{x \in \mathbb{R}^n} f(x) + \|x\|_1 \tag{2.1}$$

where $f \in \mathbb{C}^2$ is a twice continuously differentiable function and the $\mathcal{L}_1$-norm of $x \in \mathbb{R}^n$ is defined by $\|x\|_1 := \sum_{i=1}^{n} |x_i|$. Since (2.1) has an equivalent nonlinear programming formulation, it can be shown, see, e.g. [28], that the first order necessary KKT condition for (2.1) can be expressed as the following : for any $1 \leq i \leq n$

$$\begin{cases} (x_i) \cdot \left((\nabla f(x))_i + sign(x_i) \right) = 0 \\ \left| (\nabla f(x))_i \right| \leq 1 \end{cases} \tag{2.2}$$

where

$$sign(x) := \begin{cases} -1 & x < 0 \\ 0 & x = 0 \cdot \\ 1 & x > 0 \end{cases}$$

For the gradient component $x_i^* = 0$, the objective function (2.1) is not differentiable but condition$\left| (\nabla f(x^*))_i \right| \leq 1$ should be satisfied. Otherwise, for example, if $\left| (\nabla f(x^*))_i \right| > 1$, we can always find a descent direction along the i-th axis that will lead to the contradiction of the optimality assumption at x^*.

Define a vector $v(x) \in \mathbb{R}^n$ as

$$(v(x))_i := \begin{cases} 1 & \left| (\nabla f(x))_i \right| > 1 \\ |x_i| & otherwise \end{cases} \tag{2.3}$$

and the corresponding scaling matrix as $D(x) := diag(v(x)) \in \mathbb{R}_+^{n \times n}$. Note that $v(x)$ is a vector of the distance to each axis from the current iterate other than the component whose scaling factor is defined to be one. We choose affine scaling because affine scaling can help avoid the non-differentiable points. Otherwise, when $\left| (\nabla f(x))_i \right| \leq 1$, if we choose $(v(x))_i = |x_i|^\alpha$, then the distance from the current iterate to the corresponding break point is degenerated to be either 0 for $\alpha < 1$ or $+\infty$ for $\alpha > 1$, respectively.

The component-wise KKT condition (2.2) can be rewritten in a vector/matrix form as

$$D(x) \cdot (\nabla f(x) + sign(x)) = \mathbf{0}. \tag{2.4}$$

If we assume that at the current iterate $x_k \in \mathbb{R}^n$, x_k satisfies $(x_k)_i \neq 0, \forall i \in \{1, \cdots, n\}$, then a Newton step d_k for (2.4) can be defined as the solution to

$$\left(J_k^v \cdot diag\left(g_k\right) + diag\left(v_k\right) \cdot \nabla^2 f\left(x_k\right)\right) d_k = -diag\left(v_k\right) g_k \tag{2.5}$$

where $g_k := \nabla f(x_k) + sign(x_k)$ and $J_k^v(x) \in \mathbb{R}^{n \times n}$ is a diagonal matrix, which corresponds to the Jacobian matrix of $|v(x_k)|$ with each diagonal element $\left(J_k^v(x)\right)_{i,i}$ equal to either 0 when $(v(x_k))_i = 1$ and ±1 otherwise.

Affine scaling as defined in (2.3) is specially designed to handle the $\mathcal{L}_1$-norm. Problem (2.1) is essentially a minimization problem for a differentiable function $f(x)$ with nondifferentiable $\mathcal{L}_1$-norm $\|x\|_1$. When the Lagrangian multiplier associated with nondifferentiable hyperplane $\{x_i = 0\}$ suggests that this hyperplane should not be binding at the solution, i.e., $|(\nabla f(x))_i| > 1$, the scaling $D_{i,i} = 1$ facilitates the iterates to quickly move away from this hyperplane. On the contrary, when the Lagrangian multiplier associated with nondifferentiable hyperplane $\{x_i = 0\}$ suggests that $x_i = 0$ is likely to be binding, the scaling $D_{i,i} = |(x_k)_i|$ can prevent the iterates from crossing $\{x_i = 0\}$ prematurely and help iterates to converge to a solution.

The Newton step defined by (2.5) suggests that, if $\nabla^2 f(x_k) \approx 0$, $-D_k g_k$ can be regarded as an approximate Newton step, which motivates us to use $-diag(v_k) g_k$ as our descent direction.

From the necessary optimality condition for (2.1), we would like the iterates to satisfy $\lim_{k\to\infty} \|D_k g_k\| = 0$. We will see this in Chapter 3 by simulation, that the SG method terminates with $\lim_{k\to\infty} \|D_k g_k\| = 0$ being approximately satisfied. Moreover, both the first-order and the second-order necessary optimality conditions are satisfied for the trust region method as described in Chapter 4.

Chapter 3

A Scaled Gradient Method

3.1 Motivations of Our Research

A main attractive feature of a gradient based method is computational simplicity. A Newton type method requires a solution to an $n \times n$ linear systems of equations. For large n, this computation can be expensive. Inspired by the simplicity of the Barzilai-Borwein method and the SPG method, we want to solve the nonlinear optimization problem with $\mathcal{L}_1$ penalty using a similar approach.

For simplicity, in this investigation, we first focus on a convex quadratic plus an $\mathcal{L}_1$-norm as given below

$$\min_{x\in\mathbb{R}^n} \frac{1}{2} x^T H x - b^T x + \|x\|_1 \tag{3.1}$$

where H is an $n \times n$ positive definite matrix and b is an $n \times 1$ column vector. Problem (3.1) is a convex problem since a superposition of two convex functions is convex. It is known that, for a convex function, the local minimum and the global minimum coincide. We use $h(x)$ to denote the objective function, i.e., $h(x) = \frac{1}{2} x^T H x - b^T x + \|x\|_1$.

We want to use a gradient-type approach since only first-order information is supplied and it is easy to compute at each iteration. Similar to [10], we do not force our algorithm to be monotone because monotone gradient algorithm can converge very slowly. We want to use BB-type stepsize for the reason that computing BB-stepsize only requires Δx and Δg and it has been proved to work well for unconstrained quadratic minimization. We prefer to keep using adaptive non-monotone line search in order to ensure a non-monotone Armijo rule being satisfied.

Cauchy steplength $\alpha_k = \frac{g_k^T g_k}{g_k^T H g_k}$ where $g = \nabla h = Hx - b + sign\left(x\right)$ in the steepest descent direction does not work for (3.1) since the problem (3.1) is essentially a constrained optimization problem. Motivated by the KKT conditions (2.4), we intend to use a Scaled Steepest Descent (SSD) direction $-diag\left(v_k\right) g_k$ as our search direction to handle the second term in the objective function.

3.2 Our Proposed Method

We want to answer the following questions. Firstly, what search direction should we take? Secondly, what stepsize should be applied? Next, what type of line search should be performed? Finally, what terminating conditions should be specified? In this section, we will elaborate on how we design our affine scaling gradient based method.

3.2.1 Challenges and Solutions

3.2.1.1 Search Direction

A simple method that one might naturally want to use is the steepest descent $-g_k$ with the exact line search. We explain how the line search can be done and why this method will not work. We can still move along the steepest descent direction $-g_k = -\left(Hx_k - b + sign\left(x_k\right)\right)$ using the exact line search. The stepsize will be taking different formulation from that for unconstrained minimization because we have the $\mathcal{L}_1$-norm in the objective function (3.1). We define the change in the objective function $h\left(x\right)$ along d_k as

$$p\left(\alpha\right) \triangleq f\left(x_{k+1}\right) + \left\|x_{k+1}\right\|_1 = \alpha\left(Hx_k - b\right)^T d_k + \frac{1}{2}\alpha^2 d_k^T H d_k + \left\|x_k + \alpha d_k\right\|_1 + f\left(x_k\right)$$

where the iterate is updated by $x_{k+1} = x_k + \alpha d_k$.

The minimizer α^* of $p(\alpha)$ has the following characteristics. The optimizer α^* for $p(\alpha)$ in the negative gradient direction $d_k = -g_k$ is either $\alpha^* = \beta_k^{i^*+1}$ or $\alpha^* = \beta_k^{i^*} + \frac{\langle g_k^{i^*}, g_k^0 \rangle}{d_k^T H d_k}$ where $i^* \in \{0, \cdots, l_k - 1\}$ is the last break point that is crossed, i.e., $\beta_k^{i^*} := \max\{\beta_k^i : \beta_k^i < \alpha^*\}$ and g_k^i is the gradient immediately after crossing the i-th break point. Note that $\beta_k^i := \frac{\left|(x_k)_{j_i}\right|}{\left|(d_k)_{j_i}\right|}$ where j_i is the component index for the i-th break point along d_k.

The minimizer can only be either the local minimum of one quadratic piece or one of the break points along the SD direction. In Chapter 4, we will provide such a proof (Lemma 4.1).

We can see that if we perform line search along the SD direction, it is possible that the SD direction will become an ascent direction after crossing some break point. Therefore backtracking needs to be performed to avoid staying at those non-differentiable hyperplanes $\{x_j = 0\}, j \in \{1, \cdots, n\}$. Affine scaling can be served to avoid non-differentiable points. This is because, for instance, if component $(x_k)_i$ of the current iterate x_k is approaching zero and the necessary optimality condition for this component is satisfied, i.e., $|\nabla f((x_k)_i)| \leq 1$, then moving along the scaled steepest direction $-diag(v_k)\, g_k$ with a distance of β_k to $\{x_i = 0\}$ can ensure that

$$\liminf_{k\to\infty} \beta_k = \liminf_{k\to\infty} \frac{|(x_k)_i|}{(D_k)_{i,i} \cdot |(g_k)_i|} = \liminf_{k\to\infty} \frac{1}{|(g_k)_i|} \geq \liminf_{k\to\infty} \frac{1}{|g_k|_\infty} > 0$$

which can be interpreted as preventing the current iterate from directly approaching $\{x_i = 0\}$ by taking the scaled steepest descent direction $-diag(v_k)\, g_k$, which is nearly tangential to $\{x_i = 0\}$. In contrast, if $|\nabla f((x_k)_i)| > 1$, then the break point is nearly zero hence it allows iterates to cross the corresponding non-differentiable hyperplane.

For a large nonlinear problem, computing a Newton type step can be computationally extensive. As one way out, we want to use the SSD direction $-D_k g_k$ on the k-th iteration. This new direction is still a descent direction since the inner product of $\langle -D_k g_k, g_k \rangle =$

$-g_k^T D_k g_k \leq 0$ where D_k is a positive semidefinite diagonal matrix. We want to get reader's attention that it is possible for many components of x^* binding at zero, i.e., the optimal faces could reside on one or more of the non-differentiable hyperplanes $\{x_j = 0\}, j \in \{1, \cdots, n\}$. Suppose that the current iterate's i-th component x_i is still far from the optimal face $\{x_i = 0\}$ and $|(\nabla f(x))_i| \leq 1$, taking $|x_i|$ in as the the scaling factor for the i-th component may result in tremendous deviation from moving towards $\{x_i = 0\}$. As a result, we define the scaling matrix in our context as

$$D(x) := diag\left(v\left(x\right)\right)$$

where

$$(v(x))_i := \begin{cases} 1 & |(\nabla f(x))_i| > 1 \\ \min\{|x_i|, 1\} & |(\nabla f(x))_i| \leq 1 \end{cases} \tag{3.2}$$

By doing so, we avoid adding scaling effect when the iterate is not close to the optimal solution yet. Scaling plays an important role in handling $\mathcal{L}_1$-norm when non-differentiability issue comes into play. Note that the KKT condition can also be equivalently stated as $D(x) \cdot g(x) = 0$ where $D(x)$ is defined by (3.2).

3.2.1.2 Stepsize

Similar to the derivation in [1], to obtain a BB type stepsize for (3.1), we consider the optimization problem as defined by

$$\min_{\alpha \in \mathbb{R}^n} \left\| \tfrac{1}{\alpha} D_k \Delta x_k - D_k \Delta g_k \right\|^2 \tag{3.3}$$

where $\Delta x_k = x_k - x_{k-1}$ and $\Delta g_k = g_k - g_{k-1}$.

The minimum of (3.3) is trivially derived as

$$\alpha_k^{BB1} := \alpha^* = \frac{\langle D_k \Delta x_k, D_k \Delta x_k \rangle}{\langle D_k \Delta x_k, D_k \Delta g_k \rangle} \tag{3.4}$$

Similarly, the minimum of $\min_{\alpha \in \mathbb{R}^n} \|D_k \Delta x_k - \alpha D_k \Delta g_k\|^2$ is given by

$$\alpha_k^{BB2} := \alpha^* = \frac{\langle D_k \Delta g_k, D_k \Delta x_k \rangle}{\langle D_k \Delta g_k, D_k \Delta g_k \rangle} \tag{3.5}$$

If we substitute $\Delta x_k = x_k - x_{k-1}$ and $\Delta g_k = H \Delta x_k + sign\left(x_k\right) - sign\left(x_{k-1}\right)$ into (3.4) and (3.5), we observe that neither α_k^{BB1} nor α_k^{BB2} is Rayleigh quotient due to break point crossings $(sign\left(x_k\right) - sign\left(x_{k-1}\right) \neq 0)$.

We look into the expansion of the denominator of α_k^{BB1}

$$\langle D_k \Delta x_k, D_k \Delta g_k \rangle \quad = \quad \Delta x_k^T D_k^2 \Delta g_k = \Delta x_k^T D_k^2 H \Delta x_k + \Delta x_k^T D_k^2 \left(sign\left(x_k\right) - sign\left(x_{k-1}\right)\right).$$

We observe that the second term can pull the total sum down to negative values.

Based on the following arguments, we want our stepsize $\alpha_k^{\overline{BB}}$ to be in $[\alpha_{\min}, \alpha_{\max}]$ where $0 < \alpha_{\min} < 1 < \alpha_{\max}$.

- Stepsize needs to be bounded away from zero otherwise the algorithm may stop at non-optimal point where terminating condition can be easily satisfied since the last two iterates could be infinitely close to each other. In contrast, if $\liminf_{k \to \infty} \alpha_k^{\overline{BB}} > 0$ and the generated sequence converges, from the iterate update formula $x_{k+1} = x_k - \alpha_k^{\overline{BB}} D_k g_k$, this will lead to $\lim_{k \to \infty} \|D_k g_k\| = 0$, which means the first order necessary optimality condition is satisfied.

- Negative stepsize is not meaningful when we move in a descent direction. In [7, 9], when stepsize is negative, they choose to use a bold step $\alpha_{\max}$ instead of $\alpha_{\min}$. Whenever we meet negative stepsize in (3.4) and (3.5), the algorithm is in the situation that a break point is crossed. If we choose $\alpha_{\min}$, we may prohibit from crossing the break point since this stepsize is small and the next iterate may remain in the same orthant. Another alternative is to use Rayleigh Quotient whenever we encounter the negative stepsize.

- Large stepsize inherent in the original BB stepsize formulae may lead iterates to jump away from the minimum thus oscillate around the optimal solution. Hence we impose an upper bound $\alpha_{\max}$ on the stepsize $\alpha_k^{\overline{BB}}$.

- Unit stepsize 1 is effective when used with the Newton step. We can judge from the following claim that in some occasions, we should take stepsize 1 to make the iterate converge to the optimal face faster. Hence we let $1 < \alpha_{\max}$.

Unit stepsize 1 helps the corresponding components of iterates binding at zero when necessary optimality condition holds and the corresponding component is near the optimal face. This can be seen from discussion below.

Suppose for sufficiently large k, $\left|(\nabla f(x_k))_i\right| < 1$ holds and $\left|(x_k)_i\right| < 1$. If the iterate updating formula $x_{k+1} = x_k - D_k g_k$ holds for $k \geq \bar{k}$ where $\bar{k}$ is some fixed positive integer, we can conclude that

$$\begin{aligned}
(x_{k+1})_i &= (x_k)_i - \min\left\{\left|(x_k)_i\right|, 1\right\} \cdot (g_k)_i\,. \\
&= sign\left((x_k)_i\right) \cdot \left|(x_k)_i\right| - (g_k)_i \cdot \left|(x_k)_i\right| \\
&= -\left(\nabla f(x_k)\right)_i \cdot \left|(x_k)_i\right|
\end{aligned}$$

Therefore

$$\left|(x_{k+1})_i\right| = \left|(x_k)_i\right| \cdot \left|(\nabla f(x_k))_i\right| < \left|(x_k)_i\right|$$

hence

$$\left|(x_{k+1})_i\right| = \left|(x_k)_i\right| \cdot \left|(\nabla f(x_k))_i\right| = \left(\Pi_{j=\bar{k}}^{k} \left|(\nabla f(x_j))_i\right|\right) \cdot \left|(x_{\bar{k}})_i\right|$$

thus

$$\liminf_{k\to\infty} \left|(x_{k+1})_i\right| = 0.$$

3.2.1.3 Ensuring Convergence

For a non-convex minimization, taking BB stepsize always can lead to divergence. Non-monotone line search technique has been proposed to ensure convergence [8]. In order to do line search, on each iteration, Non-monotone Armijo-rule says we need to find the smallest natural number m such that

$$h\left(x_k - \theta^m \alpha_k^{\overline{BB}} D_k g_k\right) \le h_r + \theta^m \alpha_k^{\overline{BB}} \langle g_k, -D_k g_k \rangle$$

where $\theta \in (0,1)$ and h_r is the reference function value. In [9], they choose

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

where M is a positive integer. In addition, shrinking factor θ does not need to be a constant in $(0,1)$ in each iteration. They choose $\theta_{l+1} \in [\tau_1 \theta_l, \tau_2 \theta_l]$ where l is the index to indicate the number of shrinkings and $0 < \tau_1 < \tau_2 < 1$. The typical best value for M is chosen to be 10 [4]. Notice that if $M = 1$, the line search criterion reduces to the monotone Armijo-rule. In [9], it has been argued that the Spectral Projected Gradient method does not work well when $M = 1$.

3.2.2 Our Proposed Scaled Gradient Method

3.2.2.1 Algorithm Description

To answer the questions raised at the beginning of this section, we propose to use the scaled steepest descent direction as the search direction. Safeguarded Barzilai-Borwein stepsize is used to search for the first trial point. A non-monotone Armijo-type rule is applied to find the next iterate. Our proposed algorithm is described in Algorithm 3.1.

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 $h_{\max} \triangleq \max\left\{h\left(x_{k-j}\right) | 0 \leq j \leq \min\left\{k, M-1\right\}\right\}$ and
    $$\alpha_k^{\overline{BB}} = \frac{\left\langle D_k\left(x_k - x_{k-1}\right), D_k\left(x_k - x_{k-1}\right)\right\rangle}{\left\langle D_k\left(x_k - x_{k-1}\right), D_k\left(\left[H\left(x_k - x_{k-1}\right) + sign\left(x_k\right) - sign\left(x_{k-1}\right)\right]\right)\right\rangle}$$
Step 3: If $\alpha_k^{\overline{BB}} < 0$ then set $\alpha_k^{\overline{BB}} = \frac{\langle D_k \Delta x_k, D_k \Delta x_k\rangle}{\langle D_k \Delta x_k, D_k H \Delta x_k\rangle}$;
        set $\alpha_k^{\overline{BB}} = \min\left\{\alpha_{\max}, \max\left\{\alpha_{\min}, \alpha_k^{\overline{BB}}\right\}\right\}$
Step 4: Compute $\delta \leftarrow \left\langle g_k, -\alpha_k^{\overline{BB}} D_k g_k\right\rangle$ and set $\theta \leftarrow 1$
Step 5: While $h\left(x_k - \theta\alpha_k^{\overline{BB}} D_k g_k\right) > h_{\max} + \gamma\theta\delta$
        {set $\theta_{new} \in \left[\tau_1\theta, \tau_2\theta\right], \theta \leftarrow \theta_{new}$}
Step 6: $x_{k+1} = x_k - \theta\alpha_k^{\overline{BB}} D_k g_k$ and goto step 1
```

3.2.2.2 Discussions

Parameter M is used to set the reference values from

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

As M increases, it is more likely that the trial point gets accepted. If $M = 1$, then Algorithm 3.1 becomes a monotone algorithm. Parameter γ controls required decrease relative to the reference value. Note that the inner product $\delta = \left\langle g_k, -\alpha_k^{\overline{BB}} D_k g_k\right\rangle$ is always negative. Hence, as γ increases, the inequality

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

is more difficult to be satisfied. If, at some iterate x_k, it is very unlikely that the

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

is violated, we will keep shrinking the stepsize. We do not impose lower and upper bound on the stepsize $\theta\alpha_k^{\overline{BB}}$ in the while loop in step 5 because of these observations.

Our SG method combines the non-monotone line search scheme and Barzilai-Borwein stepsize with our proposed scaled steepest descent direction as the search direction to handle the $\mathcal{L}_1$-norm in the objective function. The performance of our SG method will be comprehensively examined in Section 3.3.

3.3 Computational Investigation

To illustrate the performance of our proposed SG method for minimizing a convex quadratic with the $\mathcal{L}_1$-norm regularization, we use Moré and Toraldo's method [27] to generate random test problems. We can specify dimension, condition number and sparsity of Hessian matrix H for the test problems. Throughout this investigation, we set the dimension of the test problem to be 10, condition number of H to be 3 and the sparsity of H to be 1 (a dense matrix).

3.3.1 Verification on the Appropriateness of Our Choice

Initially, we chose $\gamma = 0.5, \alpha_{\min} = 0.01, \alpha_{\max} = \infty, \tau_1 = 0.1, \tau_2 = 0.9$ and $M = 10$. The terminating condition is set to be whether or not the difference between the two consecutive objective values is less than the tolerance value $\sqrt{eps} = 10^{-8}$ or the maximum number of iterations 1000 is reached. We say an execution is successful when the norm of the scaled gradient is sufficiently small ($< 10^{-4}$ typically). If the last steplength $\theta\alpha_k^{\overline{BB}}$ is $< 10^{-4}$, it can easily make the change in $x_{k+1} = x_k - \theta\alpha_k^{\overline{BB}} D_k g_k$ be very small so that the difference in objective values can be less than $\sqrt{eps}$. Therefore the algorithm does not converge to the optimal point after termination.

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

Table 3.1: Average Number of Iterations / Success Rate for SD and SSD direction

3.3.1.1 Comparing Search Direction

We first illustrate, without using scaling, the steepest descent $-\nabla h(x_k)$ with the non-monotone stepsize rule does not work. We run 50 test cases on steepest descent direction as well to solve (3.6) for different ρ's

$$\min_{x \in \mathbb{R}^n} \rho \cdot \left(\frac{1}{2} x^T H x + b^T x \right) + \|x\|_1 \tag{3.6}$$

where ρ is the weighting factor for the quadratic term. We tried $\rho = 0.1, 1, 10, 100$ and the simulation results in terms of the average number of iterations and the success rate of the algorithm are listed in Table 3.1.

When the SD direction is used, although the algorithm terminates on average in hundreds of steps, the stepsize indeed converges to zero for most cases and the optimality condition $\|D_k g_k\| \to 0$ does not hold. Thus the algorithm does not stop at the optimal solution. We can also see that when the quadratic term becomes more dominant as ρ increases, it is possible for the algorithm that takes safeguarded BB stepsize along the steepest descent direction to work in some cases. This is entirely reasonable since problems become increasing similar to a convex quadratic minimization problem, as ρ increases.

3.3.1.2 Safeguarded BB Stepsize

We observed by simulation that for a large $\rho \geq 1000$, current choice of $\alpha_{\min} = 0.01$ often fails to solve a problem (3.6). We used the same quadratic function (but without $\mathcal{L}_1$ term) to test the Barzilai-Borwein method, the same thing happened. This indicates that the lower bound of the stepsize should be problem dependent. For large ρ, since the value of $\|D_k g_k\|$

might be very large, we should set $\alpha_{\min}$ to be some reasonable small positive value in regard to $\|D_0 g_0\|$ to make the algorithm converge with $\|D_k g_k\| \to 0$ being satisfied.

3.3.1.3 Adaptive Line Search

We can set θ to be any value in $(0, 1)$ and it turns out that there is no significant difference in terms of convergence speed when θ varies. As an alternative, we can do quadratic interpolations to find the

$$\theta_{new} = -\frac{1}{2}\theta^2 \delta / \left(h\left(x_+\right) - h\left(x_k\right) - \theta\delta\right)$$

where $h\left(x_+\right)$ is the objective value at the trial point x_+ as suggested in [9]. We observed that the algorithm does not always converge fast in our context. For $\rho = 10$, this choice of the shrinking factor θ works well and the algorithm terminates with an average number of 156 iterations. For $\rho = 100$, there are 4 out of 50 test cases, which terminate at the 1000-th iteration without meeting the required precision for the objective values. The rest 46 test problems terminate at the optimal solution.

3.3.2 Evaluate Impact of Parameters on Performance

3.3.2.1 The Impact of γ

The parameter γ is used to control the satisfiability of $h\left(x_k - \theta\alpha_k^{\overline{BB}} D_k g_k\right) > h_{\max} + \gamma\theta\delta$ in step 5. It takes value in $[0, 1]$. We run simulations for different γ in order to choose the best value of γ in terms of convergence speed.

Table 3.2 reports the number of iterations the SG algorithm takes to converge for 50 random test problems for $\gamma = 0.2, 0.4, 0.5, 0.6, 0.8$ and $\rho = 10$. Table 3.3 records the same information except for $\rho = 10$. We observe that, for $\gamma < 0.5$ when $\rho = 100$, there is not much difference in the number of iterations the algorithm takes to converge. We set the γ value in $(0.4, 0.8)$. We run the same test samples for $\rho = 10$ case. We observe that for

Test #	$\gamma = 0.2$	$\gamma = 0.4$	$\gamma = 0.5$	$\gamma = 0.6$	$\gamma = 0.8$	Test #	$\gamma = 0.2$	$\gamma = 0.4$	$\gamma = 0.5$	$\gamma = 0.6$	$\gamma = 0.8$
1	592	352	323	526	576	26	321	329	412	363	413
2	328	377	329	283	336	27	771	698	669	391	824
3	195	340	235	235	394	28	499	499	664	664	377
4	127	189	189	154	181	29	300	288	229	272	250
5	132	131	174	152	107	30	131	134	134	238	170
6	176	177	200	200	121	31	298	310	255	283	333
7	390	256	251	161	363	32	299	270	319	296	462
8	178	185	178	178	218	33	364	335	386	383	522
9	177	183	183	137	172	34	248	184	184	223	238
10	173	145	214	182	206	35	181	188	196	160	211
11	349	325	258	216	313	36	220	297	321	263	263
12	271	244	203	203	191	37	396	381	314	303	339
13	575	581	487	507	621	38	262	275	428	255	336
14	261	291	291	328	389	39	402	532	447	342	498
15	200	190	152	152	172	40	207	191	215	233	285
16	202	379	228	198	265	41	278	221	221	151	194
17	139	139	194	129	126	42	594	683	669	450	693
18	307	292	285	340	319	43	273	276	369	319	355
19	288	340	360	360	342	44	490	446	396	452	457
20	196	251	258	229	452	45	232	270	281	252	370
21	222	174	149	149	163	46	334	316	420	306	286
22	246	266	279	238	299	47	333	406	284	319	277
23	348	198	449	228	268	48	139	136	136	128	146
24	389	345	394	390	423	49	198	214	214	214	185
25	308	257	252	257	228	50	443	211	325	234	270

	$\gamma = 0.2$	$\gamma = 0.4$	$\gamma = 0.5$	$\gamma = 0.6$	$\gamma = 0.8$
Average #	300	294	300	273	320

Table 3.2: Number of Iterations vs. γ , $\rho = 100$

$\gamma \geq 0.8$, usually it takes more iterations to converge. Throughout Section 3.3.2, our SG method terminates with the optimality condition $\lim_{k\to\infty} \|D_k g_k\| = 0$ being approximately satisfied, i.e., $\|D_k g_k\| \leq 10^{-4}$ at termination.

As a result of simulations shown in Table 3.2 and 3.3, we choose $\gamma = 0.5$ since there is no much difference when $\gamma < 0.5$ and there might be fewer iterations on average required for large ρ.

Test #	$\gamma = 0.2$	$\gamma = 0.4$	$\gamma = 0.5$	$\gamma = 0.6$	$\gamma = 0.8$	Test #	$\gamma = 0.2$	$\gamma = 0.4$	$\gamma = 0.5$	$\gamma = 0.6$	$\gamma = 0.8$
1	283	129	122	122	94	26	97	88	85	80	147
2	76	62	62	68	581	27	204	204	191	191	214
3	302	136	117	163	238	28	51	51	41	41	41
4	87	87	87	87	129	29	265	323	323	323	313
5	165	88	77	121	126	30	74	126	126	126	91
6	196	231	280	280	120	31	259	259	259	259	271
7	60	53	53	53	61	32	143	123	123	113	57
8	46	157	157	157	107	33	193	193	179	68	287
9	511	508	814	467	341	34	94	94	94	169	169
10	389	388	330	464	233	35	41	42	42	42	96
11	75	94	74	87	92	36	85	106	106	106	122
12	112	239	239	261	337	37	125	97	93	267	86
13	56	56	76	76	60	38	97	101	67	111	92
14	84	84	84	137	988	39	108	108	108	93	93
15	257	168	168	261	226	40	135	135	371	371	169
16	42	411	334	106	225	41	113	69	69	69	82
17	150	188	188	75	219	42	118	118	118	150	187
18	104	170	73	65	102	43	58	62	62	74	196
19	607	383	400	216	249	44	63	63	63	63	94
20	80	91	94	71	59	45	104	46	46	46	108
21	25	25	25	25	38	46	107	101	101	44	44
22	228	179	285	236	235	47	45	43	38	38	40
23	366	339	390	392	408	48	50	45	45	45	49
24	163	163	163	74	74	49	60	60	60	60	60
25	215	266	269	149	61	50	206	247	207	249	281

	$\gamma = 0.2$	$\gamma = 0.4$	$\gamma = 0.5$	$\gamma = 0.6$	$\gamma = 0.8$
Average #	152	152	160	159	176

Table 3.3: Number of Iterations vs. γ , $\rho = 10$

Test #	$M = 1$	$M = 4$	$M = 7$	$M = 10$	Test #	$M = 1$	$M = 4$	$M = 7$	$M = 10$
1	122	122	122	122	26	157	70	83	85
2	40	64	66	62	27	167	191	191	191
3	143	67	67	117	28	27	43	44	41
4	91	87	87	87	29	323	323	323	323
5	36	51	83	77	30	126	126	126	126
6	81	72	105	280	31	482	259	259	259
7	60	53	53	53	32	104	122	123	123
8	140	157	157	157	33	239	332	179	179
9	263	261	310	814	34	148	279	94	94
10	195	173	329	330	35	41	41	41	42
11	95	74	67	74	36	103	107	105	106
12	474	130	239	239	37	73	123	130	93
13	80	76	76	76	38	104	95	67	67
14	399	353	709	84	39	59	167	52	108
15	233	209	168	168	40	203	371	371	371
16	48	330	334	334	41	101	80	76	69
17	76	188	188	188	42	95	108	118	118
18	58	56	65	73	43	50	54	54	62
19	225	400	400	400	44	43	47	63	63
20	83	63	54	94	45	48	46	46	46
21	25	25	25	25	46	36	51	73	101
22	255	377	313	285	47	32	33	38	38
23	419	451	428	390	48	44	45	45	45
24	295	163	163	163	49	79	60	60	60
25	122	60	269	269	50	202	287	249	207

	$M = 1$	$M = 4$	$M = 7$	$M = 10$
Average #	143	151	158	160

Table 3.4: Number of Iterations vs. M

3.3.2.2 The Impact of M

In [7], they reported to use best value of M to avoid any performance degradation when Armijo-type rule with GLL line search is applied. We study the $\rho = 10$ case for various M values.

Table 3.4 reports the number of iterations the SG algorithm takes to converge for $M = 1, 4, 7, 10$ for the same 50 test problems. We can see in Table 3.4 that we do not increase too much computational burden when choosing $M = 10$. It usually takes the least number of iterations for the algorithm to converge when $M = 1$. This is not surprising since the

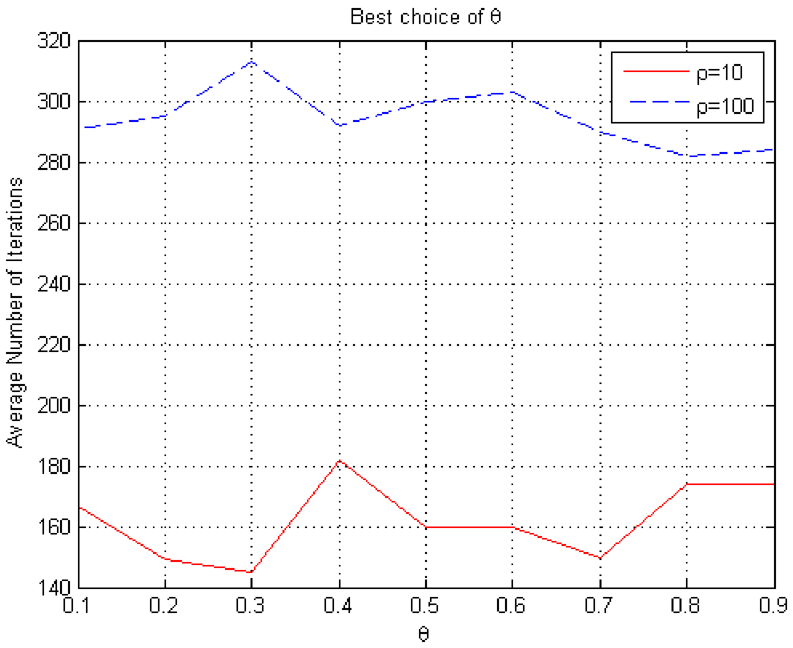

Figure 3.1: Impact of Shrinking Factor θ

algorithm becomes monotone and the condition number of the Hessian is only 3. We take a conservative value $M = 10$ to handle the more ill-posed problems.

3.3.2.3 The Impact of Shrinking Factor θ in Line Search

We tried different shrinking factor θ for $\rho = 10$ and 100 case, respectively. The average number of iterations for $\theta = 0.1, 0.2, \cdots, 0.9$ and $\rho = 10, 100$ against θ are plotted in Figure 3.1.

We choose $\theta = 0.5$ since we observe that there is no significant difference in terms of number of iterations when θ varies.

3.3.2.4 Lower and Upper Bound on the Stepsize

It is more important to set a proper lower bound than an upper bound on the stepsize as observed from our simulation results. Indeed, when $\rho \leq 100$, $\alpha_{\min} = 0.01$ works very well. When $\rho > 100$, none of the components of the optimal solution are binding at zero. This means the quadratic term is overwhelming compared to the $\mathcal{L}_1$ regularization term. The

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 $\left|f\left(x_k - \theta\alpha_k^{\overline{BB}} g_k\right) - f(x_k)\right| < tol$ 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
        $$Proj(x_k - \alpha_k^{\overline{BB}} g_k)$$
        where $Proj(x)$ is the orthogonal projection of $x$ onto the boundary
        of the feasible region $\mathcal{F} := \{x : \|x\|_1 \leq t\}$.
Step 3: Calculate
        $$f_{\max} \triangleq \max\{f(x_{k-j})\,|0 \leq j \leq \min\{k, M-1\}\}$$
        and
        $$\delta \leftarrow \left\langle g_k, Proj(x_k - \alpha_k^{\overline{BB}} g_k) - x_k\right\rangle$$
        and set $\theta \leftarrow 1$
Step 4: While $f\left(x_k - \theta\alpha_k^{\overline{BB}} g_k\right) > f_{\max} + \gamma\theta\delta$
        {set $\theta_{new} \in [\tau_1\theta, \tau_2\theta]$, $\theta \leftarrow \theta_{new}$}
Step 5: $x_{k+1} = x_k - \theta\alpha_k^{\overline{BB}} g_k$ and goto step 1
```

regularization effect is compromised by large ρ.

3.3.3 Performance Comparison

We compare our algorithm against the SPG method [9]. The detailed implementation of the SPG algorithm is given in Algorithm 3.2 where $Proj(x)$ is the orthogonal projection of x onto the boundary of the feasible region whenever it lies outside. For the SPG method in [9], projection operation is performed at most once in each iteration because the adaptive line search is performed on the line segment connecting the current iterate to the projected trial point and the feasible region $\mathcal{F} = \{x : \|x\|_1 \leq t\}$ is convex.

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

The equivalent optimization for the SPG method to work on is given by

$$\begin{aligned} & \min_{x\in\mathbb{R}^n} \rho \cdot \left(\tfrac{1}{2}x^T Hx + b^T x\right) \\ st. \quad & \|x\|_1 \leq t \end{aligned}$$

where $t = \|x^*\|_1$ and x^* is computed using our proposed scaled gradient method. We observed that when our proposed SG method and Raydan's SPG method are able to find the minimum x^* of (3.6), the optimal objective value differs by $\|x^*\|_1$ as expected.

We compare the number of iterations each algorithm takes to converge. We observe that, for $\rho = 100$, the SPG method only works for 9 out of 50 test problems. The number of iterations for those successful executions are recorded in Table 3.5. The SPG method fails to converge for the other 41 test problems.

In Table 3.5, it shows that for the 9 problems that the SPG method are successful, our SG method usually takes more iterations to converge but the SPG method does not always find the minimum under the parameters setting.

For the test case #20, the SPG and SG method both converge at the same optimal solution. The objective values versus iteration number for both the SPG and SG method are plotted in Figure 3.2. In Figure 3.2, we can see that both the SPG and SG method are non-monotone algorithm. The SPG method outperforms SG method in terms of convergence speed for this test problem.

Although the SPG method, when it works, seems to require fewer iterations to converge, it takes projection step on certain iterations which is not cheap. For many cases, under the same parameter setting of $M, \alpha_{\min}, \alpha_{\max}, \gamma$, 41 out of 50 test cases failed to converge with

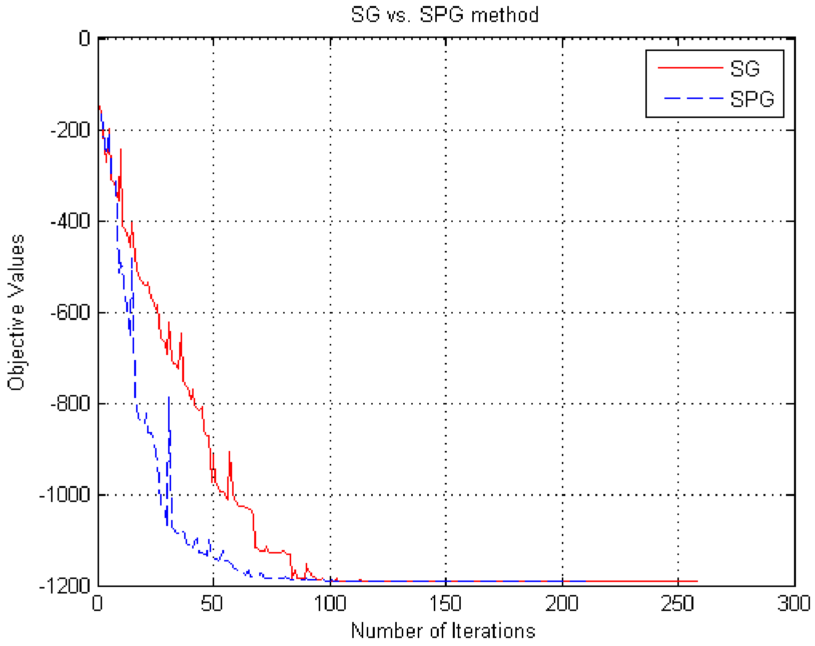

Figure 3.2: Trajectory of Objective Values for Test # 20, SG vs. SPG Method

the last steplength approaching zero. For $\rho = 10$, SPG method does not solve any of those 50 test problems.

3.4 Conclusions

In this chapter, we have proposed a scaled steepest descent method with non-monotone line search to ensure convergence. This affine scaling gradient based method is a non-monotone algorithm using the safeguarded Barzilai-Borwein stepsize along the scaled steepest descent direction. 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 ($\rho \geq 100$). We also compared our algorithm against the SPG method [9] and we observed that although SPG may converge faster than SG method, it is very sensitive to the settings of the parameters. Most of the time, SPG does not work for our sample test problems.

Chapter 4

Convergence of the Trust Region Method [26]

The scaled steepest descent method in Chapter 3 is simple, but it may be ineffective for nonconvex and very nonlinear problems. Coleman, Li and Wang proposed an affine scaling trust region method in [26] to solve a minimization problem for a twice continuously differentiable function with an $\mathcal{L}_1$-norm. This trust region algorithm as shown in Algorithm 4.1 is split up into 2 parts: step calculation and trust region size update. For steepest descent type algorithms, the complexity is on the order of $\mathcal{O}\left(n\right)$. In contrast, for trust region type algorithms, the complexity is at least $\mathcal{O}\left(n^3\right)$. However, higher rate of convergence is expected to be achieved.

Trust region method approximates a Newton step within a certain region in which model function (often a quadratic) is a good approximation to the objective function. Agreement factor is used to measure how close the current model function approximates the objective function. If the approximation is good, the trust region is expanded. Conversely, if the approximation is poor then the region is contracted.

For the simplicity of the proof, we define a vector $v\left(x\right) \in \mathbb{R}^n$ as

$$\left(v\left(x\right)\right)_i := \begin{cases} 1 & i = j^* := argmax_{j \in \{1, \cdots, n\} \wedge \left|\left(\nabla f\left(x\right)\right)_j\right| > 1}(\left|\left(\nabla f\left(x\right)\right)_j\right|) \\ \left|x_i\right| & otherwise \end{cases} \tag{4.1}$$

and the corresponding scaling matrix as $D\left(x\right) := diag\left(v\left(x\right)\right) \in \mathbb{R}_+^{n \times n}$. Note that $v\left(x\right)$ is a vector of the distance to each axis from the current iterate other than the component whose scaling factor is defined to be one. The component $\left(v\left(x\right)\right)_{j^*}$ corresponds to the largest violation of the condition $\left|\nabla f\left(x\right)\right| \leq 1$.

Define

$$\begin{aligned} s_k &= D_k^{-\frac{1}{2}} d_k \\ \hat{g}_k &= D_k^{\frac{1}{2}} g_k \\ \hat{M}_k &= D_k^{\frac{1}{2}} \nabla^2 f\left(x_k\right) D_k^{\frac{1}{2}} + diag\left(g_k\right) \cdot J_k^v \end{aligned} \tag{4.2}$$

we get the Newton direction s_k in the scaled space satisfying

$$\left(J_k^v \cdot diag\left(g_k\right) + diag\left(v_k\right) \cdot \nabla^2 f\left(x_k\right)\right) d_k = -diag\left(v_k\right) g_k$$

which is equivalent to

$$D_k^{-\frac{1}{2}} \left(J_k^v \cdot diag\left(g_k\right) D_k^{\frac{1}{2}} + D_k \nabla^2 f\left(x_k\right) D_k^{\frac{1}{2}}\right) D_k^{-\frac{1}{2}} d_k = -D_k^{\frac{1}{2}} g_k$$

or

$$\hat{M}_k s_k = -\hat{g}_k.$$

Thus we consider a trust region sub-problem at each iteration in the scaled space as given by

$$\min_{s_k \in \mathbb{R}^n} \left\{\hat{\psi}_k\left(s_k\right) : \left\|s_k\right\|_2 \leq \Delta_k\right\} \tag{4.3}$$

where

$$\hat{\phi}_k\left(s_k\right) := \hat{g}_k^T s_k + \frac{1}{2} s_k^T \hat{M}_k s_k.$$

A quadratic approximation function for the original objective function $f\left(x\right) + \left\|x\right\|_1$ is given below

$$\phi_k\left(d\right) = \nabla f\left(x_k\right)^T d + \left\|x_k + d\right\|_1 - \left\|x_k\right\|_1 + \frac{1}{2} d^T M_k d. \tag{4.4}$$

In a similar fashion as defined in Coleman, Li and Wang's paper [26], we define

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

and

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

where $\beta_k^0 = 0$ and $\beta_k^i, i \in \{1, \cdots, l_k\}$ defines the i-th break point along d_k , l_k is the index for the last break point and j_i is the component index of the i-th break point, i.e., $(\left(x_k\right)_{j_i} + \beta_k^i \left(d_k\right)_{j_i} = 0)$ in the original space where

$$\beta_k^i := \min_{j_i \in \{1, \cdots, n\} \setminus \{j_1, \cdots, j_{i-1}\}} \left\{-\frac{\left(x_k\right)_{j_i}}{\left(d_k\right)_{j_i}} : sign\left(\left(x_k\right)_{j_i}\right) = sign\left(\left(d_k\right)_{j_i}\right)\right\} = \frac{\left|\left(x_k\right)_{j_i}\right|}{\left|\left(d_k\right)_{j_i}\right|}$$

4.1 Proof Highlights

The chapter is dedicated to establishing the global and local convergence of the trust region method proposed in [26] for minimizing nonlinear function with the $\mathcal{L}_1$-norm regularization. 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 $f\left(x\right) + \left\|x\right\|_1$ is compact over $\mathcal{F}$.

Algorithm 4.1 Affine Scaling Trust Region Algorithm

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

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

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

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

where $dif\left\{\mathcal{F}\right\}$ is defined to be the union of the region in $\mathcal{F}$ that is
differentiable.
Step $3$.Compute

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

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.

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

Lemma 4.1 and 4.3, Theorem 4.5 and Theorem 4.6 prove that the first-order necessary optimality condition is satisfied assuming Assumption 1 to 4 hold. Lemma 4.7 to Theorem 4.11 are used to show that the second-order necessary optimality conditions are also satisfied. We prove next in Subsection 4.2.1 and 4.2.2 the first and the second-order necessary optimality conditions are satisfied, respectively.

4.2 Proof of the Convergence for the Trust Region Method

We make the following assumptions for the proof of the convergence.

Assumption 1: Given an initial point $x_0 \in \mathbb{R}^n$ and assume that $(x_0)_i \neq 0, \forall i \in \{1, \cdots, n\}$, 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 χ_B such that $\|B_k\| \leq \chi_B, \forall k$.

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

Assumption 4: Assume that $\phi(d_k) < \beta_g \phi_k^* [-D_k g_k]$, $\left\| D_k^{-\frac{1}{2}} d_k \right\|_2 \leq \Delta_k$, $x_k + d_k \in dif(\mathcal{F})$ where $\beta_g > 0$.

We require that $\phi(d_k)$ be less than a fraction of the minimum of $\phi_k(-D_k g_k)$ along the scaled gradient $-D_k g_k$, within the differentiable trust region. This assumption can be easily satisfied since that we can solve the following trust region problem

$$\min \left\{ \phi_k(d) : D_k^{-1} d = -\nu g_k, \left\| D_k^{-\frac{1}{2}} d \right\|_2 \leq \Delta_k, x_k + d \in dif(\mathcal{F}) \right\} \tag{4.5}$$

where ν is some positive number. Let d_k be the optimal solution to (4.5) with a possible small step-back to maintain strict differentiability and we can see that assumption 4 holds for such d_k.

4.2.1 First-Order Necessary Optimality Conditions

Let β_k^i denote the stepsize from the current iterate x_k to the i-th break point where $i \in \{0,1,2\}$. Notice that $\beta_k^0 = 0$ and β_k^2 could be infinity. Our piecewise quadratic approximation model is given by

$$\phi_k(\alpha d_k) := \alpha \nabla f(x_k)^T d_k + \frac{1}{2}\alpha^2 d_k^T M_k d_k + \|x_k + \alpha d_k\|_1 - \|x_k\|_1 \tag{4.6}$$

Suppose that $\alpha \in \left[\beta_k^i, \beta_k^{i+1}\right), \forall i \in \{0,1\}$, we observe that

$$\begin{aligned}
& \|x_k + \alpha d_k\|_1 - \|x_k\|_1 \\
= \ & \|x_k + \alpha d_k\|_1 - \left\|x_k + \beta_k^i d_k\right\|_1 + \left\|x_k + \beta_k^i d_k\right\|_1 - \left\|x_k + \beta_k^{i-1} d_k\right\|_1 \cdots + \left\|x_k + \beta_k^1 d_k\right\|_1 - \|x_k\|_1 \\
= \ & sign\left(x_k + \beta_k^{i^+} d_k\right)^T \left(x_k + \alpha d_k - \left(x_k + \beta_k^i d_k\right)\right) + \cdots + sign\left(x_k\right)^T \left(x_k + \beta_k^1 d_k - x_k\right) \\
= \ & sign\left(x_k + \beta_k^{i+} d_k\right)^T \left(\alpha - \beta_k^i\right) d_k + \sum_{j=1}^{i} sign\left(x_k + \beta_k^{(j-1)+} d_k\right)\left(\beta_k^j - \beta_k^{j-1}\right) d_k \qquad (4.7)
\end{aligned}$$

where β_k^{i+} is used to indicate that, immediately after crossing the i-th break point along d_k, what sign should the point $x_k + \alpha d_k$ take where $\alpha \in \left(\beta_k^i, \beta_k^{i+1}\right)$. Define $\phi_k^j(d) := \left(g_k^{j-1}\right)^T d + \frac{1}{2} d^T M_k d, \forall j \in \{1,2\}$ where g_k^j is the gradient immediately after crossing the j-th break point. That is,

$$g_k^j := \nabla f(x_k) + sign\left(x_k + \beta_k^{j+} d_k\right) + \beta_k^j M_k d_k \tag{4.8}$$

We get

$$\phi_k(\alpha d_k) = \phi_k^{i+1}\left(\left(\alpha - \beta_k^i\right) d_k\right) + \sum_{j=1}^{i} \phi_k^j\left(\left(\beta_k^j - \beta_k^{j-1}\right) d_k\right), \alpha \in \left[\beta_k^i, \beta_k^{i+1}\right) \tag{4.9}$$

Notice that

$$\begin{aligned}
g_k^i &= \nabla f(x_k) + sign\left(x_k + \beta_k^{i+} d_k\right) + \beta_k^i M_k d_k \\
&= g_k^{i-1} + \left(\beta_k^i - \beta_k^{i-1}\right) M_k d_k - 2 sign\left(x_{kj}\right) e_n^j \qquad (4.10)
\end{aligned}$$

where j corresponds to the axis for the i-th break point and $e_n^j = \begin{bmatrix} 0 & \cdots & \underbrace{1}_{j-th\, component} & \cdots & 0 \end{bmatrix}^T$ and x_{kj} is the j-th component of x_k .

Lemma 4.1 is intended to show that once Assumption 1 to 4 hold, the trust region algorithm will achieve a sufficient decrease which is bounded from below by some positive number, if first order optimality condition does not hold asymptotically.

Lemma 4.1. *Assume assumption* $1-3$ *hold and* d_k *satisfies assumption* 4*, let* χ *be the minimum*

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

then

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

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

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

and

$$\mu_k := d_k^T M_k d_k = s_k^T \hat{M}_k s_k$$

Proof. Suppose that the i-th break point lies on the hyperplane $x_j = 0$. We consider three possible cases below.

Case I: The optimizer along the negative scaled gradient direction $s_k = -\frac{\hat{g}_k^0}{\left\|\hat{g}_k^0\right\|}$ is strictly in the interior of two consecutive break points $\left(\beta_k^0, \beta_k^1\right)$ or $\left(\beta_k^1, \beta_k^2\right)$. i.e., $\alpha^* \in \left(\beta_k^{i^*}, \min\left\{\Delta_k, \beta_k^{i^*+1}\right\}\right)$.

This implies $\left\langle g_k^0, g_k^{i^*} \right\rangle > 0$ because of crossing of the break point $\beta_k^{i^*}$. We have

$$\begin{aligned}
\phi_k\left(\alpha d_k\right)' &= \phi_k^{i^*+1}\left(\left(\alpha-\beta_k^{i^*}\right)d_k\right)' + \sum_{j=1}^{i^*}\phi_k^j\left(\left(\beta_k^j-\beta_k^{j-1}\right)d_k\right)' \\
&= \left(\left(g_k^{i^*}\right)^T\cdot\left(\alpha-\beta_k^{i^*}\right)d_k + \frac{1}{2}\left(\alpha-\beta_k^{i^*}\right)^2 d_k^T M_k d_k\right)' \qquad (4.11) \\
&= \left(g_k^{i^*}\right)^T d_k + \left(\alpha-\beta_k^{i^*}\right)\mu_k
\end{aligned}$$

Substitute $d_k = D_k^{\frac{1}{2}} s_k = -D_k^{\frac{1}{2}} \frac{\hat{g}_k^0}{\|\hat{g}_k^0\|}$ into the above equation and derivative equals to zero at the minimizer, we can get

$$\alpha^* = \beta_k^{i^*} + \frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle}{\mu_k \left\|\hat{g}_k^0\right\|} > \beta_k^{i^*} \qquad (4.12)$$

which also implies $\mu_k \geq 0$. Therefore the optimal value is

$$\begin{aligned}
\phi\left(\alpha^* d_k\right) &= \sum_{j=1}^{i^*}\phi_k^j\left(\left(\beta_k^j-\beta_k^{j-1}\right)d_k\right) - \frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle^2}{\mu_k \left\|\hat{g}_k^0\right\|^2} + \frac{1}{2}\left(\frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle}{\mu_k \left\|\hat{g}_k^0\right\|}\right)^2 \mu_k \\
&= \sum_{j=1}^{i^*}\phi_k^j\left(\left(\beta_k^j-\beta_k^{j-1}\right)d_k\right) - \frac{1}{2}\frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle^2}{\mu_k \left\|\hat{g}_k^0\right\|^2} \qquad (4.13)
\end{aligned}$$

For each $j \in \{1, \cdots, i^*\}$, we have

$$\phi_k^j\left(\left(\beta_k^j-\beta_k^{j-1}\right)d_k\right) = -\left(\beta_k^j-\beta_k^{j-1}\right)\frac{\left\langle \hat{g}_k^{j-1}, \hat{g}_k^0 \right\rangle}{\left\|\hat{g}_k^0\right\|} + \frac{1}{2}\left(\beta_k^j-\beta_k^{j-1}\right)^2 \mu_k$$

and notice that $\beta_k^{j-1} + \frac{\langle \hat{g}_k^{j-1}, \hat{g}_k^0 \rangle}{\mu_k \|\hat{g}_k^0\|} > \beta_k^j$ (The minimum is not in $\left[\beta_k^{j-1}, \beta_k^j\right)$), so we can get

$$\begin{aligned}
& \phi_k^j\left(\left(\beta_k^j - \beta_k^{j-1}\right) d_k\right) \\
= & -\left(\beta_k^j - \beta_k^{j-1}\right) \frac{\left\langle \hat{g}_k^{j-1}, \hat{g}_k^0 \right\rangle}{\|\hat{g}_k^0\|} + \frac{1}{2}\left(\beta_k^j - \beta_k^{j-1}\right) \underbrace{\left(\beta_k^j - \beta_k^{j-1}\right)\mu_k}_{\left(\beta_k^j - \beta_k^{j-1}\right)\mu_k < \frac{\left\langle \hat{g}_k^{j-1}, \hat{g}_k^0 \right\rangle}{\|\hat{g}_k^0\|}} \\
\leq & -\left(\beta_k^j - \beta_k^{j-1}\right) \frac{\left\langle \hat{g}_k^{j-1}, \hat{g}_k^0 \right\rangle}{\|\hat{g}_k^0\|} + \frac{1}{2}\left(\beta_k^j - \beta_k^{j-1}\right) \frac{\left\langle \hat{g}_k^{j-1}, \hat{g}_k^0 \right\rangle}{\|\hat{g}_k^0\|} \qquad (4.14) \\
= & -\frac{1}{2}\left(\beta_k^j - \beta_k^{j-1}\right) \frac{\left\langle \hat{g}_k^{j-1}, \hat{g}_k^0 \right\rangle}{\|\hat{g}_k^0\|}
\end{aligned}$$

Therefore we get an upper bound on $\phi_k(\alpha^* d_k)$ when $\beta_k^{i^*} + \frac{\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \rangle}{\mu_k \|\hat{g}_k^0\|} < \beta_k^{i^*+1}, \forall i^* \in \{0, 1\}$

$$\phi(\alpha^* d_k) \leq -\frac{1}{2}\sum_{j=1}^{i^*}\left(\beta_k^j - \beta_k^{j-1}\right)\frac{\left\langle \hat{g}_k^{j-1}, \hat{g}_k^0 \right\rangle}{\|\hat{g}_k^0\|} - \frac{1}{2}\frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle^2}{\mu_k \|\hat{g}_k^0\|^2} \qquad (4.15)$$

Case II: The local minimum along the scaled steepest descent direction is located at the boundary of the trust region. Assume that $\Delta_k \in \left(\beta_k^{i^*}, \beta_k^{i^*+1}\right), i^* \in \{0, 1\}$, similarly we have $\langle g_k^0, g_k^1 \rangle, \cdots, \langle g_k^0, g_k^{i^*} \rangle > 0$, $\beta_k^{i^*} + \frac{\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \rangle}{\mu_k \|\hat{g}_k^0\|} \geq \Delta_k > \beta_k^{i^*}$ and $\beta_k^{j-1} + \frac{\langle \hat{g}_k^{j-1}, \hat{g}_k^0 \rangle}{\mu_k \|\hat{g}_k^0\|} > \beta_k^j, \forall j \in \{1, \cdots i^*\}$ when $\mu_k > 0$. Thus we can derive the following

$$\begin{aligned}
\phi_k(\alpha^* d_k) = & \sum_{j=1}^{i^*} \phi_k^j\left(\left(\beta_k^j - \beta_k^{j-1}\right) d_k\right) - \left(\Delta_k - \beta_k^{i^*}\right) \frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle}{\|\hat{g}_k^0\|} \\
& + \frac{1}{2}\left(\Delta_k - \beta_k^{i^*}\right) \underbrace{\left(\Delta_k - \beta_k^{i^*}\right)\mu_k}_{\left(\Delta_k - \beta_k^{i^*}\right)\mu_k \leq \frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle}{\|\hat{g}_k^0\|}} \\
\leq & -\frac{1}{2}\sum_{j=1}^{i^*}\left(\beta_k^j - \beta_k^{j-1}\right)\frac{\left\langle \hat{g}_k^{j-1}, \hat{g}_k^0 \right\rangle}{\|\hat{g}_k^0\|} - \left(\Delta_k - \beta_k^{i^*}\right) \frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle}{\|\hat{g}_k^0\|} \qquad (4.16) \\
& + \frac{1}{2}\left(\Delta_k - \beta_k^{i^*}\right) \frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle}{\|\hat{g}_k^0\|} \\
= & -\frac{1}{2}\sum_{j=1}^{i^*}\left(\beta_k^j - \beta_k^{j-1}\right)\frac{\left\langle \hat{g}_k^{j-1}, \hat{g}_k^0 \right\rangle}{\|\hat{g}_k^0\|} - \frac{1}{2}\left(\Delta_k - \beta_k^{i^*}\right) \frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle}{\|\hat{g}_k^0\|}
\end{aligned}$$

Next we show that the same result holds when $\mu_k \leq 0$. This is because we have $\langle g_k^0, g_k^1 \rangle, \cdots, \langle g_k^0, g_k^{i^*} \rangle > 0$ and $\beta_k^{j-1} + \frac{\langle \hat{g}_k^{j-1}, \hat{g}_k^0 \rangle}{\mu_k \|\hat{g}_k^0\|} < \beta_k^{j-1}, \forall j \in \{1, \cdots i^* + 1\}$. Notice that

$$
\begin{aligned}
& \phi_k^j \left(\left(\beta_k^j - \beta_k^{j-1} \right) d_k \right) \\
= \; & -\left(\beta_k^j - \beta_k^{j-1} \right) \frac{\langle \hat{g}_k^{j-1}, \hat{g}_k^0 \rangle}{\|\hat{g}_k^0\|} + \frac{1}{2} \underbrace{\left(\beta_k^j - \beta_k^{j-1} \right)^2 \mu_k}_{\leq 0} \\
\leq \; & -\frac{1}{2} \left(\beta_k^j - \beta_k^{j-1} \right) \frac{\langle \hat{g}_k^{j-1}, \hat{g}_k^0 \rangle}{\|\hat{g}_k^0\|}
\end{aligned}
$$

Hence we have when $\mu_k \leq 0$

$$
\begin{aligned}
\phi_k \left(\alpha^* d_k \right) = \; & \sum_{j=1}^{i^*} \phi_k^j \left(\left(\beta_k^j - \beta_k^{j-1} \right) d_k \right) - \left(\Delta_k - \beta_k^{i^*} \right) \frac{\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \rangle}{\|\hat{g}_k^0\|} \\
\leq \; & + \frac{1}{2} \left(\Delta_k - \beta_k^{i^*} \right)^2 \mu_k \\
& - \frac{1}{2} \sum_{j=1}^{i^*} \left(\beta_k^j - \beta_k^{j-1} \right) \frac{\langle \hat{g}_k^{j-1}, \hat{g}_k^0 \rangle}{\|\hat{g}_k^0\|} - \frac{1}{2} \left(\Delta_k - \beta_k^{i^*} \right) \frac{\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \rangle}{\|\hat{g}_k^0\|}
\end{aligned}
\tag{4.17}
$$

Case III: In the last case we assume that the minimizer equals one of the break points $\left\{ \beta_k^1, \cdots \beta_k^{i^*+1} \right\}$. WLOG we assume $\alpha^* = \beta_k^{i^*+1}, i^* \in \{0, \cdots, 1\}$. We still have $\langle g_k^0, g_k^1 \rangle, \cdots, \langle g_k^0, g_k^{i^*} \rangle > 0$ and $\beta_k^{j-1} + \frac{\langle \hat{g}_k^{j-1}, \hat{g}_k^0 \rangle}{\mu_k \|\hat{g}_k^0\|} > \beta_k^j, \forall j \in \{1, \cdots, i^* + 1\}$ when $\mu_k > 0$. Thus, when $\mu_k > 0$,

$$
\begin{aligned}
\phi_k \left(\alpha^* d_k \right) = \; & \sum_{j=1}^{i^*} \phi_k^j \left(\left(\beta_k^j - \beta_k^{j-1} \right) d_k \right) - \left(\beta_k^{i^*+1} - \beta_k^{i^*} \right) \frac{\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \rangle}{\|\hat{g}_k^0\|} \\
& + \frac{1}{2} \left(\beta_k^{i^*+1} - \beta_k^{i^*} \right) \quad \underbrace{\left(\beta_k^{i^*+1} - \beta_k^{i^*} \right) \mu_k}_{\left(\beta_k^{i^*+1} - \beta_k^{i^*} \right) \mu_k \leq \frac{\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \rangle}{\|\hat{g}_k^0\|}} \\
\leq \; & -\frac{1}{2} \sum_{j=1}^{i^*} \left(\beta_k^j - \beta_k^{j-1} \right) \frac{\langle \hat{g}_k^{j-1}, \hat{g}_k^0 \rangle}{\|\hat{g}_k^0\|} - \frac{1}{2} \left(\beta_k^{i^*+1} - \beta_k^{i^*} \right) \frac{\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \rangle}{\|\hat{g}_k^0\|}
\end{aligned}
\tag{4.18}
$$

when $\mu_k > 0$.

For $\mu_k \leq 0$, we still get

$$\begin{aligned}
& \phi_k\left(\alpha^* d_k\right) && (4.19) \\
= \;& \sum_{j=1}^{i^*} \phi_k^j\left(\left(\beta_k^j - \beta_k^{j-1}\right) d_k\right) - \left(\beta_k^{i^*+1} - \beta_k^{i^*}\right) \frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle}{\left\|\hat{g}_k^0\right\|} + \frac{1}{2}\left(\beta_k^{i^*+1} - \beta_k^{i^*}\right)^2 \mu_k \\
\leq \;& -\frac{1}{2} \sum_{j=1}^{i^*} \left(\beta_k^j - \beta_k^{j-1}\right) \frac{\left\langle \hat{g}_k^{j-1}, \hat{g}_k^0 \right\rangle}{\left\|\hat{g}_k^0\right\|} - \frac{1}{2}\left(\beta_k^{i^*+1} - \beta_k^{i^*}\right) \frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle}{\left\|\hat{g}_k^0\right\|}
\end{aligned}$$

In conclusion, we have the following

$$-\phi\left(d_k\right) \geq -\beta_g \phi_k^*\left[-D_k g_k^0\right] \;\geq\; \frac{\beta_g}{2}\left\{\sum_{j=1}^{i^*}\left(\beta_k^j - \beta_k^{j-1}\right)\frac{\left\langle \hat{g}_k^{j-1}, \hat{g}_k^0 \right\rangle}{\left\|\hat{g}_k^0\right\|} + \chi\left(\mu_k, \Delta_k, \beta_k^{i^*}\right)\right\} \quad (4.20)$$

□

Notice that if the k-th iteration is successful, then

$$\rho_k^f = \frac{f\left(x_k + \alpha_k d_k\right) - f\left(x_k\right) + \left\|x_k + \alpha_k d_k\right\|_1 - \left\|x_k\right\|_1 + \frac{1}{2}\alpha_k^2 d_k^T C_k d_k}{\phi_k\left(\alpha_k d_k\right)} \geq \mu,$$

Hence

$$h\left(x_k + \alpha_k d_k\right) - h\left(x_k\right) \leq \mu\phi_k\left(\alpha_k d_k\right) - \underbrace{\frac{1}{2}\alpha_k^2 d_k^T C_k d_k}_{C_k \succeq 0} < \mu\phi_k\left(\alpha d_k\right) \quad (4.21)$$

Lemma 4.2 illustrates that if a Newton step within the trust region is taken, what decomposition properties does this global solution to the trust region sub-problem have.

Lemma 4.2. *Let p_k denote a global solution to the trust region problem*

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

Then there exists a parameter λ_k *and upper triangle matrix* $R_k \in \mathbb{R}^{n\times n}$ *such that*

$$\hat{M}_k + \lambda_k I = R_k^T R_k, \left(\hat{M}_k + \lambda_k I\right) D_k^{-\frac{1}{2}} p_k = -\hat{g}_k, \lambda_k \geq 0 \tag{4.22}$$

with $\lambda_k \left(\Delta_k - \left\|D_k^{-\frac{1}{2}} p_k\right\|\right) = 0$*. We can also write out above equation equivalently as*

$$\left(\lambda_k I + D_k^{\frac{1}{2}} C_k D_k^{\frac{1}{2}}\right) p_k = -D_k \left(g_k + B_k p_k\right) \tag{4.23}$$

Proof. Please refer to [12]. The Lagrangian of above problem is given by

$$\mathcal{L}\left(d, \lambda_k\right) = g_k^T d + \frac{1}{2} d^T M_k d + \lambda_k \left(d^T D_k^{-\frac{1}{2}} D_k^{-\frac{1}{2}} d - \Delta_k^2\right)$$

then

$$\nabla_d \mathcal{L}\left(d, \lambda_k\right) = g_k + M_k d + 2\lambda_k D_k^{-\frac{1}{2}} D_k^{-\frac{1}{2}} d = 0$$

hence we get

$$\left(M_k D_k^{\frac{1}{2}} + \lambda_k' D_k^{-\frac{1}{2}}\right) D_k^{-\frac{1}{2}} p_k = -g_k$$

where $\lambda_k' = 2\lambda_k$.

Pre-multiplying both sides with $D_k^{\frac{1}{2}}$ then we get

$$\left(\hat{M}_k + \lambda_k' I\right) D_k^{-\frac{1}{2}} p_k = -\hat{g}_k$$

and the complementary slackness condition is given by

$$\lambda_k' \left(\Delta_k - \left\|D_k^{-\frac{1}{2}} p_k\right\|\right) = 0$$

We need to show that $\hat{M}_k + \lambda_k' I$ is positive semi-definite. Since p_k solves

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

then it also solves

$$\min_{w_k \in \mathbb{R}^n} \left\{ \psi_k\left(w_k\right) : \left\| D_k^{-\frac{1}{2}} w_k \right\| = \left\| D_k^{-\frac{1}{2}} p_k \right\| \right\}$$

and it follows that

$$\begin{aligned} \psi_k\left(w_k\right) \geq \psi_k\left(p_k\right), & \qquad \forall w_k \\ st. \quad \left\| D_k^{-\frac{1}{2}} w_k \right\| = \left\| D_k^{-\frac{1}{2}} p_k \right\| & \end{aligned}$$

and together with

$$\left(\hat{M}_k + \lambda_k' I\right) D_k^{-\frac{1}{2}} p_k = -\hat{g}_k$$

we have

$$\psi_k\left(w_k\right) = \hat{g}_k^T D_k^{-\frac{1}{2}} w_k + \frac{1}{2}\left(D_k^{-\frac{1}{2}} w_k\right)^T \hat{M}_k \left(D_k^{-\frac{1}{2}} w_k\right) \geq \hat{g}_k^T D_k^{-\frac{1}{2}} p_k + \frac{1}{2}\left(D_k^{-\frac{1}{2}} p_k\right)^T \hat{M}_k \left(D_k^{-\frac{1}{2}} p_k\right) = \psi_k\left(p_k\right)$$

Substitute $\hat{g}_k = -\left(\hat{M}_k + \lambda_k' I\right) D_k^{-\frac{1}{2}} p_k$ into above and re-arrange the terms to get

$$\begin{aligned} & \frac{1}{2}\left(D_k^{-\frac{1}{2}} w_k - D_k^{-\frac{1}{2}} p_k\right)^T \left(\hat{M}_k + \lambda_k' I\right) \left(D_k^{-\frac{1}{2}} w_k - D_k^{-\frac{1}{2}} p_k\right) \\ \geq \; & \frac{\lambda_k'}{2}\left(\left(D_k^{-\frac{1}{2}} w_k\right)^T \left(D_k^{-\frac{1}{2}} w_k\right) - \left(D_k^{-\frac{1}{2}} p_k\right)^T \left(D_k^{-\frac{1}{2}} p_k\right)\right) \\ = \; & 0 \end{aligned}$$

Above inequality implies that

$$\left(\hat{M}_k + \lambda_k' I\right) \succeq 0$$

whenever $\left\| D_k^{-\frac{1}{2}} p_k \right\| \neq 0$. Otherwise if $\left\| D_k^{-\frac{1}{2}} p_k \right\| = 0$, then from $\left(\hat{M}_k + \lambda_k' I\right) D_k^{-\frac{1}{2}} p_k = -\hat{g}_k$ we get

$$\hat{g}_k = 0$$

Therefore $D_k^{-\frac{1}{2}} p_k = 0$ solves the following optimization problem

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

hence $\hat{M}_k \succeq 0$ otherwise 0 can not be the optimal value for the above optimization problem.

By complementary slackness $\lambda_k \left(\Delta_k - \left\| D_k^{-\frac{1}{2}} p_k \right\| \right) = 0$ we get $\lambda_k = \frac{1}{2} \lambda'_k = 0$ since $\Delta_k - \left\| D_k^{-\frac{1}{2}} p_k \right\| \neq 0$. In this case, we still have

$$\hat{M}_k = \hat{M}_k + \lambda'_k I \succeq 0$$

□

The following lemma provides a property of the trust region size when asymptotically the step is always successful.

Lemma 4.3. *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.*

Proof. By assumption, there exists $\bar{k}$ such that when $k \geq \bar{k}$, we always have $\rho_k^f \geq \eta$. We will prove by induction that for $k \geq \bar{k}$,

$$\Delta_k \geq \min \left\{ \gamma_0 \Lambda_l, \Delta_{\bar{k}} \right\}$$

Assume above inequality is true for $k \geq \bar{k}$. From Step 3 in trust region updating rule in Algorithm 4.1, if $\Delta_k \leq \Lambda_l$, we get

$$\Delta_{k+1} \geq \Delta_k \geq \min \left\{ \gamma_1 \Lambda_l, \Delta_{\bar{k}} \right\}.$$

If $\Delta_k > \Lambda_l$, we have

$$\Delta_{k+1} \geq \min \left\{ \gamma_1 \Lambda_l, \Delta_{\bar{k}} \right\}$$

Hence the trust region size is bounded away from zero. □

Lemma 4.4. Assume *that $f : \mathbb{R}^n \mapsto \mathbb{R}$ is continuously differentiable on $dif\left(\mathcal{F}\right)$ and assumptions $1 \sim 3$ hold. Assume* $\liminf_{k\to\infty} \|\hat{g}_k^0\| > 0$ *and strict complementarity condition holds, i.e., at the limit point x_* of $\{x_k\}$ generated by Algorithm* 4.1*,* $v\left(x_*\right)_j$ and $\nabla f\left(x_*\right)_j + sign\left(x_*\right)_j$ *can not both be zero in $v\left(x_*\right)_j^{\frac{1}{2}} \cdot \left(\nabla f\left(x_*\right)_j + sign\left(x_*\right)_j\right) = 0, \forall j \in \{1, \cdots, n\}$, then there exists an $\bar{\epsilon} > 0$ such that*

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

and

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

Moreover, if the k-th iteration is successful, for sufficiently large k, the following holds

$$h\left(x_k\right) - h\left(x_{k+1}\right) > \frac{1}{2}\beta_g\mu \cdot \frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle}{\|\hat{g}_k^0\|} \min\left\{ \frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle}{\mu_k \|\hat{g}_k^0\|}, \Delta_k - \beta_k^{i^*}, \beta_k^{i^*+1} - \beta_k^{i^*} \right\} = \frac{1}{2}\beta_g\mu\bar{\epsilon}\left(\Delta_k\right)$$

Proof. Suppose that there does not exist a subsequence such that $\lim_{k\to\infty} \beta_k^{i^*} = 0$ (use k instead of k_i for simplicity). By definition of $\beta_k^{i^*}$, we can get the following facts:

$$\beta_k^{i^*} \; : \; = \; \frac{\left|\left(x_k\right)_{j_{i^*}}\right| \cdot \|\hat{g}_k^0\|}{\left(D_k\right)_{j_{i^*}, j_{i^*}} \cdot \left|\left(g_k^0\right)_{j_{i^*}}\right|} \tag{4.24}$$

The assumption $\liminf_{k\to\infty} \beta_k^{i^*} > 0$ implies $i^* = 1$ for sufficiently large k since $\beta_k^0 = 0$. Then there exists an $\epsilon_0' > 0$ such that for sufficiently large k, $\beta_k^{i^*} \geq \epsilon_0'$. From (4.20) and (4.21) in Lemma 4.1 and $\chi\left(\mu_k, \Delta_k, \beta_k^{i^*}\right) \geq 0$, we get

$$h\left(x_k\right) - h\left(x_{k+1}\right) > \frac{1}{2}\beta_g\mu \left\|\hat{g}_k^0\right\| \beta_k^{i^*}.$$

Based on the assumption $\liminf_{k\to\infty} \|\hat{g}_k^0\| > 0$, there exists an $\epsilon_0 > 0$ such that, for sufficiently large k, $\|\hat{g}_k^0\| \geq \epsilon_0$. Hence

$$\lim_{k_i \to \infty} h(x_k) - h(x_{k+1}) \geq \lim_{k \to \infty} \frac{1}{2} \beta_g \mu \left\| \hat{g}_k^0 \right\| \beta_k^{i^*} \geq \frac{1}{2} \beta_g \mu \epsilon_0 \cdot \epsilon_0' > 0 \tag{4.25}$$

which contradicts to $\lim_{k \to \infty} h(x_k) - h(x_{k+1}) = 0$.

We show next that the subsequence corresponding to $\lim_{k \to \infty} \beta_k^{i^*} = 0$ has the property that $\frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle}{\left\| \hat{g}_k^0 \right\|}$ is bounded away from zero under strict complementarity condition, i.e., $\liminf_{k \to \infty} \frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle}{\left\| \hat{g}_k^0 \right\|} > 0$. For notational simplification, we still denote the subsequence by k.

Using (4.24), we can deduce that for sufficient large k, $(D_k)_{j_{i^*}, j_{i^*}} = 1$. Otherwise $\beta_k^{i^*} = \frac{\|\hat{g}_k^0\|}{\left| \left(g_k^0\right)_{j_{i^*}} \right|}$ is bounded away from zero. We can conclude that for sufficiently large k, $\left| \left(\nabla f(x_k)\right)_{j_p} \right| > 1, \forall p \in \{1, \cdots, i^*\}$. Hence we have

$$\begin{aligned} \left\| \hat{g}_k^0 \right\| : &= \sqrt{\sum_{i=1}^{n} \left(\hat{g}_k^0\right)_i^2} \\ &= \sqrt{\sum_{p \notin \{j_1, \cdots j_{i^*}\}} \left(\hat{g}_k^0\right)_p^2 + \sum_{p \in \{j_1, \cdots, j_{i^*}\}} \left(\hat{g}_k^0\right)_p^2} \\ &= \sqrt{\sum_{p \notin \{j_1, \cdots j_{i^*}\}} \left(\hat{g}_k^0\right)_p^2 + \sum_{p \in \{j_1, \cdots, j_{i^*}\}} \left(g_k^0\right)_p^2} \end{aligned} \tag{4.26}$$

Similarly,

$$\begin{aligned} \frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle}{\left\| \hat{g}_k^0 \right\|} &= \frac{1}{\left\| \hat{g}_k^0 \right\|} \left(\sum_{p \notin \{j_1, \cdots, j_{i^*}\}} \left(\hat{g}_k^{i^*}\right)_p \cdot \left(\hat{g}_k^0\right)_p + \sum_{p \in \{j_1, \cdots, j_{i^*}\}} \left(\hat{g}_k^{i^*}\right)_p \cdot \left(\hat{g}_k^0\right)_p \right) \\ &= \frac{1}{\left\| \hat{g}_k^0 \right\|} \left(\sum_{p \notin \{j_1, \cdots, j_{i^*}\}} \left(\hat{g}_k^0\right)_p^2 + \sum_{p \in \{j_1, \cdots, j_{i^*}\}} \left(g_k^{i^*}\right)_p \cdot \left(g_k^0\right)_p \right) \end{aligned} \tag{4.27}$$

where the second equality uses $(D_k)_{j_{i^*}, j_{i^*}} = 1$.

Note that

$$\begin{aligned} & \sum_{p\in\{j_1,\cdots,j_{i^*}\}} \left(g_k^{i^*}\right)_p \cdot \left(g_k^0\right)_p \\ = & \sum_{p\in\{j_1,\cdots,j_{i^*}\}} \left(\left(\nabla f\left(x_k\right)\right)_p - sign\left(\left(x_k\right)_p\right)\right) \cdot \left(\left(\nabla f\left(x_k\right)\right)_p + sign\left(\left(x_k\right)_p\right)\right) \\ = & \sum_{p\in\{j_1,\cdots,j_i\}} \underbrace{\left(\left(\nabla f\left(x_k\right)\right)_p^2 - 1\right)}_{>0} \\ > & \ 0 \end{aligned} \tag{4.28}$$

Bear in mind that for any $j_q \in \{j_1, \cdots, j_{i^*}\}$,

$$\begin{aligned} sign\left(\left(x_k\right)_{j_q}\right) &= sign\left(\left(g_k^0\right)_{j_q}\right) \\ &= sign\left(\underbrace{\left(\nabla f\left(x_k\right)\right)_{j_q}}_{|\cdot|>1} + \underbrace{sign\left(x_k\right)_{j_q}}_{|\cdot|=1}\right) \\ &= sign\left(\left(\nabla f\left(x_k\right)\right)_{j_q}\right) \end{aligned}$$

hence $\left|\left(g_k^0\right)_{j_q}\right| = \left|\left(\nabla f\left(x_k\right)\right)_{j_q} + sign\left(x_k\right)_{j_q}\right| > 2.$

We hereby examine the value of $\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle$ in the following two cases:

Case I: $\liminf_{k\to\infty} \sum_{p\in\{j_1,\cdots,j_{i^*}\}} \left(g_k^0\right)_p^2 > 0$. From (4.28), $\liminf_{k\to\infty} \frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle}{\left\|\hat{g}_k^0\right\|} = 0$ implies that

$$\liminf_{k\to\infty} \sum_{p\in\{j_1,\cdots,j_i\}} \left(\left(\nabla f\left(x_k\right)\right)_p^2 - 1\right) = 0$$

and

$$\liminf_{k\to\infty} \sum_{p\notin\{j_1,\cdots j_{i^*}\}} \left(\hat{g}_k^0\right)_p^2 = 0.$$

Recall that we are still under the assumption that $\lim_{k\to\infty} \beta_k^{i^*} = 0$. Since $\lim_{k\to\infty} \beta_k^{i^*} = 0$,

$x_{i^*} = 0$ and $\left|\nabla f^k_{j_{i^*}}\right| > 1$, under strict complementarity condition $\liminf_{k\to\infty} \left|\nabla f^k_{j_{i^*}}\right| > 1$, thus

$$\liminf_{k\to\infty} \sum_{p\in\{j_1,\cdots,j_{i^*}\}} \left(g_k^{i^*}\right)_p \cdot \left(g_k^0\right)_p > 0$$

This results a contradiction to $\liminf_{k\to\infty} \sum_{p\in\{j_1,\cdots,j_i\}} \left(\left(\nabla f\left(x_k\right)\right)_p^2 - 1\right) = 0$.

Hence

$$\liminf_{k\to\infty} \frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle}{\left\|\hat{g}_k^0\right\|} > 0.$$

Case II: $\liminf_{k\to\infty} \sum_{p\in\{j_1,\cdots,j_{i^*}\}} \left(g_k^0\right)_p^2 = 0$. By our assumption $\liminf_{k\to\infty} \left\|\hat{g}_k^0\right\| > 0$, this will imply

$$\liminf_{k\to\infty} \sum_{p\notin\{j_1,\cdots j_{i^*}\}} \left(\hat{g}_k^0\right)_p^2 > 0$$

hence $\liminf_{k\to\infty} \frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle}{\left\|\hat{g}_k^0\right\|} > 0$.

In both cases, we arrive at the same conclusion that $\liminf_{k\to\infty} \frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle}{\left\|\hat{g}_k^0\right\|} > 0$ if $\liminf_{k\to\infty} \left\|\hat{g}_k^0\right\| > 0$.

Therefore, for sufficiently large k_i, there exists $\bar{\epsilon} > 0$ such that $\frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle}{\left\|\hat{g}_k^0\right\|} \geq \bar{\epsilon}$. Thus we have

$$h\left(x_k\right) - h\left(x_{k+1}\right) > \frac{1}{2}\beta_g \mu \cdot \frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle}{\left\|\hat{g}_k^0\right\|} \min\left\{ \frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle}{\mu_k \left\|\hat{g}_k^0\right\|}, \Delta_k - \beta_k^{i^*}, \beta_k^{i^*+1} - \beta_k^{i^*} \right\} = \frac{1}{2}\beta_g \mu \bar{\epsilon}\left(\Delta_k\right) \tag{4.29}$$

The reason why the above minimum in (4.29) takes the form as Δ_k is because of $\lim_{k_i\to\infty} \beta_k^{i^*} = 0$, $\liminf_{k\to\infty} \left(\beta_k^{i^*+1} - \beta_k^{i^*}\right) > 0$ and $\liminf_{k\to\infty} \frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0 \right\rangle}{\mu_k \left\|\hat{g}_k^0\right\|} > 0$. □

Theorem 4.5. Assume *that* $f : \mathbb{R}^n \mapsto \mathbb{R}$ *is continuously differentiable on* $\mathcal{F}$ *and assumptions* $1 \sim 3$ *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*

$$\liminf_{k\to\infty} \left\|\hat{g}_k^0\right\| = 0$$

Proof. Suppose $\liminf_{k\to\infty} \left\|\hat{g}_k^0\right\| > 0$, assume that there exists an $\epsilon > 0$ such that $\left\|\hat{g}_k^0\right\| \geq \epsilon$

for sufficiently large k. We will show that $\sum_{k=1}^{\infty}\Delta_k < \infty$. Let's consider two cases:

Case I: If there are only finitely many successful iterations, then for sufficiently large $k \geq \bar{k}$, based on the trust region update rule, we have $\Delta_{k+1} \leq \gamma_1 \Delta_k$. Hence

$$\sum_{k=1}^{\infty}\Delta_k = \sum_{k=1}^{\bar{k}-1}\Delta_k + \sum_{k=\bar{k}}^{\infty}\Delta_k \leq \sum_{k=1}^{\bar{k}-1}\Delta_k + \Delta_{\bar{k}}\left(1+\gamma_1+\cdots\right) = \sum_{k=1}^{\bar{k}-1}\Delta_k + \frac{\Delta_{\bar{k}}}{1-\gamma_1} < \infty \tag{4.30}$$

Case II: Otherwise, suppose that there is an infinite sub-sequence $\{k_i\}$ of successful iterations. Since$\{h\left(x_k\right)\}_{k=1}^{\infty}$ is non-increasing by our algorithm, and is bounded from below by assumption 1, the limit of sequence $\{h\left(x_k\right)\}_{k=1}^{\infty}$ exists. Note that

$$0 \leq \sum_{k=0}^{\infty}\left(h\left(x_k\right) - h\left(x_{k+1}\right)\right) = h\left(x_0\right) - \lim_{k\to\infty} h\left(x_k\right) < \infty \tag{4.31}$$

By Lemma 4.4, we can see that, for subsequence $\{x_{k_i}\}$,

$$h\left(x_{k_i}\right) - h\left(x_{k_i+1}\right) > \frac{1}{2}\beta_g\mu\cdot\frac{\left\langle \hat{g}_{k_i}^{i^*}, \hat{g}_{k_i}^{0}\right\rangle}{\left\|\hat{g}_{k_i}^{0}\right\|}\min\left\{\frac{\left\langle \hat{g}_{k_i}^{i^*}, \hat{g}_{k_i}^{0}\right\rangle}{\mu_{k_i}\left\|\hat{g}_{k_i}^{0}\right\|}, \Delta_{k_i} - \beta_{k_i}^{i^*}, \beta_{k_i}^{i^*+1} - \beta_{k_i}^{i^*}\right\} = \frac{1}{2}\beta_g\mu\bar{\epsilon}\left(\Delta_{k_i}\right)$$

This also implies for sufficiently large k_i, namely, $k_i \geq \bar{k}_i$, we have

$$\sum_{\bar{k}_i}^{\infty}\Delta_{k_i} < \frac{2}{\beta_g\mu\bar{\epsilon}}\sum_{k_i=\bar{k}_i}^{\infty}\left(h\left(x_{k_i}\right) - h\left(x_{k_i+1}\right)\right) \leq \frac{2}{\beta_g\mu\bar{\epsilon}}\left(h\left(x_{\bar{k}_i}\right) - h\left(x_{\infty}\right)\right) < \infty$$

That is,

$$\sum_{i=1}^{\infty}\Delta_{k_i} < \infty \tag{4.32}$$

Because $\liminf_{k\to\infty}\beta_k^{i^*} = 0$, we have

$$\sum_{k=1}^{\infty}\Delta_k \leq \left(1 + \frac{\gamma_2}{1-\gamma_1}\right)\sum_{i=1}^{\infty}\left(\Delta_{k_i} - \beta_{k_i}^{i^*}\right) < \infty \tag{4.33}$$

We show next that this will imply $\left|\rho_k^f - 1\right| \to 0$.

Note that $\|x_{k+1}-x_k\| = \|d_k\| = \left\| D_k^{\frac{1}{2}} D_k^{-\frac{1}{2}} d_k \right\| = \left\| D_k^{\frac{1}{2}} s_k \right\| \leq \chi_D \Delta_k$. But$\left\| D_k^{\frac{1}{2}} \right\| \leq \chi_D$ is bounded from above based on the Assumption 1. We have already established that $\sum_{k=1}^{\infty} \Delta_k$ converges so that $\lim_{k\to\infty} \Delta_k = 0$. Since $\|x_{k+1}-x_k\| \leq \chi_D \Delta_k \to 0$ the sequence $\{x_k\}_{k=1}^{\infty}$ converges. We observe that for sufficiently large k, the trust region can be arbitrarily small.

Because $\left\| D_k^{-\frac{1}{2}} d_k \right\| = \|s_k\| \leq \Delta_k \to 0$ and $f(x_k)$ are twice continuously differentiable at differentiable point thus

$$f(x_{k+1}) - f(x_k) = (\nabla f(x_k + \xi_k d_k))^T d_k$$

we get

$$\begin{aligned} & \left| h(x_{k+1}) - h(x_k) + \tfrac{1}{2} d_k^T C_k d_k - \phi_k(d_k) \right| \\ = & \left| f(x_{k+1}) - f(x_k) - \tfrac{1}{2} d_k^T B_k d_k - \nabla f(x_k)^T d_k \right| \\ = & \left| (\nabla f(x_k + \xi_k d_k) - \nabla f(x_k))^T d_k - \tfrac{1}{2} d_k^T B_k d_k \right| \\ \leq & \tfrac{1}{2} \chi_B \chi_D^2 \Delta_k^2 + \chi_D \Delta_k \|\nabla f(x_k + \xi_k d_k) - \nabla f(x_k)\| \end{aligned}$$

Based on above inequality, the continuity of $\nabla f(x)$, the convergence of $\{x_k\}_{k=1}^{\infty}$ and $-\phi_k(d_k) \geq \frac{1}{2} \beta_g \bar{\epsilon} \Delta_k$ for sufficiently large k, we can conclude that $\left\{ \left| \rho_k^f - 1 \right| \right\}$ converges to zero.

From Lemma 4.3, $\rho_k^f \geq \eta$ for sufficiently large k, then $\{\Delta_k\}$ is bounded away from zero, which contradicts our former result $\Delta_k \to 0$. Hence $\liminf_{k\to\infty} \|\hat{g}_k^0\| = 0$. □

We will prove next a stronger claim that $\lim_{k\to\infty} \|\hat{g}_k^0\| = 0$.

Theorem 4.6. *Assume assumption* $1 \sim 3$ *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*

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

Proof. We prove this theorem by contradiction. Based on Theorem 4.5, we know that $\liminf_{k\to\infty}\|\hat{g}_k^0\| = 0$. If we can show that $\limsup_{k\to\infty}\|\hat{g}_k^0\| = 0$, then

$$\lim_{k\to\infty}\left\|\hat{g}_k^0\right\| = \liminf_{k\to\infty}\left\|\hat{g}_k^0\right\| = \limsup_{k\to\infty}\left\|\hat{g}_k^0\right\| = 0.$$

We assume that $\limsup_{k\to\infty}\|\hat{g}_k^0\| > 0$, i.e., there exists a sub-sequence indexed by $\{m_i\}$ such that, for sufficiently large m_i, $\left\|\hat{g}_{m_i}^0\right\| \geq \epsilon_1$ where $\epsilon_1 \in (0,1)$ is arbitrarily chosen. By Theorem 4.5, for any $\epsilon_2 \in (0,\epsilon_1)$, there must exist a sub-sequence pair $\left\{m_{i_j}, l_j\right\}$ such that (WLOG, assume $m_{i_j} = m_j$), for sufficiently large k, the following holds

$$\begin{cases} \|\hat{g}_k^0\| \geq \epsilon_2 & m_i \leq k < l_i \\ \|\hat{g}_k^0\| < \epsilon_2 & k = l_i \end{cases}$$

and $\lim_{l_i\to\infty}\left\|\hat{g}_{l_i}^0\right\| = 0$. Otherwise we get $\liminf_{k\to\infty}\|\hat{g}_k^0\| > 0$. For sufficiently large k (in the range $[m_i, l_i)$) and if the k-th iteration is successful, by (4.20) and (4.21) in Lemma 4.1 we get

$$h(x_k) - h(x_{k+1}) \quad > \quad \frac{1}{2}\beta_g\mu\frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0\right\rangle}{\|\hat{g}_k^0\|}\min\left\{\frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0\right\rangle}{\chi_{\hat{M}}\left\|\hat{g}_k^0\right\|}, \Delta_k - \beta_k^{i^*}, \beta_k^{i^*+1} - \beta_k^{i^*}\right\} \tag{4.34}$$

By Lemma 4.4, we can derive that $\|\hat{g}_k^0\| \geq \epsilon_2, k \in [m_i, l_i)$ implies $\frac{\left\langle \hat{g}_k^{i^*}, \hat{g}_k^0\right\rangle}{\left\|\hat{g}_k^0\right\|} \geq \bar{\epsilon}_2$ where $\bar{\epsilon}_2$ is a positive number. Therefore the decrease in (4.34) can be simplified as

$$h(x_k) - h(x_{k+1}) \geq \frac{1}{2}\beta_g\mu\bar{\epsilon}_2\min\left\{\frac{\bar{\epsilon}_2}{\chi_{\hat{M}}}, \Delta_k - \beta_k^{i^*}, \beta_k^{i^*+1} - \beta_k^{i^*}\right\}.$$

Because $h(x_k)$ is bounded below on $\mathcal{F}$ and the sequence $\{h(x_k)\}$ is non-decreasing, $\{h(x_k)\}$ converges and $\{h(x_k) - h(x_{k+1})\}$ converges to zero. In addition, $\lim_{k\to\infty}\beta_k^{i^*} = 0$

as in the proof of Lemma 4.4, for sufficiently large k, we have

$$h\left(x_k\right) - h\left(x_{k+1}\right) > \frac{1}{2}\beta_g \mu \bar{\epsilon}_2 \left(\Delta_k\right), m_i \leq k < l_i.$$

From $\|x_{k+1} - x_k\| = \|d_k\| \leq \chi_D \Delta_k$, we now have for sufficiently large $k \in [m_i, l_i)$

$$\begin{aligned} h\left(x_k\right) - h\left(x_{k+1}\right) &\geq \frac{1}{2}\beta_g \mu \bar{\epsilon}_2 \Delta_k \\ &\geq \frac{\beta_g \mu \bar{\epsilon}_2}{2\chi_D} \left\|x_{k+1} - x_k\right\| \qquad (4.35) \\ &= \epsilon_3 \left\|x_{k+1} - x_k\right\| \end{aligned}$$

where $\epsilon_3 := \frac{\beta_g \mu \bar{\epsilon}_2}{2\chi_D}$. By triangle inequality and (4.35), for sufficiently large i, we can easily get

$$\begin{aligned} h\left(x_{m_i}\right) - h\left(x_{k_i}\right) &= \sum_{j=m_i}^{k_i-1} \left(h\left(x_j\right) - h\left(x_{j+1}\right)\right) \qquad (4.36) \\ &\geq \sum_{j=m_i}^{k_i-1} \epsilon_3 \left\|x_{j+1} - x_j\right\| \\ &\geq \epsilon_3 \left\|\sum_{j=m_i}^{k_i-1} \left(x_{j+1} - x_j\right)\right\| \\ &= \epsilon_3 \left\|x_{k_i} - x_{m_i}\right\| \end{aligned}$$

for any $k_i \in [m_i, l_i]$. By the uniform continuity of $\nabla f(x)$ and the convergence of $\{h(x_k)\}$, which implies $\|x_{m_i} - x_{l_i}\| \to 0$ from (4.36), we have

$$\|\nabla f_{m_i} - \nabla f_{l_i}\| \leq \epsilon_2$$

for sufficiently large i. Note that $\hat{g}_k^0$ can be expanded as the following:

$$
\begin{aligned}
D_{m_i}^{\frac{1}{2}} g_{m_i} &= D_{m_i}^{\frac{1}{2}} \left(\nabla f_{m_i} + sign\left(x_{m_i}\right)\right) \\
&= \underbrace{D_{m_i}^{\frac{1}{2}} \left(\nabla f_{m_i} + sign\left(x_{m_i}\right)\right) - D_{m_i}^{\frac{1}{2}} \left(\nabla f_{l_i} + sign\left(x_{m_i}\right)\right)}_{(i)} \\
&\quad + \underbrace{D_{m_i}^{\frac{1}{2}} \left(\nabla f_{l_i} + sign\left(x_{m_i}\right)\right) - D_{l_i}^{\frac{1}{2}} \left(\nabla f_{l_i} + sign\left(x_{l_i}\right)\right)}_{(ii)} \\
&\quad + \underbrace{D_{l_i}^{\frac{1}{2}} \left(\nabla f_{l_i} + sign\left(x_{l_i}\right)\right)}_{(iii)} \qquad (4.37)
\end{aligned}
$$

Note that (i) in (4.37) can be arbitrarily small since $\|\nabla f_{m_i} - \nabla f_{l_i}\|$ can be arbitrarily small. (iii) in (4.37) is approaching zero by the construction of such $\{x_{l_i}\}$ subsequence. Rewrite (ii) in (4.37) as the following

$$
\begin{aligned}
& D_{m_i}^{\frac{1}{2}} \left(\nabla f_{l_i} + sign\left(x_{m_i}\right)\right) - D_{l_i}^{\frac{1}{2}} \left(\nabla f_{l_i} + sign\left(x_{l_i}\right)\right) \\
= & \left(D_{m_i}^{\frac{1}{2}} - D_{l_i}^{\frac{1}{2}}\right) \nabla f_{l_i} + \left(D_{m_i}^{\frac{1}{2}} sign\left(x_{m_i}\right) - D_{l_i}^{\frac{1}{2}} sign\left(x_{l_i}\right)\right)
\end{aligned}
$$

WLOG suppose that the full sequence$\{x_{l_i}\}$ converges to x_*, then $\{x_{m_i}\}$ must converge to the same x_* by $\|x_{m_i} - x_{l_i}\| \to 0$.

By the strict complementarity condition, we can claim that, at the limit point x_*, there is no $j \in \{1, \cdots, n\}$ such that $\left|\nabla f\left(x_*\right)_j\right| > 1$. This can be seen from discussion below.

If there exists $j \in \{1, \cdots, n\}$ such that $\left|\nabla f\left(x_*\right)_j\right| > 1$, then by definition of our scaling matrix, $\left(v\left(x_*\right)\right)_{j_{i^*}} = 1$ and $\left|\nabla f\left(x_*\right)_{j_{i^*}}\right| > 1$ (j^* and j_{i^*} could coincide). This will lead to $\left|\left(D_* g_*\right)_{j_{i^*}}\right| \neq 0$, which is a contradiction to $\|D_* g_*\| = 0$.

We study the following two cases:

Case I: if $x_{*j} \neq 0, j \in \{1, \cdots, n\}$

Based on the above claim, $\left|\nabla f\left(x_*\right)_j\right| \leq 1, \forall j \in \{1, \cdots, n\}$. For sufficiently large l_i,

$$\left(D_{m_i}^{\frac{1}{2}} - D_{l_i}^{\frac{1}{2}}\right)_{jj} = \left(\sqrt{\left|\left(x_{m_i}\right)_j\right|} - \sqrt{\left|\left(x_{l_i}\right)_j\right|}\right) \to 0$$

by $\|x_{m_i} - x_{l_i}\| \to 0$ and $\|\nabla f_{l_i}\|_\infty \leq \chi_f$ by assumption 3, we get $\left(\left(D_{m_i}^{\frac{1}{2}} - D_{l_i}^{\frac{1}{2}}\right)\nabla f_{l_i}\right)_j \to 0$. Moreover,

$$\left(D_{m_i}^{\frac{1}{2}} sign\left(x_{m_i}\right) - D_{l_i}^{\frac{1}{2}} sign\left(x_{l_i}\right)\right)_j = \left(\sqrt{\left|\left(x_{m_i}\right)_j\right|} - \sqrt{\left|\left(x_{l_i}\right)_j\right|}\right) sign\left(x_{l_i}\right)_j \to 0.$$

Case II: $x_{*j} = 0, j \in \{1, \cdots, n\}$

We get

$$\left(\left(D_{m_i}^{\frac{1}{2}} - D_{l_i}^{\frac{1}{2}}\right)\nabla f_{l_i}\right)_j \to 0$$

since

$$\left(D_{m_i}^{\frac{1}{2}}\right)_{jj} - \left(D_{l_i}^{\frac{1}{2}}\right)_{jj} = \left(\sqrt{\left|\left(x_{m_i}\right)_j\right|} - \sqrt{\left|\left(x_{l_i}\right)_j\right|}\right) \to 0$$

and

$$\left(D_{m_i}^{\frac{1}{2}} sign\left(x_{m_i}\right) - D_{l_i}^{\frac{1}{2}} sign\left(x_{l_i}\right)\right)_j \to 0$$

since

$$\left|\left(D_{m_i}^{\frac{1}{2}} sign\left(x_{m_i}\right) - D_{l_i}^{\frac{1}{2}} sign\left(x_{l_i}\right)\right)_j\right| \leq \sqrt{\left|\left(x_{m_i}\right)_j\right|} + \sqrt{\left|\left(x_{l_i}\right)_j\right|} \to 0.$$

Thus for sufficiently large l_i, we then establish the following

$$\begin{aligned}
\epsilon_1 \leq \|\hat{g}_{m_i}\| &\leq \left\|D_{m_i}^{\frac{1}{2}}\left(\nabla f_{m_i} + sign\left(x_{m_i}\right)\right) - D_{m_i}^{\frac{1}{2}}\left(\nabla f_{l_i} + sign\left(x_{m_i}\right)\right)\right\| \\
&\quad + \left\|D_{m_i}^{\frac{1}{2}}\left(\nabla f_{l_i} + sign\left(x_{m_i}\right)\right) - D_{l_i}^{\frac{1}{2}}\left(\nabla f_{l_i} + sign\left(x_{l_i}\right)\right)\right\| \\
&\quad \left\|D_{l_i}^{\frac{1}{2}}\left(\nabla f_{l_i} + sign\left(x_{l_i}\right)\right)\right\| \\
&\leq \chi_D\epsilon_2 + \epsilon_2 + \epsilon_2 \\
&= \left(\chi_D + 2\right)\epsilon_2
\end{aligned}$$

Since ϵ_2 can be chosen to be any number in $(0, \epsilon_1)$, we already arrive at a contradiction. Therefore, our base assumption $\limsup_{k\to\infty} \|\hat{g}_k^0\| \neq 0$ is not correct. □

4.2.2 Second-Order Necessary Optimality Conditions

In Section 4.2.1, we have proved that the first-order necessary condition $\lim_{k\to\infty} \|\hat{g}_k^0\| = 0$ is satisfied. Hereby we list assumption 5 and 6 to show the satisfiability for the 2-nd order necessary conditions. In this section, we suppress the stepsize α_k into d_k. That is, we want to solve the sub-problem as shown in assumption 6.

Assumption 5. Assume $\psi_k(d) := \left(g_k^0\right)^T d + \frac{1}{2} d^T \left(\nabla^2 f(x_k) + C_k\right) d$, i.e., $B_k = \nabla^2 f(x_k)$. We need B_k to be the exact Hessian matrix at x_k.

Assumption 6. Assume that p_k is a global solution to the problem

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

and β^q is a positive number, then d_k satisfies

$$\phi_k(d_k) \le \beta^q \phi_k^\star [d_k], \left\| D_k^{-\frac{1}{2}} d_k \right\| \le \Delta_k, x_k + d_k \in dif(\mathcal{F})$$

where

$$\phi_k^\star [d_k] := \min_\tau \left\{ \psi_k(\tau d_k) : \left\| \tau D_k^{-\frac{1}{2}} d \right\| \le \Delta_k, 0 \le \tau \le \beta_k^1 \right\}.$$

Lemma 4.7. *Let x_* be an isolated limit point of a sequence $\{x_k\}$ in $\mathbb{R}^n$. If $\{x_k\}$ does not converge, then there is a sub-sequence $\{x_{l_j}\}$, which converges to x_*, and an $\epsilon > 0$ such that*

$$\left\| x_{l_j+1} - x_{l_j} \right\| \ge \epsilon$$

Proof. Please refer to [23]. By the assumption of x_*, i.e., it is an isolated limit point of $\{x_k\}_{k=1}^\infty$, we can choose a ball $B(x^*, \bar{\epsilon})$, which is centered at x^* with radius $\bar{\epsilon} > 0$, such that, if $\|x - x_*\| \leq \bar{\epsilon}$ and x is a limit point, then $x = x_*$. For any sub-sequence $\left\{x_{k_j}\right\}$ which converges to x_*, if $\left\|x_{k_j} - x_*\right\| \leq \bar{\epsilon}$, then we can define l_j to be the maximum index from k_j such that $\left\|x_{l_j} - x_*\right\| \leq \bar{\epsilon}$, i.e.,

$$l_j := \max\left\{l : \|x_i - x_*\| \leq \bar{\epsilon}, i = k_j, \cdots l\right\}.$$

For each k_j, such l_j is well defined, i.e., $l_j < \infty$. Otherwise all the points following x_{k_j} will fall within $B(x^*, \bar{\epsilon})$. For the sequence $\{x_i\}_{i=k_j}^{\infty}$, because x^* is the only limit point in $B(x^*, \bar{\epsilon})$, we can deduce that$\{x_i\}_{i=k_j}^{\infty}$ converges, which contradicts the assumption in the lemma that $\{x_k\}$ does not converge. Thus the sub-sequence $\left\{x_{l_j}\right\}$ has the following property

$$\left\|x_{l_j} - x_*\right\| \leq \bar{\epsilon}, \left\|x_{l_j+1} - x_*\right\| > \bar{\epsilon}$$

Since $\left\{x_{l_j}\right\}$ is within $B(x^*, \bar{\epsilon})$ and there is only one limit point in this ball, $\left\{x_{l_j}\right\}$ must converge to x_*. Therefore for sufficiently large l_j, we have

$$\left\|x_{l_j} - x_*\right\| \leq \frac{1}{2}\bar{\epsilon}$$

We can see that

$$\left\|x_{l_j+1} - x_{l_j}\right\| \geq \left\|x_{l_j+1} - x_*\right\| - \left\|x_{l_j} - x_*\right\| \geq \bar{\epsilon} - \frac{1}{2}\bar{\epsilon} = \frac{1}{2}\bar{\epsilon}$$

so we can choose $\epsilon := \frac{1}{2}\bar{\epsilon}$ as required. □

Lemma 4.7 says if a sequence does not converge and one of its limit point is an isolated limit point, then we can always find a sub-sequence such that it converges to this isolated limit point. Moreover, for any point indexed by sufficiently large number in this sub-sequence, the distance between this point and the succeeding point in the original sequence $\{x_k\}$ is at least $\epsilon > 0$ away from each other.

Lemma 4.8. *Assume that assumption* 6 *holds. Then*

$$-\phi_k\left(d_k\right) \geq -\beta^q \phi_k^\star\left[d_k\right] \geq \frac{\beta^q}{2}\left[\min\left\{1,\left(\beta_k^1\right)^2\right\}\lambda_k\Delta_k^2 + \min\left\{1,\beta_k^1\right\}\left\|R_k D_k^{-\frac{1}{2}} p_k\right\|^2\right]$$

where β_k^1 *is along* p_k*, which is the global solution to* $\min_{d\in\mathbb{R}^n}\left\{\psi_k\left(d\right) : \left\|D_k^{-\frac{1}{2}} d\right\| \leq \Delta_k\right\}$.

Proof. By the definition of $\phi_k^\star\left[d_k\right]$, we get

$$\phi_k^\star\left[p_k\right] = \min_{\tau\in\mathbb{R}^n}\left\{\psi_k\left(\tau p_k\right) : \left\|\tau D_k^{-\frac{1}{2}} p_k\right\| \leq \Delta_k, \tau \leq \beta_k^1\right\}$$

where β_k^1 is the stepsize from the current iterate along p_k to the first break point. By definition of $\phi_k^\star\left[p_k\right]$. We can define $\Phi\left(\tau\right) := \left\{\psi_k\left(\tau p_k\right) : \tau \leq \beta_k^1\right\}$ and derive the following (use g_k to denote g_k^0 since we don't need to deal with break point crossing)

$$\begin{aligned}
\Phi\left(\tau\right) &= \phi_k\left(\tau p_k\right) \\
&= \psi_k\left(\tau p_k\right) \\
&= \tau g_k^T p_k + \frac{1}{2}\tau^2 p_k^T M_k p_k \\
&= \tau \hat{g}_k^T D_k^{-\frac{1}{2}} p_k + \frac{1}{2}\tau^2\left(D_k^{-\frac{1}{2}} p_k\right)^T \hat{M}_k D_k^{-\frac{1}{2}} p_k \\
&= -\tau\underbrace{\left(R_k^T R_k D_k^{-\frac{1}{2}} p_k\right)}_{From\,Lemma2}^T D_k^{-\frac{1}{2}} p_k - \frac{1}{2}\tau^2\left(D_k^{-\frac{1}{2}} p_k\right)^T \underbrace{\left(\hat{g}_k + \lambda_k D_k^{-\frac{1}{2}} p_k\right)}_{From\,Lemma2} \\
&= -\tau\left\|R_k D_k^{-\frac{1}{2}} p_k\right\|^2 + \underbrace{\frac{1}{2}\tau^2\left\|R_k D_k^{-\frac{1}{2}} p_k\right\|^2}_{From\,Lemma2} - \frac{1}{2}\tau^2\lambda_k\left\|D_k^{-\frac{1}{2}} p_k\right\|^2 \\
&= -\tau\left\|R_k D_k^{-\frac{1}{2}} p_k\right\|^2 + \frac{1}{2}\tau^2\left\|R_k D_k^{-\frac{1}{2}} p_k\right\|^2 - \frac{1}{2}\tau^2\underbrace{\lambda_k\Delta_k^2}_{From\,Lemma2} \\
&= \left(\frac{1}{2}\tau^2 - \tau\right)\left\|R_k D_k^{-\frac{1}{2}} p_k\right\|^2 - \frac{1}{2}\tau^2\lambda_k\Delta_k^2
\end{aligned}$$

Note that both $\left(\frac{1}{2}\tau^2 - \tau\right)$ and $-\frac{1}{2}\tau^2\lambda_k\Delta_k^2$ decrease as τ increases. Therefore

$$\tau^* = argmin\left\{\Phi\left(\tau\right) : \tau \in \left[0, \min\left\{1,\beta_k^1\right\}\right]\right\} = \min\left\{1,\beta_k^1\right\}$$

since $\Phi(\tau)$ is monotonic decreasing as τ increases in $[0, \min\{1, \beta_k^1\}]$.

Keep in mind that

$$\begin{aligned} \phi_k^*[p_k] &= \Phi(\tau^*) \\ &= \left(\frac{1}{2}(\tau^*)^2 - \tau^*\right)\left\|R_k D_k^{-\frac{1}{2}} p_k\right\|^2 - \frac{1}{2}(\tau^*)^2 \lambda_k \Delta_k^2 \\ &\leq -\frac{\tau^*}{2}\left\|R_k D_k^{-\frac{1}{2}} p_k\right\|^2 - \frac{1}{2}(\tau^*)^2 \lambda_k \Delta_k^2 \end{aligned} \tag{4.38}$$

Based on assumption 6 and (4.29) , we get

$$-\phi_k(d_k) \geq -\beta^q \phi_k^*[p_k] \geq \frac{\beta^q}{2}\left[\min\left\{1, \left(\beta_k^1\right)^2\right\}\lambda_k \Delta_k^2 + \min\left\{1, \beta_k^1\right\}\left\|R_k D_k^{-\frac{1}{2}} p_k\right\|^2\right] \tag{4.39}$$

□

Lemma 4.9 further extends the lower bound in Lemma 4.8 by checking into the lower bound of the first break point β_k^1.

Lemma 4.9. *Assume that the conditions in Theorem* 4.6 *hold and Assumption* 6 *is satisfied for d_k. Furthermore, $\{x_k\}$ is any sequence generated by the trust region Algorithm* 4.1*. If we assume that the objective function at every limit point of $\{x_k\}$ is differentiable, then there exists* $0 < \epsilon_0 < 1$ *such that, for sufficiently large k,*

$$-\phi_k(d_k) \geq \frac{\beta^q}{2}\min\left\{1, \frac{\lambda_k^2 \epsilon_0^2}{\left[\chi_D^2(\chi_g + \chi_B\chi_D\Delta_k)\right]^2}, \frac{\lambda_k^2}{(\chi_g + \chi_B\chi_D\Delta_k)^2}\right\} \cdot \lambda_k \Delta_k^2$$

where λ_k is defined in Lemma 4.2*.*

Moreover, if the k-th iteration is successful,

$$h(x_k) - h(x_{k+1}) > \frac{\beta^q}{2}\mu \min\left\{1, \frac{\lambda_k^2 \epsilon_0^2}{\left[\chi_D^2(\chi_g + \chi_B\chi_D\Delta_k)\right]^2}, \frac{\lambda_k^2}{(\chi_g + \chi_B\chi_D\Delta_k)^2}\right\} \cdot \lambda_k \Delta_k^2$$

Proof. Since the second term in (4.39) is also non-negative, we get

$$-\phi_k(d_k) \geq \frac{\beta^q}{2} \min\left\{1, \left(\beta_k^1\right)^2\right\} \lambda_k \Delta_k^2 \tag{4.40}$$

where β_k^1 is defined by

$$\beta_k^1 := \min_{i \in \{1, \cdots, n\}} \left\{ -\frac{x_{ki}}{p_{ki}} : -\frac{x_{ki}}{p_{ki}} \geq 0 \right\} \tag{4.41}$$

If we assume that the problem is differentiable at every limit point, then there must exist $0 < \epsilon_0 < 1$, for sufficiently large k, such that

$$|x_k| + |g_k| > 2\epsilon_0 e \tag{4.42}$$

where $e = (1, \cdots, 1)^T \in \mathbb{R}^n$. Otherwise, if such ϵ_0 does not exist, then for every $\epsilon \in (0,1)$, no matter how large k is, there exists $\tilde{k} \geq k$ such that $|x_{\tilde{k}}| + |g_{\tilde{k}}| > 2\epsilon e$ is violated, i.e., for some $j \in \{1, \cdots, n\}$, $\left|(x_{\tilde{k}})_j\right| + \left|(g_{\tilde{k}})_j\right| \leq 2\epsilon$. In another word, at the limit point x_*, for some component $\left|(x_*)_j\right| + \left|(g_*)_j\right| \leq 2\epsilon$. When $\left|(g_*)_j\right| = 0$, we can deduce that $\left|(x_*)_j\right| \leq 2\epsilon$ for arbitrary $\epsilon \in (0,1)$. This means the limit point is not differentiable, which is a contradiction to the assumption. On the other hand, if inequality (4.42) is satisfied, whenever $\left|(g_*)_j\right| = 0$, we have

$$\left|(x_*)_j\right| > 2\epsilon_0 > 0$$

which guarantees that the limit point is away from the corresponding break point.

From Theorem 4.6, in which it says $\lim_{k\to\infty} \left\|D_k^{\frac{1}{2}} g_k\right\| = 0$, and $\left\|D_k^{\frac{1}{2}}\right\| \leq \chi_D$, we can deduce that, for sufficiently large k,

$$\|D_k g_k\|_\infty < \epsilon_0^2. \tag{4.43}$$

Assume that

$$\beta_k^1 = \min_{i \in \{1, \cdots, n\}} \left\{ -\frac{x_{ki}}{p_{ki}} : -\frac{x_{ki}}{p_{ki}} \geq 0 \right\} = -\frac{x_{kj}}{p_{kj}} \tag{4.44}$$

for some $j \in \{1, \cdots, n\}$.

We study two cases:

Case I: If $|g_{kj}| \leq \epsilon_0$, by (4.42), we have $|x_{kj}| > \epsilon_0$. Therefore we get

$$\beta_k^1 = \frac{|x_{kj}|}{|p_{kj}|} \geq \frac{\epsilon_0}{|p_{kj}|}$$

Note that

$$\begin{aligned} \|g_k + B_k p_k\|_\infty &\leq \|g_k\|_\infty + \|B_k p_k\|_\infty \\ &\leq \chi_g + \|B_k p_k\|_2 \\ &\leq \chi_g + \chi_B \chi_D \Delta_k \end{aligned} \tag{4.45}$$

Keep in mind that $D_k^{\frac{1}{2}} C_k D_k^{\frac{1}{2}} = diag\left(g_k\right) J_k^v \succeq 0$ where J_k^v is a diagonal Jacobian matrix of $|v\left(x_k\right)|$. That is because

$$\left(diag\left(g_k\right) J_k^v\right)_{ii} = \begin{cases} 0 & \left|\left(\nabla f\left(x_k\right)\right)_i\right| > 1 \\ \left(\left(\nabla f\left(x_k\right)\right)_i + sign\left(x_{ki}\right)\right) \cdot sign\left(x_{ki}\right) \geq 0 & otherwise \end{cases}$$

Based on (4.45) and

$$\left(\lambda_k I + D_k^{\frac{1}{2}} C_k D_k^{\frac{1}{2}}\right) p_k = -D_k\left(g_k + B_k p_k\right)$$

in Lemma 4.2, we get

$$\begin{aligned} \left\|\left(\lambda_k I + diag\left(g_k\right) J_k^v\right) p_k\right\|_\infty &= \left\|-D_k\left(g_k + B_k p_k\right)\right\|_\infty \\ &\leq \chi_D^2\left(\chi_g + \chi_B \chi_D \Delta_k\right) \end{aligned} \tag{4.46}$$

Note that

$$\left\|\left(\lambda_k I + diag\left(g_k\right) J_k^v\right) p_k\right\|_\infty \geq \left\|\lambda_k p_k\right\|_\infty \geq \lambda_k\left|p_{kj}\right| \tag{4.47}$$

since $I \succeq 0$ and $diag\left(g_k\right) J_k^v \succeq 0$.

Based on (4.46) and (4.47), we can easily get

$$\left|p_{kj}\right| \leq \frac{\chi_D^2\left(\chi_g + \chi_B \chi_D \Delta_k\right)}{\lambda_k}$$

thus

$$\beta_k^1 \geq \frac{\lambda_k \epsilon_0}{\chi_D^2\left(\chi_g + \chi_B \chi_D \Delta_k\right)}$$

Case II: If $\left|g_{kj}\right| > \epsilon_0$, then by $\left\|D_k g_k\right\|_\infty < \epsilon_0^2$ for sufficiently large k, we have

$$\left|v_{kj} \cdot g_{kj}\right| \leq \left\|D_k g_k\right\|_\infty < \epsilon_0^2$$

thus $\left|v_{kj}\right| < \epsilon_0$. By definition of v_{kj}, the value of it is either 1 or $\left|x_{kj}\right|$, hence it must be the case that

$$\left|x_{kj}\right| = \left|v_{kj}\right| < \epsilon_0 < 1$$

Then we get

$$\left(\lambda_k + \left(D_k^{\frac{1}{2}} C_k D_k^{\frac{1}{2}}\right)_{jj}\right) p_{kj} = -\left|x_{kj}\right| \left(g_{kj} + \left(B_k p_k\right)_j\right)$$

from the j-th component of the following equation

$$\left(\lambda_k I + D_k^{\frac{1}{2}} C_k D_k^{\frac{1}{2}}\right) p_k = -D_k\left(g_k + B_k p_k\right)$$

Thus we have

$$\begin{aligned} \left|p_{kj}\right| &= \frac{\left|x_{kj}\right| \cdot \left|g_{kj} + \left(B_k p_k\right)_j\right|}{\left|\lambda_k + \left(D_k^{\frac{1}{2}} C_k D_k^{\frac{1}{2}}\right)_{jj}\right|} \\ &\leq \frac{\left|x_{kj}\right| \cdot \left(\chi_g + \chi_B \chi_D \Delta_k\right)}{\lambda_k} \end{aligned}$$

Therefore we can conclude that

$$\begin{aligned}\beta_k^1 &= \frac{|x_{kj}|}{|p_{kj}|}\\ &\geq \frac{\lambda_k |x_{kj}|}{|x_{kj}| \cdot (\chi_g + \chi_B \chi_D \Delta_k)}\\ &= \frac{\lambda_k}{\chi_g + \chi_B \chi_D \Delta_k}\end{aligned}$$

and

$$-\phi_k(d_k) \geq \frac{\beta^q}{2} \min\left\{1, \frac{\lambda_k^2 \epsilon_0^2}{\left[\chi_D^2 (\chi_g + \chi_B \chi_D \Delta_k)\right]^2}, \frac{\lambda_k^2}{(\chi_g + \chi_B \chi_D \Delta_k)^2}\right\} \cdot \lambda_k \Delta_k^2 \tag{4.48}$$

□

If $\hat{M}_k$ is positive definite, we can define the Newton step for

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

by

$$d_k^N := -D_k^{\frac{1}{2}} \hat{M}_k^{-1} \hat{g}_k, \; i.e, \hat{M}_k D_k^{-\frac{1}{2}} d_k^N = -\hat{g}_k \tag{4.49}$$

Lemma 4.10. *Assume assumption* $1, 4, 5, 6$ *hold and* $f(x)$ *is twice continuously differentiable on* $\mathcal{F}$. *If the sequence of trust region sub-problem* $\min_{d \in \mathbb{R}^n} \left\{\psi_k(d) : \left\|D_k^{-\frac{1}{2}} d\right\| \leq \Delta_k\right\}$ *solution* $\{p_k\}$ *converges to zero,* $\{x_k\}$ *converges to* x_* *and* $\hat{M}_*$ *is positive definite, then*

$$\liminf_{k \to \infty} \frac{\phi_k(\beta_k^*[p_k])}{\phi_k^\star[d_k]} \geq 1, \liminf_{k \to \infty} \frac{\phi_k^\star[d_k]}{\phi_k(p_k)} \geq 1$$

where $\beta_k^*[p_k] := \theta_k \tau_k^* p_k$ *is the step obtained from* p_k *with a possible step-back to maintain strict differentiability.*

Moreover, for sufficiently large k,

$$\left|\phi_k^* \left[p_k\right]\right| \geq \epsilon \min \left\{ \Delta_k^2, \left\| D_k^{-\frac{1}{2}} d_k^N \right\|^2 \right\}$$

for some constant $\epsilon > 0$.

Proof. Let $\beta_k^1 := \min_{i \in \{1,\cdots,n\}} \left\{ -\frac{x_{ki}}{p_{ki}} : -\frac{x_{ki}}{p_{ki}} \geq 0 \right\}$ be as before. By Lemma 4.7, $\tau_k^* = \min\left\{1, \beta_k^1\right\}$ is a minimizer of $\Phi\left(\tau\right)$. Back-tracking is performed when necessary with $\theta_k \in [\theta_l, 1)$ and $\theta_k - 1 = \mathcal{O}\left(\|p_k\|\right)$.

We first prove $\liminf_{k\to\infty} \beta_k^1 \geq 1$.

Since $\hat{M}_*$ is positive definite and $f\left(x\right) \in \mathbb{C}^2$, for sufficiently large k, $\hat{M}_k$ is also positive definite.

Case I: All components of x_* are non-zeros, then

$$\liminf_{k\to\infty} \beta_k^1 = \liminf_{k\to\infty} \min_{i \in \{1,\cdots,n\}} \left\{ -\frac{x_{ki}}{p_{ki}} : -\frac{x_{ki}}{p_{ki}} \geq 0 \right\} \to +\infty$$

This is reasonable since at the differentiable point x_*, global solution $p_* = \left(0, \cdots, 0\right)^T$ by the assumption $\|p_k\| \to 0$.

Case II: There exists at lease one component of x_* which is zero. Based on$\left(\lambda_k I + D_k C_k\right) p_k = -D_k \left(g_k + B_k p_k\right)$ in Lemma 4.2 ($D_k^{\frac{1}{2}} C_k D_k^{\frac{1}{2}} = D_k C_k$ since diagonal), we can show that $\liminf_{k\to\infty} \beta_k^1 \geq 1$. The reasons are as follows: since $\left\{B_k p_k\right\}$ converges to zero, $\lambda_k I + D_k C_k$ is a positive semi-definite diagonal matrix, we can see that, for any j with $v_{*j} = 0$ and for sufficiently large k, p_{kj} and $-g_{kj}$ must have the same sign. Therefore if β_k^1 is defined by some v_{*j} with $g_{*j} \neq 0$, i.e., $\beta_k^1 = \frac{\left|x_{kj}\right|}{\left|p_{kj}\right|} = \frac{\left|v_{kj}\right|}{\left|p_{kj}\right|}$ (Because $D_*^{\frac{1}{2}} g_* = 0$ and $g_{*j} \neq 0$, we get $\lim_{k\to\infty} v_{kj} \to 0$. Hence for sufficiently large k, $v_{kj} = \left|x_{kj}\right|$) , we can deduce that

$$\beta_k^1 = \frac{\left|v_{kj}\right|}{\left|p_{kj}\right|} = \frac{\lambda_k + \left(D_k C_k\right)_{jj}}{\left|g_{kj} + \left(B_k p_k\right)_j\right|} = \frac{\lambda_k + \left(diag\left(g_k\right) J_k^v\right)_{jj}}{\left|g_{kj} + \left(B_k p_k\right)_j\right|}$$

Note that $J_k^v\left(x_k\right) \in \mathbb{R}^{n\times n}$ is a diagonal matrix, which corresponds to the Jacobian matrix of

$|v(x_k)|$. We already exclude the case $v_{kj} = 1$ for sufficiently large k. Therefore

$$J_k^v(x_k)_{jj} = sign(x_{kj})$$

By $\beta_k^1 = -\frac{x_{kj}}{p_{kj}}$, we get $sign(x_{kj}) = -sign(p_{kj}) = sign(g_{kj})$. Therefore for sufficiently large k, β_k^1 can be written as

$$\beta_k^1 = \frac{\lambda_k + g_{kj} sign(g_{kj})}{\left| g_{kj} + (B_k p_k)_j \right|} = \frac{\lambda_k + |g_{kj}|}{\left| g_{kj} + (B_k p_k)_j \right|} \tag{4.50}$$

Since $\lambda_k \geq 0$ and $\{B_k p_k\} \to 0$, we can conclude that $\liminf_{k\to\infty} \beta_k^1 \geq 1$.

Using the second term in (4.39) in Lemma 4.8 , we see that

$$-\phi_k^*[p_k] \geq \frac{1}{2}\min\left\{1, \beta_k^1\right\} \left\| R_k D_k^{-\frac{1}{2}} p_k \right\|^2$$

Since $\hat{M}_*$ is a positive definite matrix, a lower bound ϵ' on all eigenvalues of $\hat{M}_*$ must be positive.

From the conclusions in Lemma 4.2, $\lambda_k I + \hat{M}_k = R_k^T R_k$, $\left(\lambda_k I + \hat{M}_k\right) D_k^{-\frac{1}{2}} p_k = -\hat{g}_k$ and $\lambda_k \geq 0$, we can derive the following

$$\begin{aligned}
\left\| R_k D_k^{-\frac{1}{2}} p_k \right\|^2 &= \left(R_k D_k^{-\frac{1}{2}} p_k\right)^T \left(R_k D_k^{-\frac{1}{2}} p_k\right) \\
&= \left(D_k^{-\frac{1}{2}} p_k\right)^T R_k^T R_k \left(D_k^{-\frac{1}{2}} p_k\right) \\
&= \left(D_k^{-\frac{1}{2}} p_k\right)^T \left(\hat{M}_k + \lambda_k I\right) \left(D_k^{-\frac{1}{2}} p_k\right) \\
&\geq \epsilon' \left\| D_k^{-\frac{1}{2}} p_k \right\|_2^2 + \lambda_k \left\| D_k^{-\frac{1}{2}} p_k \right\|_2^2
\end{aligned}$$

Notice that $\left\| D_k^{-\frac{1}{2}} p_k \right\| \leq \Delta_k$ and for sufficiently large k, $p_k = d_k^N$ if $\left\| D_k^{-\frac{1}{2}} p_k \right\| < \Delta_k$. Based on these observations, we can derive that

$$\begin{aligned}
\left|\phi_k^*\left[p_k\right]\right| &\geq \frac{1}{2}\min\left\{1,\beta_k^1\right\}\left\|R_k D_k^{-\frac{1}{2}} p_k\right\|^2 \\
&\geq \frac{1}{2}\min\left\{1,\beta_k^1\right\}\epsilon'\left\|D_k^{-\frac{1}{2}} p_k\right\|^2 \\
&= \begin{cases} \frac{1}{2}\epsilon'\min\left\{1,\beta_k^1\right\}\Delta_k^2 & \left\|D_k^{-\frac{1}{2}} p_k\right\| = \Delta_k \\ \frac{1}{2}\epsilon'\min\left\{1,\beta_k^1\right\}\left\|D_k^{-\frac{1}{2}} d_k^N\right\|^2 & \left\|D_k^{-\frac{1}{2}} p_k\right\| < \Delta_k \end{cases} \\
&\geq \frac{1}{2}\epsilon'\min\left\{1,\beta_k^1\right\}\min\left\{\Delta_k^2,\left\|D_k^{-\frac{1}{2}} d_k^N\right\|^2\right\} \\
&\geq \epsilon\min\left\{1,\beta_k^1\right\}\min\left\{\Delta_k^2,\left\|D_k^{-\frac{1}{2}} d_k^N\right\|^2\right\}
\end{aligned}$$

where $\epsilon \in \left(0,\frac{1}{2}\epsilon'\right)$. In addition, when k is sufficiently large , based on $\liminf_{k\to\infty}\beta_k^1 \geq 1$ as proved earlier, we can judge that

$$\left|\phi_k^*\left[p_k\right]\right| \geq \epsilon\min\left\{\Delta_k^2,\left\|D_k^{-\frac{1}{2}} d_k^N\right\|^2\right\} \tag{4.51}$$

Now we can conclude that, based on the fact that $\left(\tau_k^*\right)^2 \leq \tau_k^* = \min\left\{1,\beta_k^1\right\} \leq 1$, $p_k^T M_k p_k > 0$ for sufficiently large k, $\psi_k^*\left[p_k\right] = \Phi\left(\tau^*\right)$ and $\phi_k\left(p_k\right) < 0$, we get

$$\begin{aligned}
\liminf_{k\to\infty}\frac{\phi_k^*\left[p_k\right]}{\phi_k\left(p_k\right)} &= \liminf_{k\to\infty}\frac{\Phi\left(\tau^*\right)}{\phi_k\left(p_k\right)} \\
&= \liminf_{k\to\infty}\frac{\tau_k^* g_k^T p_k + \frac{1}{2}\left(\tau_k^*\right)^2 p_k^T M_k p_k}{g_k^T p_k + \frac{1}{2} p_k^T M_k p_k} \\
&\geq \liminf_{k\to\infty}\tau_k^* \\
&= \liminf_{k\to\infty}\min\left\{1,\beta_k^1\right\} \\
&= 1
\end{aligned}$$

Hence

$$\liminf_{k\to\infty}\frac{\phi_k^*\left[p_k\right]}{\phi_k\left(p_k\right)} \geq 1$$

Moreover, we get

$$\begin{aligned}
\liminf_{k\to\infty} \frac{\phi_k\left(\beta_k^*\left[p_k\right]\right)}{\phi_k^*\left[p_k\right]} &= \liminf_{k\to\infty} \frac{\tau_k^*\theta_k g_k^T p_k + \frac{1}{2}\left(\tau_k^*\right)^2 \theta_k^2 p_k^T M_k p_k}{\tau_k^* g_k^T p_k + \frac{1}{2}\left(\tau_k^*\right)^2 p_k^T M_k p_k} \\
&\geq \liminf_{k\to\infty} \frac{\tau_k^*\theta_k g_k^T p_k + \frac{1}{2}\left(\tau_k^*\right)^2 \theta_k p_k^T M_k p_k}{\tau_k^* g_k^T p_k + \frac{1}{2}\left(\tau_k^*\right)^2 p_k^T M_k p_k} \\
&= \liminf_{k\to\infty} \theta_k \\
&= \lim_{k\to\infty} \theta_k \\
&= 1
\end{aligned}$$

The first equality follows from that $\theta_k^2 \leq \theta_k < 1$ $(\theta_k \to 1)$, both the numerator and the denominator have the negative sign in value and $p_k^T M_k p_k > 0$ from

$$p_k^T M_k p_k = \left(D_k^{-\frac{1}{2}} p_k\right)^T \hat{M}_k \left(D_k^{-\frac{1}{2}} p_k\right) \geq 0$$

The last equality comes from $\lim_{k\to\infty}\left(\theta_k - 1\right) = \lim_{k\to\infty} O\left(\|p_k\|\right) \to 0$ because $\|p_k\| \to 0$ by the assumption in the lemma. □

The last theorem will establish the fact that the first and second-order necessary conditions are satisfied achieved by Algorithm 4.1.

Theorem 4.11. *Assume assumption* 1 *holds and* $f : \mathcal{F} \mapsto \mathbb{R}$ *is twice continuously differentiable on* $\mathcal{F}$*. Let* $\{x_k\}$ *be the sequence generated by Algorithm* 4.1 *under assumption* 5 *on the model* ψ_k*, and under assumption* 4 *and* 6 *on the step* d_k*. Assume strict complementarity holds, then*

(i) The sequence $\{\hat{g}_k\}$ *converges to zero.*

(ii) There is a limit point x_* *with* $\hat{M}_*$ *positive semi-definite.*

(iii) If x_* *is an isolated limit point satisfies strict complementarity, then* $\hat{M}_*$ *is positive semi-definite.*

(iv) If $\hat{M}_*$ *is non-singular for some limit point* x_* *of* $\{x_k\}$*, then* $\hat{M}_*$ *is positive definite,*

$\{x_k\}$ converges to x_*, all iterations are eventually successful, and $\{\Delta_k\}$ is bounded away from zero.

Proof. (i) This claim has already been proved in Theorem 4.6.

(ii) We first consider the case when $\liminf_{k\to\infty} \lambda_k = 0$. Let $\lambda_{\min}\left(\hat{M}_k\right)$ denote the minimum eigenvalue of $\hat{M}_k$. Since $\lambda_k I + \hat{M}_k \succeq 0$, $\lambda_k \geq \max\left\{-\lambda_{\min}\left(\hat{M}_k\right), 0\right\}$ must be true. If $\liminf_{k\to\infty} \lambda_k = 0$, for sequence $\{x_k\}$, there must exist a sub-sequence $\{x_{k_i}\}$ such that $\{x_{k_i}\} \to x_*$ with $\{\lambda_{k_i}\} \to 0$. From $0 \leftarrow \lambda_{k_i} \geq \max\left\{-\lambda_{\min}\left(\hat{M}_{k_i}\right), 0\right\}$, we have $\lambda_{\min}\left(\hat{M}_{k_i}\right) \geq 0$ for sufficiently large k_i. Hence $\lambda_{\min}\left(\hat{M}_*\right) \geq 0$.

We will prove by contradiction that $\liminf_{k\to\infty} \lambda_k$ can only take value 0. Suppose not, assume that there exists an $\epsilon > 0$ such that for sufficiently large k, $\lambda_k \geq \epsilon > 0$. Using (4.48) in Lemma 4.8, for sufficiently large k,

$$-\phi_k\left(d_k\right) \geq \frac{\beta^q}{2} \min\left\{1, \frac{\lambda_k^2 \epsilon_0^2}{\left[\chi_D^2\left(\chi_g + \chi_B\chi_D\Delta_k\right)\right]^2}, \frac{\lambda_k^2}{\left(\chi_g + \chi_B\chi_D\Delta_k\right)^2}\right\} \cdot \lambda_k \Delta_k^2$$

We get

$$-\phi_k\left(d_k\right) \geq \frac{\beta^q}{2} \min\left\{1, \frac{\epsilon^2 \epsilon_0^2}{\left[\chi_D^2\left(\chi_g + \chi_B\chi_D\Delta_k\right)\right]^2}, \frac{\epsilon^2}{\left(\chi_g + \chi_B\chi_D\Delta_k\right)^2}\right\} \epsilon \Delta_k^2$$

Let $\hat{\epsilon}_k := \min\left\{1, \frac{\epsilon^2 \epsilon_0^2}{\left[\chi_D^2\left(\chi_g + \chi_B\chi_D\Delta_k\right)\right]^2}, \frac{\epsilon^2}{\left(\chi_g + \chi_B\chi_D\Delta_k\right)^2}\right\}$, for sufficiently large k, if iteration is successful, we get

$$h\left(x_k\right) - h\left(x_{k+1}\right) > \frac{\beta^q}{2} \mu \hat{\epsilon}_k \epsilon \Delta_k^2 \tag{4.52}$$

Case I: If there are only finite many successful iterations, $\{\Delta_k\} \to 0$ by the trust region update rule.

Case II: Otherwise let $\{k_i\}$ be the indices for the infinite sequence of successful iterations. Based on the definition of $\hat{\epsilon}_k$, (4.52) acting on $\{k_i\}$, and $\{h\left(x_k\right) - h\left(x_{k+1}\right)\}$converges to zero, we can conclude that there must exist some $\epsilon_1 > 0$ such that $\hat{\epsilon}_{k_i} > \epsilon_1$. This will indicate that

$$\sum_{i=1}^{\infty} \Delta_{k_i}^2 < \infty$$

It is easy to verify

$$\sum_{k=1}^{\infty} \Delta_k^2 \leq \left(1 + \frac{\gamma_2^2}{1-\gamma_1}\right) \sum_{i=1}^{\infty} \Delta_{k_i}^2 < \infty$$

hence we get $\Delta_k \to 0$. Thus,

$$\|d_k\| = \left\| D_k^{\frac{1}{2}} D_k^{-\frac{1}{2}} d_k \right\| \leq \chi_D \Delta_k \to 0$$

and

$$\|p_k\| = \left\| D_k^{\frac{1}{2}} D_k^{-\frac{1}{2}} p_k \right\| \leq \chi_D \Delta_k \to 0$$

Since $\{\Delta_k\}$ converges to zero, there exists some $\hat{\epsilon} > 0$ such that $\hat{\epsilon}_k \geq \hat{\epsilon}$. Therefore

$$-\phi_k(d_k) \geq \frac{\beta^q}{2} \hat{\epsilon} \epsilon \Delta_k^2$$

Now we examine the value of the agreement factor ρ_k^f

$$\begin{aligned} & \left| h(x_k + d_k) - h(x_k) + \frac{1}{2} d_k^T C_k d_k - \phi_k(d_k) \right| \\ = & \left| f(x_k + d_k) - f(x_k) - \nabla f(x_k)^T d_k - \frac{1}{2} d_k^T B_k d_k \right| \end{aligned}$$

Since $\{\Delta_k\} \to 0$ thus $\|d_k\| = \left\| D_k^{\frac{1}{2}} D_k^{-\frac{1}{2}} d_k \right\| \leq \chi_D \Delta_k \to 0$, we have

$$\begin{aligned} & \left| f(x_k + d_k) - f(x_k) - \nabla f(x_k)^T d_k - \frac{1}{2} d_k^T B_k d_k \right| \\ \leq & \|d_k\|^2 \cdot \max_{\hat{\xi} \in [0,1]} \left\| \nabla^2 f\left(x_k + \hat{\xi} d_k\right) - \nabla^2 f(x_k) \right\| \end{aligned}$$

Since $\{d_k\} \to 0$ and $f \in \mathbb{C}^2$, then $\max_{\hat{\xi} \in [0,1]} \left\| \nabla^2 f\left(x_k + \hat{\xi} d_k\right) - \nabla^2 f(x_k) \right\| \to 0$. Therefore

$\left\{\left|\rho_k^f - 1\right|\right\} \to 0$. By Lemma 4.3, $\{\Delta_k\}$ must be bounded away from zero if for sufficiently large k, the iteration is always successful. We then arrive at a contradiction with $\{\Delta_k\} \to 0$. In conclusion, there exists a limit point x_* with the corresponding $\hat{M}_*$ being positive semi-definite and $\liminf_{k\to\infty} \lambda_k = 0$.

(iii) If $\{x_k\}$ converges, then it must converge to x_*. From (ii), $\hat{M}_*$ is positive semi-definite. If $\{x_k\}$ does not converge, we can take a sub-sequence $\left\{x_{l_j}\right\}$ which is generated as described in Lemma 4.6, then $\left\{x_{l_j}\right\}$ converges to x_* and $\left\|x_{l_j+1} - x_{l_j}\right\| \geq \epsilon$ for sufficiently large l_j. Note that $\left\|x_{l_j+1} - x_{l_j}\right\| = \left\|d_{l_j}\right\| \leq \chi_D \Delta_{l_j}$, hence, for sufficiently large l_j, we have $\Delta_{l_j} \geq \frac{\epsilon}{\chi_D} > 0$. By Lemma 4.9, it can be verified that (l_j-th iteration is successful)

$$h\left(x_{l_j}\right) - h\left(x_{l_j+1}\right) > \frac{\beta^q}{2}\mu \min\left\{1, \frac{\lambda_{l_j}^2 \epsilon_0^2}{\left[\chi_D^2\left(\chi_g + \chi_B \chi_D \Delta_{l_j}\right)\right]^2}, \frac{\lambda_{l_j}^2}{\left(\chi_g + \chi_B \chi_D \Delta_{l_j}\right)^2}\right\} \cdot \lambda_{l_j} \Delta_{l_j}^2$$

Since $\Delta_{l_j} \geq \frac{1}{\chi_D}\epsilon$, there exist positive constants $\epsilon_1, \epsilon_2, \epsilon_3$ such that

$$h\left(x_{l_j}\right) - h\left(x_{l_j+1}\right) > \frac{\beta^q}{2}\mu \min\left\{\epsilon_1 \lambda_{l_j}, \epsilon_2 \lambda_{l_j}^3, \epsilon_3^3 \lambda_{l_j}^3\right\}$$

Since$\{h\left(x_k\right)\}$ is non-increasing and bounded below, $\left\{h\left(x_{l_j}\right) - h\left(x_{l_j+1}\right)\right\} \to 0$. Then it must be the case that λ_{l_j} converges to zero. Thus we get

$$\lim_{j\to\infty} \hat{M}_{l_j} + \lambda_{l_j} I = \hat{M}_* \succeq 0$$

(iv) If $\hat{M}_*$ is non-singular at a limit point x_* of $\{x_k\}$, then x_* is an isolated limit point and $\hat{M}_*$ is positive definite following from (iii). Since

$$\phi_k\left(d_k\right) = g_k^T d_k + \frac{1}{2} d_k^T M_k d_k < 0$$

which is equivalent to

$$\hat{g}_k^T D_k^{-\frac{1}{2}} d_k < -\frac{1}{2}\left(D_k^{-\frac{1}{2}} d_k\right)^T \hat{M}_k \left(D_k^{-\frac{1}{2}} d_k\right) < 0$$

whenever $\hat{M}_k \succ 0$ and $D_k^{-\frac{1}{2}} d_k \neq \mathbf{0}$. Therefore we can derive the following

$$\frac{1}{2}\left(D_k^{-\frac{1}{2}} d_k\right)^T \hat{M}_k \left(D_k^{-\frac{1}{2}} d_k\right) < -\hat{g}_k^T D_k^{-\frac{1}{2}} d_k \leq \left\|D_k^{-\frac{1}{2}} d_k\right\| \left\|\hat{g}_k\right\|$$

whenever $\hat{M}_k \succ 0$. ($|\langle A, B\rangle| \leq \|A\| \cdot \|B\|$ Cauchy-Schwartz Inequality)

But it is easy to check $\left(D_k^{-\frac{1}{2}} d_k\right)^T \hat{M}_k D_k^{-\frac{1}{2}} d_k \geq \frac{1}{\|\hat{M}_k^{-1}\|} \left\|D_k^{-\frac{1}{2}} d_k\right\|^2$. This means that

$$\frac{1}{2}\left\|D_k^{-\frac{1}{2}} d_k\right\| \leq \left\|\hat{M}_k^{-1}\right\| \left\|\hat{g}_k\right\|$$

whenever $\hat{M}_k$ is positive definite. We can derive the following

$$\frac{1}{2}\left\|d_k\right\| \leq \frac{1}{2}\chi_D \left\|D_k^{-\frac{1}{2}} d_k\right\| \leq \chi_D \left\|\hat{M}_k^{-1}\right\| \left\|\hat{g}_k\right\|$$

whenever $\hat{M}_k \succ 0$. By Theorem 4.6, $\lim_{k\to\infty} \|\hat{g}_k\| = 0$. We can deduce that $\{x_k\}$ converges based on Lemma 4.7. Otherwise, if $\{x_k\}$ does not converge, there exists a sub-sequence $\left\{x_{l_j}\right\}$, which converges to this isolated limit point x_* with property $\left\|x_{l_j+1} - x_{l_j}\right\| \geq \epsilon$ and $\hat{M}_* \succ 0$. This will lead to a contradiction as follows

$$0 < \frac{1}{2}\epsilon \leq \frac{1}{2}\left\|d_{l_j}\right\| \leq \frac{1}{2}\chi_D \left\|D_{l_j}^{-\frac{1}{2}} d_{l_j}\right\| \leq \chi_D \left\|\hat{M}_{l_j}^{-1}\right\| \left\|\hat{g}_{l_j}\right\| \to 0$$

for sufficiently large l_j. In conclusion, $\{x_k\}$ converges to x_* and $\hat{M}_*$ is positive definite, $\{p_k\}$ and $\{d_k\}$ both converge to zero.

Using (4.51) in Lemma 4.10, there exists $\epsilon > 0$ such that, for sufficiently large k,

$$\left|\phi_k^* \left[p_k\right]\right| \geq \epsilon \min\left\{\Delta_k^2, \left\|D_k^{-\frac{1}{2}} d_k^N\right\|^2\right\}$$

Recall that we have just established the fact that whenever $\hat{M}_k \succ 0$, we have

$$\frac{1}{2}\left\|D_k^{-\frac{1}{2}}d_k\right\| \leq \left\|\hat{M}_k^{-1}\right\| \|\hat{g}_k\|$$

Let κ be an upper bound on the condition number of $\hat{M}_k$. From $\hat{g}_k = -\hat{M}_k D_k^{-\frac{1}{2}} d_k^N$, we get

$$\frac{1}{2}\left\|D_k^{-\frac{1}{2}}d_k\right\| \leq \kappa \left\|D_k^{-\frac{1}{2}}d_k^N\right\|$$

Hence we can use $|\phi_k^*[p_k]| \geq \epsilon \min\left\{\Delta_k^2, \left\|D_k^{-\frac{1}{2}}d_k^N\right\|^2\right\}$ and Assumption 6, there exists $\bar{\epsilon} > 0$ such that for sufficiently large k

$$\begin{aligned}
-\phi_k(d_k) &\geq \beta^q |\phi_k^*[p_k]| \\
&\geq \beta^q \epsilon \min\left\{\Delta_k^2, \left\|D_k^{-\frac{1}{2}}d_k^N\right\|^2\right\} \\
&= \beta^q \epsilon \left\|D_k^{-\frac{1}{2}}d_k^N\right\|^2 \\
&\geq \frac{\beta^q \epsilon}{4\kappa^2}\left\|D_k^{-\frac{1}{2}}d_k\right\|^2 \\
&= \bar{\epsilon}\left\|D_k^{-\frac{1}{2}}d_k\right\|^2
\end{aligned}$$

where $\bar{\epsilon} := \frac{\beta^q \epsilon}{4\kappa^2} > 0$.

Hence we can get

$$-\phi_k(d_k) \geq \bar{\epsilon}\left\|D_k^{-\frac{1}{2}}d_k\right\|^2 \geq \frac{\bar{\epsilon}}{\chi_D^2}\|d_k\|^2$$

Based on above we can derive

$$\left|h(x_k + d_k) - h(x_k) + \frac{1}{2}d_k^T C_k d_k - \phi_k(d_k)\right| \leq \frac{1}{2}\|d_k\|^2 \cdot \max_{\hat{\xi}\in[0,1]}\left\|\nabla^2 f\left(x_k + \hat{\xi}d_k\right) - \nabla^2 f(x_k)\right\| \to 0$$

since $\|d_k\| \to 0$ thus $\max_{\hat{\xi}\in[0,1]}\left\|\nabla^2 f\left(x_k + \hat{\xi}d_k\right) - \nabla^2 f(x_k)\right\| \to 0$. This implies that for sufficiently large k, we always get

$$\rho_k^f \geq \eta$$

then $\{\Delta_k\}$ is bounded away from zero based on Lemma 4.3. □

Chapter 5

Concluding Remarks

In this thesis, we have proposed an affine scaling gradient based algorithm with BB stepsize for minimizing nonlinear function regularized by $\mathcal{L}_1$-norm. The novelty of this algorithm comes from taking a new scaled descent direction with safeguarded Barzilai-Borwein stepsizes. In order to ensure convergence, an adaptive line search is performed on each iteration. This algorithm does not require exact line search and the overall computational cost is cheap. We investigated the computational performance of our proposed scaled steepest descent method and impact of parameter choice. We also compared our method against the spectral projected gradient method in illustrating the effectiveness of our algorithm. We have established the 1-st order convergence properties of the affine scaling trust region method proposed in [26]. The 2-nd order convergence properties are presented but the proof is omitted due to its strong similarity to the proof in [11].

Due to time limit, we haven't investigated much on the performance of the SG method for minimizing a differentiable function plus an $\mathcal{L}_1$-norm. Application of our proposed affine scaling gradient based method to practical problems including volatility surface calibration has not been studied yet. We also believe that we can establish the proof of the convergence of the proposed scaled gradient method in more general cases. All these work will be done in our future work.

Bibliography

[1] Barzilai, J. and Borwein, J.M. , "Two Point Step Size Gradient Methods," in IMA Journal of Numerical Analysis Vol.8, pp.$141 - 148$, 1988 1, 7, 18

[2] Marcos Raydan, "On the Barzilai and Borwein Choice of Steplength for the Gradient Method," in IMA Journal of Numerical Analysis, Vol.13, pp.$321 - 326$, 1993 7

[3] Y.H.Dai and L.Z.Liao, "R-linear Convergence of the Barzilai and Borwein Gradient Method," in IMA Journal of Numerical Analysis, Vol.22, pp.$1 - 10$, 2002 7

[4] R.Fletcher, "On the Barzilai-Borwein Method," in Numerical Analysis Report, Chapter 10, pp. $235 - 256$ 7, 21

[5] A.Friedlander, J.M. Martinez, B.Molina, and M.Raydan, "Gradient Method with Retards and Generalizations," in SIAM Journal of Numerical Analysis, Vol.36, pp. $275 - 289$, 1999 8

[6] Bin Zhou, Li Gao and Yuhong Dai, "Gradient Methods with Adaptive Step-sizes," in Computational Optimization and Applications, Vol.35, pp. $69 - 86$, 2006 8

[7] Yu-Hong Dai and Roger Fletcher, "Projected Barzilai-Borwein Methods for Large-scale Box-constrained Quadratic Programming," in Numerical Mathematics, Vol.100, pp. $21-47$, 2005 8, 19, 28

[8] L. Grippo, F. Lampariello and S. Lucidi, “A Non-monotone Line Search Technique for Newton’s Method,” in SIAM, Journal of Numerical Analysis, Vol. 23, pp. $707-716$, 1986 5, 8, 21

[9] E.G.Birgin, J.M.Martinez and M.Raydan, “Spectral Projected Gradient Methods,” in Encyclopedia of Optimization, 2nd ed., Springer US, pp. $3652-3659$, 2009 8, 19, 21, 25, 30, 32

[10] Almut Eisenträger, “Non-monotone Line Search and Trust Region Methods for Optimization,” University of Oxford, 2007 15

[11] 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. $418-445$ 10, 75

[12] J. J. More, “Recent Developments in Algorithms and Software for Trust Region Methods” in Mathematical Programming: The State of the Art, M.G.A. Bachem and E.B. Dorte, Springer-Verlag, Berlin, 1983 44

[13] Dimitri P. Bertsekas, “Nonlinear Programming,” ISBN $1-886529-00-0$, 2nd printing, 2003 2, 5, 11

[14] R.Tibshirani, “Regression Shrinkage and Selection via the Lasso”, technical report, University of Toronto, 1994 2

[15] Scott Shaobing Chen, David L. Donoho and Michael A. Saunders, “Atomic Decomposition by Basis Pursuit,” in SIAM Journal on Scientific Computing, Vol.20, No.1, pp. $33-61$, 1999 3

[16] M.Osborne, B.Presnell and B.Turlach, “A New Approach to Variable Selection in Least Squares Problems,” in IMA Journal of Numerical Analysis, Vol.20, No. 3, pp. $389-403$, 2000 3

[17] Jorge Nocedal and Stephen J. Wright, “Numerical Optimization,” Second Edition, ISBN: $978-0-387-30303-1$, 2006 6

[18] Emmanuel Candès, Justin Romberg, and Terence Tao, “Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information,” in IEEE Trans. on Information Theory, Vol.52, No.2, pp. $489-509$, February 2006 2

[19] Yin Zhang, “A Simple Proof for Recoverability of $\mathcal{L}_1$-minimization: Go Over or Under?” in Rice CAAM Department Technical Report TR05-09, 2005 2

[20] Yin Zhang, “A Simple Proof for Recoverability of $\mathcal{L}_1$-minimization (II): The Nonnegative Case,” in Rice CAAM Department Technical Report TR05-10, 2005 2

[21] Simon Perkins, Kevin lacker, and James Theiler, “Grafting: Fast, Incremental Feature Selection by Gradient Descent in Function Space,” in Journal of Machine Learning Research, Vol.3, pp. $1333-1356$, 2003 4

[22] Shirish Krishnaj Shevade and S.Sathiya Keerthi, “A Simple and Efficient Algorithm for Gene Selection Using Sparse Logistic Regression,” in Bioinformatics, Vol. 19, pp. $2246-2253$, 2003 4

[23] Jorge J. Moré and D. C. Sorensen, “Computing a Trust Region Step,” in SIAM, Journal of Sci. and Stat. Comput., Vol.4, No.3, pp. $553-572$, 1983 57

[24] Thomas F. Coleman and Yuying Li, “A Trust Region and Affine Scaling Interior Point Method for Nonconvex Minimization with Linear Inequality Constraints,” in Math. Program, Ser A 88, pp.$1-31$, 2000 1

[25] Jianqing Fan and Runze Li, “Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties,” in Journal of the American Statistical Association, Vol.96, No.456 , pp. $1348-1360$,2001 4

[26] Thomas F. Coleman, Yuying Li and Cheng Wang, “Stable Local Volatility Function Calibration Using Kernel Splines,” in print, 2010 iii, vii, 5, 6, 9, 10, 33, 35, 75

[27] Jorge J. Moré and Gerardo Toraldo, “On the Solution of Large Quadratic Programming Problems with Bound Constraints,” in SIAM Journal of Optimization, Vol.1, No.1, pp. 93 – 113, 1991 23

[28] R. Fletcher, “Practical Methods of Optimization: Volume 2, Constrained Optimization,” John Wiley and Sons, 1981 12

[29] Levent Tunçel, “Polyhedral and Semidefinite Programming Methods in Combinatorial Optimization,” Fields Institute Monograph Series, AMS, 2008 3