Notes on Numerical Optimization

University of Chicago, 2014

Vivak Patel

October 18, 2014

Contents

List of Algorithms

Part I
Fundamentals of Optimization

1 Overview of Numerical Optimization

1.1 Problem and Classification

1. Problem:
$$\arg\min_{z\in\mathbb{R}^n} f(z) : \begin{cases} c_i(z) = 0 & i \in \mathcal{E} \\ c_i(z) \geq 0 & i \in \mathcal{I} \end{cases}$$
   (a) $f : \mathbb{R}^n \to \mathbb{R}$ is known as the objective function
   (b) $\mathcal{E}$ are equality constraints
   (c) $\mathcal{I}$ are inequality constraints
2. Classifications
   (a) Unconstrained vs. Constrained. If $\mathcal{E} \cup \mathcal{I} = \emptyset$ then it is an unconstrained problem.
   (b) Linear Programming vs. Nonlinear Programming. When $f(z)$ and $c_i(z)$ are all linear functions of z then this is a linear programming problem. Otherwise, it is a nonlinear programming problem.
   (c) Note: This is not the same as having linear or nonlinear equations.

1.2 Theoretical Properties of Solutions

1. Global Minimizer. Weak Local Minimizer (Local Minimizer). Strict Local Minimizer. Isolated Local Minimizer.

   Definition 1.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$*
   (a) x^* is a global minimizer of f if $f(x^*) \leq f(x)$ for all $x \in \mathbb{R}^n$
   (b) x^* is a local minimizer of f if for some neighborhood, N of x^*, $f(x^*) \leq f(x)$ for all $x \in N(x^*)$.
   (c) x^* is a strict local minimizer of f if for some neighborhood, $N(x^*)$, $f(x^*) < f(x)$ for all $x \in N(x^*)$.
   (d) x^* is an isolated local minimizer of f if for some neighborhood, $N(x^*)$, there are no other minimizers of f in $N(x^*)$.

   Note 1.1. *Every isolated local minimizer is a strict minimizer, but it is not true that a strict minimizer is always an isolated local minimizer.*

2. Taylor's Theorem

   Theorem 1.1. *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. Let $p \in \mathbb{R}^n$. Then for some $t \in [0, 1]$:*
$$f(x + p) = f(x) + \nabla f(x + tp)^T p \tag{1}$$

If f is twice continuously differentiable, then:

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+tp)p dt \tag{2}$$

And for some $t \in [0,1]$:

$$f(x+p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x+tp)p \tag{3}$$

3. Necessary Conditions

   (a) First Order Necessary Conditions

   Lemma 1.1. *Let f be continuously differentiable. If x^* is a minimizer of f then $\nabla f(x^*) = 0$.*

   Proof. Suppose $\nabla f(x^*) \neq 0$. Let $p = -\nabla f(x^*)$. Then, $p^T \nabla f(x^*) < 0$ and by continuity there is some $T > 0$ such that for any $t \in [0,T)$, $p^T \nabla f(x^* + tp) < 0$. We now apply **Equation 1** of Taylor's Theorem to get a contradiction.

   Let $t \in [0,T)$ and $q = x + tp$. Then $\exists t' \in [0,t]$ such that:

$$f(x^* + q) = f(x^*) + t\nabla f(x^* + t'tp)^T p < f(x^*)$$

   □

   (b) Second Order Necessary Conditions

   Lemma 1.2. *Suppose f is twice continuously differentiable. If x^* is a minimizer of f then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succeq 0$ (i.e. positive semi-definite)*

   Proof. The first conclusion follows from the first order necessary condition. We proceed with the second one by contradiction and Taylor's Theorem. Suppose $\nabla^2 f(x^*) \prec 0$. Then there is a p such that:

$$p^T \nabla^2 f(x^*) p < 0$$

   By continuity, there is a $T > 0$ such that $p^T \nabla^2 f(x^* + tp)p < 0$ for $t \in [0,T]$. Fix $t \in [0,T]$ and let $q = tp$. Then, by **Equation 3** from Taylor's Theorem, $\exists t' \in [0,t]$ such that:

$$f(x^* + q) = f(x^*) + \frac{1}{2} t^2 p^T \nabla^2 f(x^* + t'tp)p < f(x^*)$$

   □

4. Second Order Sufficient Condition

   Lemma 1.3. *Let $x^* \in \mathbb{R}^n$. Suppose $\nabla^2 f$ is continuous, $\nabla^2 f(x^*) \succ 0$, and $\nabla f(x^*) = 0$. Then x^* is a strict local minimizer.*

Proof. By continuity, there is a ball of radius T about x^*, B, in which $\nabla^2 f(x) \succ 0$. Let $\|p\| < T$. Then, there exists a $t \in [0,1]$ such that:

$$f(x^* + p) = f(x^*) + \frac{1}{2} p^T \nabla^2 f(x^* + tp) p > f(x^*)$$

□

5. Convexity and Local Minimizers

Lemma 1.4. *When f is convex any local minimizer x^* is a global minimizer. If, in addition, f is differentiable then any stationary point is a global minimizer.*

Proof. Suppose x^* is not a global minimizer. Let z be the global minimizer. Then $f(x^*) > f(z)$. By convexity, for any $\lambda \in [0,1]$:

$$f(x^*) > \lambda f(z) + (1-\lambda) f(x^*) \geq f(\lambda z + (1-\lambda) x^*)$$

So as $\lambda \to 0$, we see that in any neighborhood of x^* there is a point w such that $f(w) < f(x^*)$. A contradiction.

For the second part, suppose z is as above. Then:

$$\begin{aligned} \nabla f(x^*)^T (z - x^*) &= \lim_{\lambda \downarrow 0} \frac{f(x^* + \lambda(z - x^*)) - f(x^*)}{\lambda} \\ &\leq \lim_{\lambda \downarrow 0} \frac{\lambda f(z) + (1-\lambda) f(x^*) - f(x^*)}{\lambda} \\ &\leq f(z) - f(x^*) \\ &< 0 \end{aligned}$$

□

1.3 Algorithm Overview

1. In general, algorithms begin with a seed point, x_0, and locally search for decreases in the objective function, producing iterates x_k, until stopping conditions are met.
2. Algorithms typically generate a local model for f near a point x_k:

$$f(x_k + p) \approx m_k(p) = f_k + p^T \nabla f_k + \frac{1}{2} p^T B_k p$$

3. Different choices of B_k will lead to different methods with different properties:
   (a) $B_k = 0$ will lead to steepest descent methods
   (b) Letting B_k be the closest positive definite approximation to $\nabla^2 f_k$ leads to newton's methods
   (c) Iterative approximations to the Hessian given by B_k lead to quasi-newton's methods
   (d) Conjugate Gradient methods update p without explicitly computing B_k

1.4 Fundamental Definitions and Results

1. Q-convergence.

   Definition 1.2. *Let $x_k \to x$ in $\mathbb{R}^n$.*

   (a) x_k converge Q-linearly if $\exists q \in (0,1)$ such that for sufficiently large k:

   $$\frac{\|x_{k+1} - x\|}{\|x_k - x\|} \leq q$$

   (b) x_k converge Q-superlinearly if:

   $$\lim_{k\to\infty} \frac{\|x_{k+1} - x\|}{\|x_k - x\|} \to 0$$

   (c) x_k converges Q-quadratically if $\exists q^2 > 0$ such that for sufficiently large k:

   $$\frac{\|x_{k+1} - x\|}{\|x_k - x\|^2} \leq q$$

2. R-convergence.

   Definition 1.3. *Let $x_k \to x$ in $\mathbb{R}^n$*

   (a) x_k converges R-linearly if $\exists v_k \in \mathbb{R}$ converging Q-linearly such that

   $$\|x_k - x\| \leq v_k$$

   (b) x_k converges R-superlinearly if $\exists v_k \in \mathbb{R}$ converging Q-superlinearly such that

   $$\|x_k - x\| \leq v_k$$

   (c) x_k converges R-quadratically if $\exists v_k \in \mathbb{R}$ converging Q-quadratically such that

   $$\|x_k - x\| \leq v_k$$

3. Sherman-Morrison-Woodbury

   Lemma 1.5. *If A and $\tilde{A}$ are non-singular and related by $\tilde{A} = A + ab^T$ then:*

   $$\tilde{A}^{-1} = A^{-1} - \frac{A^{-1}ab^TA^{-1}}{1 + b^TA^{-1}a}$$

   Proof. We can check this by simply plugging in the formula. However, to "derive" this:

   $$\begin{aligned}
   \tilde{A}^{-1} &= \left(A + ab^T\right)^{-1} \\
   &= \left(A\left[I + A^{-1}ab^T\right]\right)^{-1} \\
   &= \left[I + A^{-1}ab^T\right]^{-1} A^{-1} \\
   &= \left[I - A^{-1}ab^T + A^{-1}ab^TA^{-1}ab^T - \cdots\right] A^{-1} \\
   &= \left[I - A^{-1}a(1 - b^TA^{-1}a + \cdots)b^T\right] A^{-1} \\
   &= A^{-1} - \frac{A^{-1}ab^TA^{-1}}{1 + b^TA^{-1}a}
   \end{aligned}$$

   □

2 Line Search Methods

2.1 Step Length

1. Descent Direction

   Definition 2.1. *$p \in \mathbb{R}^n$ is a descent direction if $p^T \nabla f_k < 0$.*

2. Problem: Find $\alpha_k \in \arg\min \phi(\alpha)$ where $\phi(\alpha) = f(x_k + \alpha p_k)$. However, it is too costly to find the exact minimizer of ϕ. Instead, we find an α which acceptably reduces ϕ.

3. Wolfe Condition

   (a) Armijo Condition. Curvature Condition. Wolfe Condition. Strong Wolfe Condition.

   Definition 2.2. *Fix x and p to be a descent direction. Let $\phi(\alpha) = f(x + \alpha p)$ and so $\phi'(\alpha) = f^T(x + \alpha p)p$. Fix $0 < c_1 < c_2 < 1$.*

   i. α satisfies the Armijo Condition if $\phi(\alpha) \leq \phi(0) + \alpha c_1 \phi'(0)$. Note $l(\alpha) := \phi(0) + \alpha c_1 \phi'(0)$.

   ii. α satisfies the Curvature Condition if $\phi'(\alpha) \geq c_2 \phi'(0)$

   iii. α satisfies the Strong Curvature Condition if $|\phi'(\alpha)| \leq c_2 |\phi'(0)|$

   iv. The Wolfe Condition is the Armijo Condition with the Curvature Condition

   v. The Strong Wolfe Condition is the Armijo Condition with the Strong Curvature Condition.

   (b) The Armijo Condition: guarantees a decrease at the next iterate by ensuring $\phi(alpha) < \phi(0)$.

   (c) Curvature Condition

   i. If $\phi'(\alpha) < c_2 \phi'(0)$, then ϕ is still decreasing at α and so we improve the reduction in f by taking a larger α

   ii. If $\phi'(\alpha) \geq c_2 \phi'(0)$, then either we are closer to a minimum where $\phi' = 0$ or $\phi' > 0$ which means we have surpassed the minimum.

   iii. The strong condition guarantees that the choice of α is closer to $\phi' = 0$.

   (d) Existence of Step satisfying Wolfe and Strong Wolfe Conditions

   Lemma 2.1. *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. Let p be a descent direction at x and assume that f is bounded from below along the ray $\{x + \alpha p : \alpha > 0\}$. If $0 < c_1 < c_2 < 1$ there exists an interval satisfying the Wolfe and Strong Wolfe Condition.*

   Proof.

   i. Satisfying Armijo Condition: Let $l(\alpha) = f(x) + c_1 \alpha p^T \nabla f(x)$. Since $l(\alpha)$ is decreasing and $\phi(\alpha)$ is bounded from below, eventually $\phi(\alpha) = l(\alpha)$. Let $\alpha_A > 0$ be the first time this intersection occurs. Then $\phi(\alpha) < l(\alpha)$ for $\alpha \in (0, \alpha_A)$.

ii. Curvature Condition: By the mean value theorem, $\exists \beta \in (0, \alpha_A)$ such that (since $l(\alpha_A) = \phi(\alpha_A)$):

$$\alpha_A \phi'(\beta) = \phi(\alpha_A) - \phi(0) = c_1 \alpha_A \phi'(0) > c_2 \alpha_A \phi'(0)$$

And since ϕ' is smooth there is an interval containing β for which this inequality holds.

iii. Strong Curvature Condition. Since $\phi'(\beta) = c_1 \phi'(0) < 0$, it follows that:

$$|\phi'(\beta)| < |c_2 \phi'(0)|$$

□

4. Goldstein Condition

(a) Goldstein Condition

Definition 2.3. *Take $c \in (0, 1/2)$. The Goldstein Condition is satisfied by α if:*

$$\phi(0) + (1-c)\alpha\phi'(0) \leq \phi(\alpha) \leq \phi(0) + c\alpha\phi'(0)$$

(b) Existence of Step satisfying Goldstein Condition

Lemma 2.2. *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. Let p be a descent direction at x and assume that f is bounded from below along the ray $\{x + \alpha p : \alpha > 0\}$. Let $c \in (0, 1/2)$. Then there exists an interval satisfying the Goldstein condition.*

Proof. Let $l(\alpha) = \phi(0) + (1-c)\alpha\phi'(0)$ and $u(\alpha) = \phi(0) + c\alpha\phi'(0)$. For $\alpha > 0$, $l(\alpha) < u(\alpha)$ since $c \in (0, 1/2)$. Since ϕ is bounded from below let α_u be the smallest point of intersection between $u(\alpha)$ and $\phi(\alpha)$ for $\alpha > 0$. And let α_l be the largest point of intersection, less than α_u, of $l(\alpha)$ and $\phi(\alpha)$. Then, for $\alpha \in (\alpha_l, \alpha_u)$, $l(\alpha) < \phi(\alpha) < u(\alpha)$. □

5. Backtracking

(a) Backtracking

Definition 2.4. *Let $\rho \in (0, 1)$ and $\bar{\alpha} > 0$. Backtracking checks over a sequence $\bar{\alpha}, \rho\bar{\alpha}, \rho^2\bar{\alpha}, \ldots$ until a α is found satisfying the Armijo condition.*

(b) Existence of Step satisfying Backtracking

Lemma 2.3. *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. Let p be a descent direction at x. Then there is a step produced by the backtracking algorithm which satisfies the Armijo condition.*

Proof. $\exists \epsilon > 0$ such that for $0 < \alpha < \epsilon$, $\phi(\alpha) < \phi(0) + c\alpha\phi'(0)$ by continuity. Since $\rho^n \bar{\alpha} \to 0$, there is an n such that $\rho^n \bar{\alpha} < \epsilon$. □

2.2 Step Length Selection Algorithms

1. Algorithms which implement the Wolfe or Goldstein Conditions exist and are ideal, but are not discussed herein.

2. Backtracking Algorithm

Algorithm 1: Backtracking Algorithm
input : Reduction rate $\rho \in (0,1)$, Initial estimate $\bar{\alpha} > 0$, $c \in (0,1)$, $\phi(\alpha)$
$\alpha \leftarrow \bar{\alpha}$
while $\phi(\alpha) \geq \phi(0) + \alpha c \phi'(0)$ **do**
  $\alpha \leftarrow \rho\alpha$
end
return α

3. Interpolation with Expensive First Derivatives

   (a) Suppose the Interpolation technique produces a sequence of guesses $\alpha_0, \alpha_1, \ldots$.

   (b) Interpolation produces α_k by modeling ϕ with a polynomial m (quadratic or cubic) and then minimizes m to find a new estimate for α. The parameters of the polynomial are given by requiring:

      i. $m(0) = \phi(0)$
      ii. $m'(0) = \phi'(0)$
      iii. $m(\alpha_{k-1}) = \phi(\alpha_{k-1})$
      iv. $m(\alpha_{k-2}) = \phi(\alpha_{k-2})$

   (c) When $k = 1$, we model using a quadratic polynomial

4. Interpolation with Inexpensive First Derivatives

   (a) Suppose interpolation produces a sequence of guesses $\alpha_0, \alpha_1, \ldots$

   (b) In this case, m is always a cubic polynomial such that, α_k is computed by:

      i. $m(\alpha_{k-1}) = \phi(\alpha_{k-1})$
      ii. $m(\alpha_{k-2}) = \phi(\alpha_{k-2})$
      iii. $m'(\alpha_{k-1}) = \phi'(\alpha_{k-1})$
      iv. $m'(\alpha_{k-2}) = \phi'(\alpha_{k-2})$

5. Interpolation Algorithm

Algorithm 2: Interpolation Algorithm
input : Feasible Search Region $[\bar{a}, \bar{b}]$, Initial estimate $\alpha_0 > 0$, ϕ
$\alpha \leftarrow 0$, $\beta \leftarrow \alpha_0$
while $\phi(\beta) \geq \phi(0) + c\beta\phi'(0)$ **do**
  Approximate m
  if *Inexpensive Derivatives* **then**
    $\alpha \leftarrow \beta$
  end
  Explicitly compute minimizer of m in feasible region. Store as β.
end
return β

2.3 Global Convergence and Zoutendjik

1. Globally Convergent. Zoutendjik Condition.

   Definition 2.5. *Suppose we have an algorithm, Ω which produces iterates (x_k) and denote $\nabla f(x_k) = \nabla f_k$.*

   (a) Ω is globally convergent if $\|\nabla f_k\| \to 0$. (That is, x_k converge to a stationary point.)

   (b) Suppose Ω produced search directions p_k such that $\|p_k\| = 1$. And let θ_k be the angle between ∇f_k and p_k. Ω satisfies the Zoutendjik Condition if

   $$\sum_{k=1}^{\infty} \cos^2(\theta_k) \|\nabla f_k\|^2 < \infty$$

2. Zoutendjik Condition & Angle Bound implies global convergence

   Lemma 2.4. *Suppose Ω produces a sequence $(x_k, p_k, \nabla f_k, \theta_k)$ such that $\exists \delta > 0$ and $\forall k \geq 1$, $\cos(\theta_k) \geq \delta$. If Ω satisfies the Zoutendjik condition then Ω is globally convergent.*

   Proof. By the Zoutendjik condition:

   $$\delta^2 \sum_{k=1}^{\infty} \|\nabla f_k\|^2 < \infty$$

   Therefore, $\|\nabla f_k\|^2 \to 0$. □

3. Example of Angle Bound

   Example 2.1. *Suppose $p_k = -B_k^{-1} \nabla f_k$ where B_k are positive definite and $\|B_k\| \left\|B_k^{-1}\right\| < M < \infty$ for all k. Letting $\|\cdot\| = \|\cdot\|_2$, $B_k = X \Lambda X^T$ be the EVD of B_k, and $X^T \nabla f_k = z$:*

   $$\begin{aligned}
   |\cos(\theta_k)| &= \left| \frac{-\nabla f_k^T B_k^{-1} \nabla f_k}{\|\nabla f_k\| \left\|B_k^{-1} \nabla f_k\right\|} \right| \\
   &= \left| \frac{z^T \Lambda^{-1} z}{\|z\| \|\Lambda^{-1} z\|} \right| \\
   &= \left| \frac{\sum_{i=1}^n \frac{z_i^2}{\lambda_i}}{\|z\| \sqrt{\sum_{i=1}^n \frac{z_i^2}{\lambda_i^2}}} \right| \\
   &\geq \frac{1/\lambda_1 \|z\|^2}{1/\lambda_n \|z\|^2} \\
   &\geq \frac{\lambda_n}{\lambda_1} \\
   &\geq \frac{1}{M}
   \end{aligned}$$

   Hence, if Ω satisfies the Zoutendjik Condition, we see that $\lim_n \|\nabla f_n\| \to 0$.

4. Wolfe Condition Line Search Satisfies Zoutendjik Condition.

Theorem 2.1. *Suppose we have an objective f satisfying:*

(a) f is bounded from below in $\mathbb{R}^n$

(b) Given an initial x_0, there is an open set N containing $\mathcal{L} = \{x : f(x) \leq f(x_0)$

(c) f is continuously differentiable on N

(d) ∇f is Lipschitz continuous on N

Suppose we have an algorithm Ω producing $(x_k, p_k, \nabla f_k, \theta_k, \alpha_k)$ such that:

(a) p_k is a descent direction (with $\|p_k\| = 1$)

(b) α_k satisfies the Wolfe Conditions

Then Ω satisfies the Zoutendjik Condition.

Proof. The general strategy is to lower bound α_k uniformly by $C|\nabla f_k^T p_k|$, where $C > 0$. Then using the descent condition:

$$\begin{aligned} f_{k+1} &\leq f_k - c\alpha_k |\nabla f_k^T p_k| \\ &\leq f_k - C|\nabla f_k^T p_k|^2 \\ &\leq f_k - C\cos^2(\theta_k)\|\nabla f_k\|_2^2 \qquad \leq f_0 - C\sum_{j=1}^{k} \cos^(\theta_j)\|\nabla f_j\|_2^2 \end{aligned}$$

And we can conclude that since f is bounded from below by l, then $f_0 - f_k \leq f_0 - l < \infty$. The result follows.

So now we set out to show that $\alpha_k \geq C|\nabla f_k^T p_k|$. We do this by leveraging the Curvature Condition.

(a) From the curvature condition, we have that:

$$(\nabla f_{k+1} - \nabla f_k)^T p_k \geq (c_2 - 1)\nabla f_k^T p_k$$

(b) From Lipschitz Continuity, with constant L:

$$|(\nabla f_{k+1} - \nabla f_k)^T p_k| \leq \|\nabla f_{k+1} - \nabla f_k\| \|p_k\| \leq L\alpha_k \|p_k\|^2 = L\alpha_k$$

Together these imply that $\frac{1-c_2}{L}|\nabla f_k^T p_k| \leq \alpha_k$. □

5. Goldstein Condition Line Search satisfies Zoutendjik Condition

Theorem 2.2. *Suppose f has the same properties as it does in **Theorem 2.1**. And suppose Ω produces $(x_k, p_k, \nabla f_k, \theta_k, \alpha_k)$ such that:*

(a) p_k is a descent direction with $\|p_k\| = 1$

(b) α_k satisfies the Goldstein conditions.

Then Ω satisfies the Zoutendjik Condition.

Proof. We use Taylor's Theorem and then Lispchitz Continuity to find the lower bound. By Taylor's theorem, for $t_k \in (0,1)$:

$$(1-c)\alpha_k \nabla f_k^T p_k \leq f_{k+1} - f_k = \nabla f(x_k + t_k \alpha_k p_k)^T \alpha_k p_k$$

Therefore,

$$\begin{aligned} c\alpha_k |\nabla f_k^T p_k| &\leq |(\nabla f(x_k + t_k \alpha_k p_k)^T - \nabla f_k^T) p_k| \\ &\leq L t_k \alpha_k \left\| p_k \right\|^2 \\ &\leq L \alpha_k \end{aligned}$$

The result follows. □

6. Backtracking Line Search satisfies Zoutendjik Condition

Theorem 2.3. *Suppose f has the same properties as it does in **Theorem 2.1**. And suppose Ω produces $(x_k, p_k, \nabla f_k, \theta_k, \alpha_k)$ such that:*

(a) p_k is a descent direction with $\|p_k\| = 1$

(b) α_k is selected by backtracking with $\bar{\alpha} = 1$

Then Ω satisfies the Zoutendjik condition.

Proof. Suppose Ω uses $\rho \in (0,1)$. There are two cases to consider. When $\alpha_k = 1$ and $\alpha_k < 1$. We denote the first subsequence by $k(j)$ and the second by $k(l)$. For $\alpha_{k(l)}$ we know at least that $\alpha_{k(l)}/\rho$ was rejected, that is:

$$f(x_k + \alpha_{k(l)} p_k/\rho) > f_k + c\alpha_{k(l)} \nabla f_k^T p_k/\rho$$

Using the same strategy as in Goldstein, $\exists t_k$ such that:

$$\nabla f(x_k + t_k \alpha_{k(l)} p_k/\rho)^T \alpha_k p_k/\rho > c\alpha_{k(l)} \nabla f_k^T p_k/\rho$$

Therefore, by Lipschitz continuity:

$$(1-c)|\nabla f_k^T p_k| \leq L t_k \alpha_{k(l)}/\rho \leq L\alpha_{k(l)}/\rho$$

We now consider $\alpha_{k(j)} = 1$ since the Armijo condition gives the sequence:

$$f_{k(j+1)} \leq f_{k(j)+1} \leq f_{k(j)} + c\nabla f_k^T p_k = f_{k(j)} - c|\cos(\theta_{k(j)s})| \left\| f_{k(j)} \right\|$$

Therefore:

$$c\sum_j |\cos(\theta_{k(j)})| \left\| \nabla f_{k(j)} \right\| < \infty$$

Then $\exists J$ such that $j \geq J$,

$$\cos^2(\theta_{k(j)}) \left\| \nabla f_{k(j)} \right\|^2 \leq |\cos(\theta_{k(j)})| \left\| \nabla f_{k(j)} \right\| < 1$$

Therefore, the Zoutendjik condition holds. □

3 Trust Region Methods

3.1 Trust Region Subproblem and Fidelity

1. The strategy of the trust region problem is as follows: at any point x we have a model $m_x(p)$ of $f(x+p)$ which we trust is a good approximation in a region $\|p\| \leq \Delta$. Minimizing this, we have a new iterate on which to continue applying the approach.

2. Trust Region Subproblem

   Definition 3.1. *Let $x \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ be an objective function with model $m_x(p)$ in a region $\|p\| \leq \Delta$. The Trust Region Subproblem is to find*

   $$\arg\min\{m_x(p) : \|p\| \leq \Delta\}$$

   Note 3.1. *Usually m_x is taken to be*

   $$m_x(p) = f(x) + \nabla f_x^T p + \frac{1}{2} p^T B p$$

   where B is the model Hessian and is not necessarily positive definite.

3. Fidelity

   Definition 3.2. *The Fidelity of m_x in a region $\|p\| \leq \Delta$ at a point $p + x$ is*

   $$\rho(m_x, f, p, \Delta) = \frac{f(x+p) - f(x)}{m_x(p) - m_x(0)}$$

4. Fidelity and Trust Region. Let $0 < c_1 < c_2 < 1$

   (a) If $\rho \geq c_2$, especially if p is intended for descent, the model is a good approximation to f and we can expand the trust region Δ

   (b) If $\rho \leq c_1$, especially if p is intended for descent, the model is a poor approximation to f and we should shrink the trust region Δ

   (c) Otherwise, the model is a sufficient approximation to f and the trust region should not be adjusted.

5. Fidelity and Iterates. Suppose p causes a descent in m_x. And let $\eta \in (0, 1/4]$

   (a) If $\rho < \eta$ then we do not have a sufficient decrease in the function and the process should be repeated with x

   (b) If $\rho \geq \eta$ then we have a sufficient decrease in f, and we should continue the process with $x + p$

6. Notation

   (a) Let $x_0, x_1, \ldots$ be the iterates produces by subsequential solutions to the trust region subproblem for some objective f

   (b) $m_{x_k}(p) =: m_k(p)$

   (c) Δ for a particular m_k is denoted Δ_k

   (d) Accepted solutions to the subproblem at x_k are p_k so that $x_{k+1} = x_k + p_k$.

3.2 Fidelity Algorithms

1. Fidelity and Trust Region

Algorithm 3: Trust Region Management Algorithm
input : Thresholds $0 < c_2 < c_1 < 1$, Trust Region Δ, $\Delta_{\max}$, ρ_x
if $\rho_x > c_1$ **then**
  $\Delta \leftarrow \min(2\Delta, \Delta_{\max})$
else if $\rho < c_2$ **then**
  $\Delta \leftarrow \Delta/2$
else
  Continue
end
return Δ

Note 3.2. *Usually* $c_1 = 1/4$ *and* $c_2 = 3/4$

2. Fidelity and Solution Acceptance

Algorithm 4: Solution Acceptance Algorithm
input : Treshold $\eta \in (0, 1/4]$, ρ,x,p
if $\rho_x(p) \geq \eta$ **then**
  $x = x + p$
else
  Continue
end
return x

3.3 Approximate Solutions to Subproblem

3.3.1 Cauchy Point

1. Cauchy Point

Definition 3.3. *Let* m_x *be a model of* f *within* $\|p\| \leq \Delta$. *Let* p^S *be the direction of steepest descent at* $m_x(0)$ *such that* $\left\|p^S\right\| = 1$, *and let* τ *be:*

$$\arg\min\{m_x(\tau p^s) : \tau \leq \Delta\}$$

Then, $p^C = \tau p^S$ *is the Cauchy Point of* m_x.

Note 3.3. *The Cauchy point is the line search minimizer of* m_k *along the direction of steepest descent subject to* $\|p\| \leq \Delta$.

2. Computing the Cauchy Point

Lemma 3.1. *Let* m_x *be a quadratic model of* f *within* $\|p\| \leq \Delta$ *such that*

$$m_x(p) = f(x) + g^T p + \frac{1}{2} p^T B p$$

Then, its Cauchy Point is:

$$p^C = \begin{cases} \frac{-g}{\|g\|} \min\left(\Delta, \frac{\|g\|^3}{g^T B g}\right) & g^T B g > 0 \\ \frac{-g}{\|g\|}\Delta & g^T B g \leq 0 \end{cases}$$

Proof. First, we compute the direction of steepest descent:

$$p^S = \frac{-g}{\|g\|}$$

Then, we have that $m_x(tp^S) = f(x) + tg^T p^S + \frac{t^2}{2}(p^S)^T B p^S$. If the quadratic term is negative, then $t = \Delta$ is the minimizer. And so one option is:

$$p^C = \frac{-\Delta g}{\|g\|}$$

If the quadratic term is positive then:

$$\begin{aligned}\frac{d}{dt} m_x(tp^S) &= \frac{d}{dt} \\ &= g^T p^S + tp^S B p^S\end{aligned}$$

Setting this to 0 and substituting back in for p^S:

$$t = \frac{\|g\|^3}{g^T B g}$$

This value could potentially be larger than Δ, so to safeguard against this, in this case:

$$\tau = \min\left(\Delta, \frac{\|g\|^3}{g^T B g}\right)$$

□

3. Using the Cauchy point does not produce the best rate of convergence

3.3.2 Dogleg Method

1. Requirement: $B \succ 0$
2. Intuitively, the dogleg method moves towards the minimizer of m_x which is $p^{\min} = -B^{-1}g$ until it reaches this point or hits the boundary of the trust region. It does this by first moving to the Cauchy Point, and then along a linear path to $p^{\min}$
3. Dogleg Path

   Definition 3.4. *Let m_x be the quadratic model with $B \succ 0$ of f with radius Δ. Let p^C be the Cauchy Point and $p^{\min}$ be the unconstrained minimizer of m_x. The Dogleg Path is $p^{DL}(t)$ where:*

$$p^{DL}(t) = \begin{cases} p^C t & t \in [0,1] \\ p^C + (t-1)(p^{\min} - p^C) & t \in [1,2] \end{cases}$$

4. Dogleg Method

   Definition 3.5. *The Dogleg Method chooses $x_{k+1} = x_k + p^{DL}(\tau)$ where $\|p^{DL}(\tau)\| = \Delta$.*

5. Properties of Dogleg Path

Lemma 3.2. *Suppose $p^{DL}(t)$ is the Dogleg path of m_x with $\Delta = \infty$ and $B \succ 0$. Then:*

(a) $m_x(p^{DL}(t))$ is decreasing

(b) $\left\|p^{DL}(t)\right\|$ is increasing

Proof. When $0 < t < 1$, this case is uninteresting since for these values we know $m_x(p^{DL}(t))$ is decreasing and $\left\|p^{DL}(t)\right\|$ is increasing. So we consider $1 < t < 2$ and reparameterize using $t - 1 = z$. We have that:

$$\begin{aligned}\frac{d}{dz} m_x(p^{DL}(t)) &= (p^{\min} - p^C)^T(g - Bp^C) + z(p^{\min} - p^C)^T B(p^{\min} - p^C) \\ &= -(1 - z)(p^{\min} - p^C)^T B(p^{\min} - p^C)\end{aligned}$$

Letting $c = (p^{\min} - p^C)^T B(p^{\min} - p^C) > 0$ since $B \succ 0$, we have that the derivative is strictly negative for $z < 1$. To show that the norm is increasing, first we show that the angle between p^C and $p^{\min}$ is between $(-\pi/2, \pi/2)$, then we show its length is longer.

(a) For the angle, $g \neq 0$:

$$\langle p^{\min}, p^C \rangle = g^T B^{-1} g \frac{\|g\|^2}{g^T B g} > 0$$

(b) Using Cauchy-Schwartz, $\|g\| \left\|p^{\min}\right\| \geq \|g\| \left\|p^C\right\|$ since:

$$\left\|B^{-1}g\right\| \|g\| \geq g^T B^{-1} g = \frac{g^T B^{-1} g g^T B g}{g^T B g} \geq \frac{\|g\|^4}{g^T B g}$$

□

3.3.3 Global Convergence of Cauchy Point Methods

1. Reduction obtained by Cauchy Point

Lemma 3.3. *The Cauchy Point p_k^c satisfies*

$$m_k(p_k^C) - m_k(0) \leq -\frac{1}{2} \|g_k\| \min\left(\Delta_k, \frac{\|g_k\|}{\|B_k\|}\right)$$

Proof. We have that

$$m_k(p_k^C) - m_k(0) = g^T p_k^C + \frac{1}{2}(p_k^C)^T B_k(p_k^C)$$

This brings in two cases when $g^T Bg \leq 0$ then (dropping subscripts)

$$\begin{aligned} m_k(p^C) - m_k(0) &= -\|g\|\Delta + \frac{1}{2\|g\|^2} g^T Bg \\ &\leq -\|g\|\Delta \\ &\leq -\frac{1}{2}\|g\| \min\left(\Delta, \frac{\|g\|}{\|B\|}\right) \end{aligned}$$

When $g^T Bg \geq 0$, and $\Delta(g^T Bg) \leq \|g\|^3$, we are again in the above case:

$$\begin{aligned} m_k(p^C) - m_k(0) &= -\|g\|\Delta + \frac{1}{2\|g\|^2} g^T Bg \\ &\leq -\|g\|\Delta + \frac{1}{2}\|g\|\Delta \\ &\leq -\frac{1}{2}\|g\| \min\left(\Delta, \frac{\|g\|}{\|B\|}\right) \end{aligned}$$

When $\Delta g^T Bg > \|g\|^3$:

$$\begin{aligned} m_k(p^C) - m_k(0) &= -\frac{\|g\|^4}{g^T Bg} + \frac{\|g\|^4}{2(g^T Bg)^2} g^T Bg \\ &= -\frac{\|g\|^4}{2g^T Bg} \\ &\leq -\frac{1}{2}\frac{\|g\|^2}{\|B\|} \\ &\leq -\frac{1}{2}\|g\| \min\left(\Delta, \frac{\|g\|}{\|B\|}\right) \end{aligned}$$

□

2. Reduction achieved by Dogleg

Corollary 3.1. *The dogleg step p_k^{DL} satisfies:*

$$m_k(p_k^{DL}) - m_k(0) \leq -\frac{1}{2}\|g_k\| \min\left(\Delta_k, \frac{\|g_k\|}{\|B_k\|}\right)$$

Proof. Since $m_k(p_k^{DL}) \leq m_k^{p_k^C}$ the result follows. □

3. Reduction achieved by any method based on Cauchy Point

Corollary 3.2. *Suppose a method uses step p_k where $m_k(p_k) - m_k(0) \leq c(m_k(p_k^C) - m_k(0))$ for $c \in (0,1)$. Then p_k satisfies:*

$$m_k(p_k) - m_k(0) \leq -\frac{1}{2}c\|g_k\| \min\left(\Delta_k, \frac{\|g_k\|}{\|B_k\|}\right)$$

4. Convergence when $\eta = 0$

   Theorem 3.1. *Let $x_0 \in \mathbb{R}^n$ and f be an objective function. Let $S = \{x : f(x) \leq f(x_0)\}$ and $S(R_0) = \{x : \|x - y\| < R_0, y \in S\}$ be a neighborhood of radius R_0 about S. Suppose the following conditions:*

   (a) f is bounded from below on S

   (b) f is Lipschitz continuously differentiable on $S(R_0)$ for some $R_0 > 0$ and some constant L.

   (c) There is a c such that for all k:

$$m_k(p_k) - m_k(0) \leq -c \|g_k\| \min\left(\Delta_k, \frac{\|g_k\|}{\|B_k\|}\right)$$

   (d) There is a $\gamma \geq 1$ such that for all k:

$$\|p_k\| \leq \gamma \Delta_k$$

   (e) $\eta = 0$ (fidelity threshold)

   (f) $\|B_k\| \leq \beta$ for all k.

   Then:

$$\liminf \|g_k\| = 0$$

5. Convergence when $\eta \in (0, 1/4)$

   Theorem 3.2. *If $\eta \in (0, 1/4)$ with all other conditions from **Theorem 3.1** holding, then:*

$$\lim \|g_k\| = 0$$

3.4 Iterative Solutions to Subproblem

3.4.1 Exact Solution to Subproblem

1. Conditions for Minimizing Quadratics

   Lemma 3.4. *Let m be the quadratic function*

$$m(p) = g^T p + \frac{1}{2} p^T B p$$

   where B is a symmetric matrix.

   (a) m attains a minimum if and only if $B \succeq 0$ and $g \in Im(B)$. If $B \succeq 0$ then every p satisfying $Bp = -g$ is a global minimizer of M.

   (b) m has a unique minimizer if and only if $B \succ 0$.

   Proof. If p^* is a minimizer of m we have from second order conditions that $0 = \nabla m(p^*) = Bp^* + g$ and $B = \nabla^2 m(p^*) \succeq 0$. Now suppose that $g \in Im(B)$, then $\exists p$ such that $Bp = -g$. Let $w \in \mathbb{R}^n$ then:

$$\begin{aligned} m(p + w) &= g^T p + g^T w + \frac{1}{2}\left(w^T B w + 2p^T B w + p^T B p\right) \\ &= m(p) + g^T w - \frac{1}{2} 2 g^T w + \frac{1}{2} w^T B w \\ &\geq m(p) \end{aligned}$$

where the inequality follows since $B \succeq 0$. This proves the first part.

Now suppose m has a unique minimizer, p^*, and suppose $\exists w$ such that $w^T B w = 0$. Since p^* is a minimizer, we have that $Bp^* = -g$ and so from the computation above $m(p+w) = m(p)$ and so $p+w, p$ are minimizers to m which is a contradiction. For the other direction, since $B \succ 0$ and letting $p^* = -B^{-1}g$, by Second Order Sufficient Conditions, m has a unique minimizer. □

2. Characterization of Exact Solution

Theorem 3.3. *The vector p^* is a solution to the trust region problem:*

$$\min_{p \in \mathbb{R}^n} f(x) + g^T p + \frac{1}{2} p^T B p : \|p\| \leq \Delta$$

if and only if p^* is feasible and $\exists \lambda > 0$ such that:

(a) $(B + \lambda I)p^* = g$
(b) $\lambda(\|p^*\| - \Delta) = 0$
(c) $B + \lambda I \succeq 0$

Proof. First we prove ($\Leftarrow$) direction. So $\exists \lambda \geq 0$ which satisfies the properties. Consider the problem

$$m^*(p) = f(x) + g^T p + \frac{1}{2} p^T (B + \lambda I) p = m(p) + \frac{1}{2} \lambda \|p\|^2$$

Moreover, $m^*(p) \geq m^*(p^*)$ which implies that:

$$m(p) \geq m(p^*) + \lambda(\|p^*\|^2 - \|p\|^2)$$

So if $\lambda = 0$ then $m(p) \geq m(p^*)$ and $p*$ is minimizer of m, or if $\lambda > 0$ then $\|p^*\| = \Delta$ and so for a feasible p, $\|p\| \leq \Delta$, which implies $m(p) \geq m(p^*)$.

Now suppose p^* is the minimizer of m. If $\|p^*\| < \Delta$ then p^* is the unconstrained minimizer of m and so $\lambda = 0$, and so $\nabla m(p^*) = Bp^* + g = 0$ and $\nabla^2 m(p^*) = B \succeq 0$ by the previous lemma. Now suppose $\|p^*\| = \Delta$. Then the second condition is automatically satisfied. We now use the Lagrangian to determine the solution. $\mathcal{L}(\lambda, p) = g^T p + \frac{1}{2} p^T B p + \frac{\lambda}{2}(p^T p - \Delta^2)$ Differentiating with respect to p and setting this equal to 0, we see that

$$(B + \lambda I)p^* + g = 0$$

Now to show that $(B + \lambda I)$ is positive semi definite, note that for any p such that $\|p\|^2 = \Delta$ and since p^* is the minimizer of m: $m(p) + \frac{\lambda}{2}\|p\|^2 \geq m(p^*) + \frac{\lambda}{2}\|p^*\|^2$ Rearranging, we have that

$$(p - p^*)^T (B - \lambda I)(p - p^*) \geq 0$$

And since the set of all normalized $\pm(p - p^*)$ where $\|p\| = \Delta$ is dense on the unit sphere, $B - \lambda I$ is positive semi-definite. The last thing to show is that $\lambda \geq 0$ which follows using the fact that if $\lambda < 0$ then p^* must be a global minimizer of m. By the previous lemma, this is a contradiction. □

3.4.2 Newton's Root Finding Iterative Solution

1. **Theorem 3.3** guarantees that a solution exists and we need on find λ which satisfies these conditions. Two cases exist from this theorem:

Case 1 $\lambda = 0$. If $\lambda = 0$ then $B \succeq 0$ and we can compute a solution p^* using QR decomposition.

Case 2 $\lambda > 0$. First, letting $Q\Lambda Q^T$, be the EVD of B, and defining:

$$p(\lambda) = -Q(\Lambda + \lambda I)^{-1} Q^T g$$

Hence:

$$\|p(\lambda)\|^2 = g^T Q(\Lambda + \lambda I)^{-2} Q^T g = \sum_{j=1}^{n} \frac{(q_j^T g)^2}{(\lambda_j + \lambda)^2}$$

From **Theorem 3.3**, we want to find $\lambda > 0$ for which:

$$\Delta^2 = \sum_{j=1}^{n} \frac{(q_j^T g)^2}{(\lambda_j + \lambda)^2}$$

2. The second case has two sub-cases which must be considered. Let Q_1 be the columns of Q which correspond to the eigenspace of λ_1.

Case 2a If $Q_1^T g \neq 0$, then we implement Newton's Root finding algorithm on:

$$f(\lambda) = \frac{1}{\Delta} - \frac{1}{\|p(\lambda)\|}$$

since as $\lambda \to \lambda_1$, $\|p(\lambda)\| \to \infty$ and so a solution exists.

Case 2b If $Q_1^T g = 0$, then applying Newton's root finding algorithm naively is not useful since $\|p(\lambda)\| \not\to \infty$ as $\lambda \to \lambda_1$. We note that when $\lambda = -\lambda_1$, $(B + \lambda I) \succeq 0$ and so we can find a τ such that:

$$\Delta^2 = \sum_{j:\lambda_1 \neq \lambda_j} \frac{(q_j^T g)^2}{(\lambda_j - \lambda_1)^2} + \tau^2$$

Computation. $\exists z \neq 0$ such that $(B - \lambda_1 I)z = 0$ and $\|z\| = 1$. Therefore, $q_j^T z = 0$ for all $j : \lambda_1 \neq \lambda_j$. Setting

$$p(\lambda_1, \tau) = \sum_{j:\lambda_1 \neq \lambda_j} \frac{(q_j^T g)^2}{(\lambda_j - \lambda_1)^2} + \tau^2$$

we need only find τ. □

3.5 Trust Region Subproblem Algorithms

4 Conjugate Gradients

1. Conjugated Gradients Overview

Algorithm 5: Overview of Conjugate Gradient Algorithm
input : x_0 a starting point, ϵ tolerance, f objective
$x \leftarrow x_0$
Compute Gradient: $r \leftarrow \nabla f(x)$
Compute Conjugated Descent Direction: $p \leftarrow -r$
while $\|r\| > \epsilon$ **do**
  Compute Optimal Step Length: α
  Compute Step: $x \leftarrow x + \alpha r$
  Compute Gradient: $r \leftarrow \nabla f(x)$
  Compute Conjugated Step Direction: p
end
return *Minimizer: x*

2. Conjugate Gradients for minimizing convex ($A \succ 0$), quadratic function

$$f(x) = \frac{1}{2}x^T Ax - b^T x$$

Algorithm 6: Conjugate Gradients Algorithm for Convex Quadratic
input : x_0 a starting point, $\epsilon = 0$ tolerance, f objective
$x \leftarrow x_0$
Compute Gradient: $r \leftarrow \nabla f(x) = Ax - b$
Compute Conjugated Descent Direction: $p \leftarrow -r$
while $\|r\| > \epsilon$ **do**
  Compute Optimal Step Length: $\alpha = \frac{r^T r}{p^T Ap}$
  Compute Step: $x \leftarrow x + \alpha r$
  Store Previous: $r_{old} \leftarrow r$
  Compute Gradient: $r \leftarrow r + \alpha Ap$
  Compute Conjugated Step Direction: $p \leftarrow -r + \frac{\|r\|^2}{\|r_{old}\|^2} p$
end
return *Minimizer: x*

3. Fletcher-Reeves for Nonlinear Functions

Algorithm 7: Fletcher Reeves CG Algorithm
input : x_0 a starting point, ϵ tolerance, f objective
$x \leftarrow x_0$
Compute Gradient: $r \leftarrow \nabla f(x)$
Compute Conjugated Descent Direction: $p \leftarrow -r$
while $\|r\| > \epsilon$ **do**
  Compute Optimal Step Length: α (Strong Wolfe Line Search)
  Compute Step: $x \leftarrow x + \alpha r$
  Store Previous: $r_{old} \leftarrow r$
  Compute Gradient: $r \leftarrow \nabla f(x)$
  Compute Conjugated Step Direction: $p \leftarrow -r + \frac{\|r\|^2}{\|r_{old}^2\|} p$
end
return *Minimizer: x*

Part II
Model Hessian Selection

5 Newton's Method

1. Requires $\nabla^2 f \succ 0$ in some region for the method to work.
2. Line Search
   (a) The search direction:
   $$p_k^N = -\nabla^2 f_k^{-1} \nabla f_k$$
   (b) Rate of Convergence

   Theorem 5.1. *Suppose f is an objective with minimum x^* satisfying the Second Order Conditions. Moreover:*

   i. Suppose f is twice continuously differentiable in a neighborhood of x^*
   ii. Specifically, suppose $\nabla^2 f$ is Lipschitz continuous in a neighborhood of x^*
   iii. Suppose $p_k = p_k^N$ and $x_{k+1} = x_k + p_k^N$
   iv. Suppose $x_k \to x^*$

   Then for x_0 sufficiently close to x^*:

   i. $x_k \to x^*$ quadratically
   ii. $\|f_k\| \to 0$ quadratically.

   Proof. Our goal is to get $x_{k+1} - x^*$ in terms of $\nabla^2 f$ and use Lipschitz continuity to provide an upper bound.

   i. We have that:

   $$\begin{aligned}
   &x_{k+1} - x^* \\
   &= x_k + p_k - x^* \\
   &= x_k - x^* - \nabla^2 f_k^{-1} \nabla f_k \\
   &= \nabla^2 f_k^{-1} \left[\nabla^2 f_k (x_k - x^*) - (\nabla f_k - \nabla f^*)\right] \\
   &= \nabla^2 f_k^{-1} \left[\nabla^2 f_k (x_k - x^*) - \int_0^1 \nabla^2 f(x_k + t(x - x^*))(x - x^*) dt\right] \\
   &= \nabla^2 f_k^{-1} \left[\int_0^1 \left(\nabla^2 f_k - \nabla^2 f(x_k + t(x_k - x^*))\right)(x - x^*) dt\right]
   \end{aligned}$$

ii. Using Lipschitz continuity:

$$\begin{aligned}
&\|x_{k+1} - x^*\| \\
&\leq \left\|\nabla^2 f_k^{-1}\right\| \int_0^1 \left\|\nabla^2 f_k - \nabla^2 f(x_k + t(x_k - x^*))\right\| \|x - x^*\| \, dt \\
&\leq \left\|\nabla^2 f_k^{-1}\right\| \int_0^1 tL \|x_k - x^*\|^2 \, dt \\
&\leq \left\|\nabla^2 f_k^{-1}\right\| \frac{L}{2} \|x_k - x^*\|^2
\end{aligned}$$

We now need to show that $\left\|\nabla^2 f_k^{-1}\right\|$ is bounded from above. Let $\eta(2r)$ be a neighborhood of radius $2r$ about x^* for which $\nabla^2 f \succ 0$. In this neighborhood, $\nabla^2 f^{-1}$ exists and $g(x) = \left\|\nabla^2 f(x)^{-1}\right\|$ is continuous. Then, on the compact closed ball $\overline{\eta(r)}$, $g(x)$ has a finite maximum. Therefore, we have quadratic convergence of $x_k \to x^*$.

iii. We now consider the first derivative:

$$\begin{aligned}
&\|\nabla f_{k+1} - \nabla f^*\| = \|\nabla f_{k+1}\| \\
&= \left\|\nabla f_{k+1} - \nabla f_k - \nabla^2 f_k p_k\right\| \\
&= \left\| \int_0^1 \left[\nabla^2 (x_k + t p_k) - \nabla^2 f_k\right] p_k dt \right\| \\
&\leq \left\| \int_0^1 Lt \|p_k\|^2 \right\| \\
&\leq \frac{L}{2} g(x_k)^2 \|\nabla f_k\|^2
\end{aligned}$$

So if $x_0 \in \eta(r)$, the result follows. □

3. Trust Region

(a) Asymptotically Similar

Definition 5.1. *A sequence p_k is asymptotically similar to a sequence q_k if*

$$\|p_k - q_k\| = o(\|q_k\|)$$

(b) The Subproblem

$$\min_{p \in \mathbb{R}^n} f(x) + \nabla f_x^T p + \frac{1}{2} p^T \nabla^2 f_x p : \|p\| \leq \Delta$$

(c) Rate of Convergence

Theorem 5.2. *Suppose f is an objective with minimum x^* satisfying the Second Order Conditions. Moreover:*

i. Suppose f is twice continuously differentiable in a neighborhood of x^*

ii. Specifically, suppose $\nabla^2 f$ is Lipschitz continuous in a neighborhood of x^*

iii. Suppose for some $c \in (0,1)$, p_k satisfy

$$m_k(p_k) - m_k(0) \leq -c \|\nabla f_k\| \min\left(\Delta_k, \frac{\|\nabla f_k\|}{\|\nabla^2 f_k\|}\right)$$

and are asymptotically similar to p_k^N when $\left\|p_k^N\right\| \leq \frac{1}{2}\Delta_k$.

iv. Suppose $x_{k+1} = x_k + p_k$ and $x_k \to x^*$

Then:

i. Δ_k becomes inactive for sufficiently large k.

ii. $x_k \to x^*$ superlinearly.

Proof. We need to show that Δ_k are bounded from below. The second part follows from this quickly since $x_k \to x^*$ implies $\left\|p_k^N\right\| \to 0$ and eventually $\left\|p_k^N\right\| \leq \frac{1}{2}\Delta_k$. So applying the asymptotic similarity:

$$\begin{aligned}
\|x_k + p_k - x^*\| &\leq \left\|x_k + p_k^N - x^*\right\| + \left\|p_k^N - p_k\right\| \\
&\leq o(\|x_k - x^*\|^2) + o(\left\|p_k^N\right\|) \\
&\leq o(\|x_k - x^*\|^2) + o(\|x - x^*\|) \\
&\leq o(\|x - x^*\|)
\end{aligned}$$

where the penultimate inequality follows from:

$$\left\|\nabla^2 f_k^{-1}\right\| \|\nabla f_k - \nabla f^*\| \leq \left\|\nabla^2 f_k^{-1}\right\| \|x_k - x^*\| \left(\sup_{x \in \eta(r)} \left\|\nabla^2 f(x)\right\|\right)$$

To get that Δ_k is bounded from below, we frame it in terms of ρ_k. The numerator of $\rho_k - 1$ satisfies:

$$\begin{aligned}
&|f(x_k + p_k) - f(x_k) - m_k(p_k) - m_k(0)| \\
&\quad \leq \left\| \frac{1}{2} p_k^T \int_0^1 \left(\nabla^2 f(x + tsp_k) - \nabla^2 f_k\right) p_k dt \right\| \\
&\quad \leq \frac{L}{4} \|p_k\|^3
\end{aligned}$$

And the denominator of $\rho_k - 1$ satisfies:

$$\begin{aligned}
|m_k(p_k) - m_k(0)| &\geq c \|\nabla f_k\| \min\left(\Delta_k, \frac{\|\nabla f_k\|}{\|\nabla^2 f_k\|}\right) \\
&\geq c \|\nabla f_k\| \min\left(\|p_k\|, \frac{\|\nabla f_k\|}{\|\nabla^2 f_k\|}\right)
\end{aligned}$$

So we need only lower bound $\|f_k\|$ by $\|p_k\|$ to translate this to a lower bound for Δ_k. If $\left\|p_k^N\right\| > \frac{1}{2}\Delta_k$ then (in some neighborhood of x^*)

$$\|p_k\| \leq \Delta_k \leq 2\left\|p_k^N\right\| \leq 2\left\|\nabla^2 f_k^{-1}\right\| \|\nabla f_k\| \leq C \|\nabla f_k\|$$

Else if $\left\|p_k^N\right\| \leq \frac{1}{2}\Delta_k$ then

$$\|p_k\| \leq \left\|p_k^N\right\| + o(\left\|p_k^N\right\|) \leq 2\left\|p_k^N\right\| \leq C\|\nabla f_k\|$$

Therefore:

$$|m_k(p_k) - m_k(0)| \geq \frac{c\|p_k\|}{C} \min\left(\|p_k\|, \frac{\|p_k\|}{\|\nabla^2 f_k\|\left\|\nabla^2 f_k^{-1}\right\|}\right)$$

Since the second term's denominator is the condition number of the matrix (which must be greater than or equal to 1), we have that:

$$|\rho_k - 1| \leq C'\|p_k\| \leq C'\Delta_k$$

Hence, if $\Delta_k < \frac{3}{4C'}$, $\rho_k > 3/4$ and Δ_k would double. So we know that $\Delta_k \geq \frac{1}{2^M}\hat{\Delta}$ for some fixed M such that $\frac{1}{2^M}\hat{\Delta} \leq \frac{3}{4C'}$. So we have that for sufficiently large k, $p_k \to 0$ and so $\rho_k \to 1$ and Δ_k becomes inactive. □

6 Newton's Method with Hessian Modification

1. If $\nabla^2 f_k \not\succ 0$, we can add a matrix E_k so that $\nabla^2 f_k + E_k =: B_k \succ 0$. As long as this model produces a descent direction, and $\|B_k\| \left\|B_k^{-1}\right\|$ are bounded uniformly, Zoutendjik guarantees convergence.

2. Newton's Method with Hessian Modification

Algorithm 8: Line Search with Modified Hessian
input : Starting point x_0, Tolerance ϵ
$k \leftarrow 0$ **while** $\|p_k\| > \epsilon$ **do**
  Find E_k so that $\nabla^2 f_k + E_k \succ 0$
  $p_k \leftarrow$ Solve $B_k p = -g_k$
  Compute α_k (Wolfe, Goldstein, Backtracking, Interpolation)
  $x_{k+1} \leftarrow x_k + \alpha_k p_k$ $k \leftarrow k+1$
end

3. Minimum Frobenius Norm
   (a) Problem: find E which satisfies $\min \|E\|_F^2 : \lambda_{\min}(A+E) \geq \delta$.
   (b) Solution: If $A = Q\Lambda Q^T$ is the EVD of A then $E = Q diag(\tau_i) Q^T$ where
   $$\tau_i = \begin{cases} 0 & \lambda_i \geq \delta \\ \delta - \lambda_i & \lambda_i < \delta \end{cases}$$

4. Minimum 2-Norm
   (a) Problem: find E which satisfies $\min \|E\|_2^2 : \lambda_{\min}(A+E) \geq \delta$
   (b) Solution: If $\lambda_{\min}(A) = \lambda_n$ then $E = \tau I$ where:
   $$\tau = \begin{cases} 0 & \lambda_n \geq \delta \\ \delta - \lambda_n & \lambda_n < \delta \end{cases}$$

5. Modified Cholesky
   (a) Compute LDL Factorization but add constants to ensure positive definiteness of A
   (b) May require pivoting to ensure numerical stability

6. Modified Block Cholesky
   (a) Compute LBL factorization where B is a block diagonal (1×1 and $2\times$ where the number of blocks is the number of positive eigenvalues and the number of 2×2 blocks is the number of negative eigenvalues)
   (b) Compute the EVD of B and modify it to ensure appropriate eigenvalues.

7 Quasi-Newton's Method

1. Fundamental Idea

   Computation. Let f be our objective and suppose it is Lipschitz twice continuously differentiable. From Taylor's Theorem:

$$\begin{aligned}\nabla f(x+p) &= \nabla f(x) + \int_0^1 \nabla^2 f(x+tp)p\,dt \\ &= \nabla f(x) + \nabla^2 f(x)p + \int_0^1 \left(\nabla^2 f(x+tp) - \nabla^2 f(x)\right) p\,dt \\ &= \nabla f(x) + \nabla^2 f(x)p + o(\|p\|)\end{aligned}$$

   Therefore, letting $x = x_k$ and $p = x_{k+1} - x_k$

$$\nabla^2 f_k(x_{k+1} - x_k) \approx \nabla f_{k+1} - \nabla f_k$$

   Quasi-Newton methods choose an approximation to $\nabla^2 f_k$, B_k which satisfies:

$$B_{k+1}(x_{k+1} - x_k) = g_{k+1} - g_k$$

   □

2. Secant Equation

   Definition 7.1. *Let x_i be a sequence of iterates when minimizing an objective function f. Let $s_k = x_{k+1} - x_k$ and $y_k = g_{k+1} - g_k$ where $g_i = \nabla f_i$. Then the secant equation is*

$$B_{k+1}s_k = y_k$$

   where B_{k+1} is some matrix.

3. Curvature Condition

   Definition 7.2. *The curvature condition is $s_k^T y_k > 0$.*

   Note 7.1. *If B_k satisfies the secant equation and s_k, y_k satisfy the curvature condition then $s_k^T B_{k+1} s_k > 0$. So the quadratic along the step direction is convex.*

4. Line Search Rate of Convergence

   Theorem 7.1. *Suppose f is an objective with minimum x^* satisfying the Second Order Conditions. Moreover:*

   (a) Suppose f is twice continuously differentiable in a neighborhood of x^*

   (b) Suppose B_k is computed using a quasi-Newton method and

$$B_k p_k = -g_k$$

   (c) Suppose $x_{k+1} = x_k + p_k$

(d) Suppose $x_k \to x^*$

When x_0 is sufficiently close to x^*, $x_k \to x^*$ superlinearly if and only if:

$$\lim_{k\to\infty} \frac{\left\|(B_k - \nabla^2 f(x^*))p_k\right\|}{\|p_k\|} \to 0$$

Proof. We show that the condition is equivalent to: $p_k - p_k^N = o(\|p_k\|)$. First, if the condition holds:

$$\begin{aligned} p_k - p_k^N &= \nabla^2 f_k^{-1}\left(\nabla^2 f_k p_k + \nabla f_k\right) \\ &= \nabla^2 f_k^{-1}\left(\nabla^2 f_k - B_k\right) p_k \\ &= \nabla^2 f_k^{-1}\left[\left(\nabla^2 f_k - \nabla^2 f(x^*)\right) p_k + \left(\nabla^2 f(x^*) - B_k\right) p_k\right] \end{aligned}$$

Taking norms and using that x_0 is sufficiently close to x^*, and continuity:

$$\left\|p_k - p_k^N\right\| = o(\|p_k\|)$$

For the other direction:

$$o(\|p_k\|) = \left\|p_k - p_k^N\right\| \geq \left\|\left(\nabla^2 f(x^*) - B_k\right) p_k\right\|$$

Now we just apply triangle inequality and properties of the Newton Step:

$$\|x_k + p_k - x^*\| \leq \left\|x_k + p_k^N - x^*\right\| + \left\|p_k - p_k^N\right\| = o(\|x_k - x^*\|^2) + o(\|p_k\|)$$

Note that $o(\|p_k\|) = o(\|x_k - x^*\|)$. □

7.1 Rank-2 Update: DFP & BFGS

1. Davidson-Fletcher-Powell Update

   (a) Let:

   $$G_k = \int_0^1 \nabla^2 f(x_k + t\alpha p_k)dt$$

   So Taylor's Theorem implies:

   $$y_k = G_k s_k$$

   (b) Let:

   $$\|A\|_W = \left\|G_k^{-1/2} A G_k^{-1/2}\right\|_F$$

   (c) Problem. Given B_k symmetric positive definite, s_k and y_k, find

   $$\arg\min \|B - B_k\|_W : B = B^T,\ Bs_k = y_k$$

   (d) Solution:

   $$B_{k+1} = \left(I - \frac{y_k s_k^T}{y_k^T s_k}\right) B_k \left(I - \frac{s_k y_k^T}{y_k^T s_k}\right) + \frac{y_k y_k^T}{y_k^T s_k}$$

(e) Inverse Hessian:

$$H_{k+1} = H_k - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} + \frac{s_k s_k^T}{y_k^T s_k}$$

2. Broyden, Fletcher, Goldfarb, Shanno Update

(a) Problem: Given H_k (inverse of B_k) symmetric positive definite, s_k and y_k, find

$$\arg\min \|H - H_k\|_W : H = H^T,\ Hy_k = s_k$$

(b) Solution:

$$H_{k+1} = \left(I - \frac{s_k y_k^T}{y_k^T s_k}\right) H_k \left(I - \frac{y_k s_k^T}{y_k^T s_k}\right) + \frac{s_k s_k^T}{y_k^T s_k}$$

(c) Hessian:

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}$$

3. Preservation of Positive Definiteness

Lemma 7.1. *If $B_k \succ 0$ and B_{k+1} is updated from the BFGS method, then $B_{k+1} \succ 0$.*

Proof. We can equivalently show that $H_{k+1} \succ 0$. Let $z \in \mathbb{R}^n \setminus \{0\}$. Then:

$$z^T H_{k+1} z = w^T H_k w + \rho_k (s_k^T z)^2$$

where $w = z - \rho_k (s_k^T z) y_k$. If $s_k^T z = 0$ then $w = z$ and $z^T H_k z > 0$. If $s_k^T z \neq 0$, then $\rho_k (s_k^T z)^2 > 0$ (regardless of if $w = 0$) and so $H_{k+1} \succ 0$. □

7.2 Rank-1 Update: SR1

1. Problem: find v such that:

$$B_{k+1} = B_k + \sigma v v^T : B_{k+1} s_k = y_k$$

2. Solution:

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^T}{s_k^T (y_k - B_k s_k)}$$

Proof. Multiplying by s_k:

$$y_k - B_k s_k = \sigma (v^T s_k) v$$

Hence, v is a multiple of $y_k - B_k s_k$. □

Part III
Specialized Applications of Optimization

8 Inexact Newton's Method

1. Requirements for search direction:
   (a) p_k is inexpensive to compute
   (b) p_k approximates the newton step closely as measured by:
   $$z_k = \nabla^2 f_k p_k - \nabla f_k$$
   (c) The error is controlled by the current gradient:
   $$\|z_k\| \leq \eta_k \|\nabla f_k\|$$
2. Typically, inexact methods keep $B_k = \nabla^2 f_k$ and improve on computing b_k
3. CG Based Algorithmic Overview

Algorithm 9: Overview of Inexact CG Algorithm

```
input : x_0 a starting point, f objective, restrictions
x ← x_0
Define Tolerance: ε Compute Gradient: r ← ∇f(x)
Compute Conjugated Descent Direction: p ← −r
while ‖r‖ > ε do
    Check Constrained Polynomial
    if p^T Bp ≤ 0 then
        Compute α given restrictions
        Compute Minimizer: x ← x + αp
        Break Loop
    else
        Compute Optimal Step Length: α
        if Restrictions on x then
            Compute τ given restrictions
            Compute Minimizer: x ← x + ταp
            Break Loop
        else
            Compute Step: x ← x + αr
        end
        Compute Gradient: r ← ∇f(x)
        Compute Conjugated Step Direction: p
    end
end
x* ← x return Minimizer: x*
```

4. As with the usual CG, we can implement methods for computing the gradient and conjugate step direction efficiently

5. Line Search Strategy: the restriction of α is simply that it satisfies the strong Wolfe Conditions

6. Trust-Region Strategy:

   (a) The restriction on α is that $\|x + \alpha p\| = \Delta$ because the restricted polynomial in the direction p is decreasing, so we should go all the way to the boundary

   (b) The restriction on x is that $\|x + \alpha p\| > \Delta$ because $x + \alpha p$ is no longer in the trust region so we need to find τ that ensures this happens.

7. An alternative is to use Lanczos' Method, which generalizes CG Methods.

9 Limited Memory BFGS

1. In normal BFGS:

$$H_{k+1} = (I - \rho_k s_k y_k^T) H_k (I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T$$

2. Substituting in for H_k until H_0 reveals that only s_k, y_k must be stored to compute H_{k+1}

3. Limited BFGS capitalizes on this by storing a fixed number of s_k, y_k and re-updating H_k from a (typically diagonal) matrix H_0^k

Algorithm 10: L-BFGS Algorithm

input : x_0 a starting point, f objective, m storage size, ϵ tolerance

$x \leftarrow x_0$

$k \leftarrow 0$

while $\|\nabla f(x)\| > \epsilon$ **do**

  Choose H_0

  Update H from $\{(s_k, y_k), \ldots, (s_{k-m}, y_{k-m})\}$

  Compute $p^u \leftarrow -H \nabla f(x)$

  Implement Line Search or Trust Region (dogleg) to find p

  $x \leftarrow x + p$

  $k \leftarrow k + 1$

end

return *Minimizer:* x

10 Least Squares Regression Methods

1. In least squares problems, the objective function has a specialized form, for example:
$$f(x) = \sum_{i=1}^{m} r_j(x)^2$$
   (a) $x \in \mathbb{R}^n$ and usually are the parameters of a particular model of interest
   (b) r_j are usually the residuals evaluated for the model ϕ with parameter x at point t_j and output y_j. That is:
$$r_j(x) = \phi(x, t_j) - y_j$$

2. Formulation
   (a) Let $r(x) = \begin{bmatrix} r_1(x) & r_2(x) & \cdots & r_m(x) \end{bmatrix}$. Then:
$$f(x) = \frac{1}{2} \|r(x)\|^2$$
   (b) The first derivative:
$$\nabla f(x) = \nabla r(x)^T r(x) = \begin{bmatrix} \nabla r_1(x)^T \\ \nabla r_2(x)^T \\ \vdots \\ \nabla r_m(x)^T \end{bmatrix} r(x) =: J(x)^T r(x)$$
   (c) The Second Derivative:
$$\nabla^2 f(x) = \nabla r(x)^T \nabla r(x) + \nabla^2 r(x) r(x) = J(x)^T J(x) + \nabla^2 r(x) r(x)$$

3. The main conceit behind Least Squares methods is to approximate $\nabla^2 f(x)$ by $J(x)^T J(x)$, thus, only first derivative information is required.

10.1 Linear Least Squares Regression

1. Suppose $\phi(x, t_j) = t_j^T x$ then:
$$r(x) = Tx - y$$
   where each row of T is t_j^T

2. Objective Function:
$$f(x) = \frac{1}{2} \|Tx - y\|^2$$
   Note 10.1. *The objective function can be solved directly using matrix factorizations if m is not too large.*

3. First Derivative:
$$\nabla f(x) = T^T T x - T^y$$
   Note 10.2. *The normal equations may be better to solve if m is large because by multiplying by the transpose, we need only solve an $n \times n$ system of equations.*

10.2 Line Search: Gauss-Newton

1. The Gauss-Newton Algorithm is a line search for non-linear least squares regression with search direction:

$$J_k^T J_k p_k = -J_k^T r_k$$

2. Notice that these are the normal equations from:

$$\frac{1}{2} \|J_k p_k + r_k\|^2$$

3. Using this search direction, we proceed with line search as per usual

10.3 Trust Region: Levenberg-Marquardt

1. The local model is:

$$m(p) = f(x) + (J(x)^T r(x))^T p + \frac{1}{2} p^T J(x)^T J(x) p \quad \|p\| \leq \Delta$$

2. Since $f(x) = \|r(x)\|^2 / 2$, the solution to minimizing $m(p)$ is equivalent to:

$$\min \frac{1}{2} \|J(x)p + r\|^2 \quad \|p\| \leq \Delta$$

Part IV
Constrained Optimization

10.4 Theory of Constrained Optimization

1. In constrained optimization, the problem is:

$$\min\{x \in \arg\min f(x) : c_i(x) = 0 \ \forall i \in \mathcal{E},\ c_j(x) \geq 0 \ \forall j \in \mathcal{I}\}$$

   (a) f is the usual objective function

   (b) $\mathcal{E}$ is the index for equality constraints

   (c) $\mathcal{I}$ is the index for inequality constraints

2. Feasible Set. Active Set.

   Definition 10.1. *The feasible set, Ω, is the set of all points which satisfy the constraints:*

$$\Omega = \{x : c_i(x) = 0 \ \forall i \in \mathcal{E},\ c_j(x) \geq 0 \ \forall j \in \mathcal{I}\}$$

   The active set at point $x \in \Omega$, $\mathcal{A}(x)$, is defined by:

$$\mathcal{A}(x) = \mathcal{E} \cup \{j \in \mathcal{I} : c_j(x) = 0\}$$

3. Feasible Sequence. Tangent. Tangent Cone.

   Definition 10.2. *Let $x \in \Omega$. A sequence $z_k \to x$ is a feasible sequence if for sufficiently large K, if $k \geq K$ then $z_k \in \Omega$. A vector d is a tangent to x if there is a feasible sequence $z_k \to x$ and a sequence $t_k \to 0$ such that:*

$$\lim_{k\to\infty} \frac{z_k - x}{t_k} = d$$

   The set of all tangents is the Tangent Cone at x, denoted $T_\Omega(x)$.

   Note 10.3. *$T_\Omega(x)$ contains all step directions which for a sufficiently small step size will remain in the feasible region Ω.*

4. Linearized Feasible Directions.

   Definition 10.3. *Let $x \in \Omega$ and $\mathcal{A}(x)$ be its active set. The set of linearized feasible directions, $\mathcal{F}(x)$ is defined as:*

$$\mathcal{F}(x) = \{d : d^T \nabla c_i(x) = 0 \ \forall i \in \mathcal{E},\ d^T \nabla c_i(x) \geq 0 \ \forall \mathcal{A}(x) \cap \mathcal{I}\}$$

   Note 10.4. *Let $x \in \Omega$ and suppose we linearize the constraints near x. The set $\mathcal{F}(x)$ contains all search directions for which the linearized approximations to $c_i(x+d)$ satisfy the constraints. For example If $c_i(x) = 0$ for $i \in \mathcal{E}$, then $0c_i(x+d) \approx c_i(x) + d^T \nabla c_i(x) = d^T \nabla c_i(x)$ for d to be a good search direction.*

5. LICQ Constrain Qualification

Definition 10.4. *Given a point x and active set $\mathcal{A}(x)$, the linear independence constrain qualification (LICQ) holds if the set of active constraint gradients, $\{\nabla c_i(x) : i \in \mathcal{A}(x)\}$, is linearly independent.*

Note 10.5. *Most algorithms use line search or trust region require that $\mathcal{F}(x) \approx T_\Omega(x)$, else there is not a good way to find another iterate.*

6. First Order Necessary Conditions

Theorem 10.1. *Let f, c_i be continuously differentiable. Suppose x^* is a local solution to the optimization problem and at x^*, LICQ holds. Then there is a Lagrange multiplier vector λ, with components λ_i for $i \in \mathcal{E} \cup \mathcal{I}$, such that the Karush-Kuhn-Tucker conditions are satisfied at (x^*, λ):*

(a) $\nabla_x \mathcal{L}(x^*, \lambda) = 0$

(b) For all $i \in \mathcal{E}$, $c_i(x^*) = 0$

(c) For all $j \in \mathcal{I}$, $c_j(x^*) \geq 0$

(d) For all $j \in \mathcal{I}$, $\lambda_j \geq 0$

(e) For all $j \in \mathcal{I}$, $\lambda_j c_j(x^*) = 0$