

MATH378 - Nonlinear Optimization

Based on lectures from Fall 2025 by Prof. Tim Hoheisel.

Notes by Louis Meunier

Contents

| | |
|---|----|
| 1 Preliminaries | 2 |
| 1.1 Terminology | 2 |
| 1.2 Convex Sets and Functions | 3 |
| 2 Unconstrained Optimization | 4 |
| 2.1 Theoretical Foundations | 4 |
| 2.1.1 Quadratic Approximation | 6 |
| 2.2 Differentiable Convex Functions | 7 |
| 2.3 Matrix Norms | 9 |
| 3 Descent Methods | 11 |
| 3.1 A General Line-Search Method | 11 |
| 3.1.1 Global Convergence of Algorithm 3.1 | 12 |
| 3.2 The Gradient Method | 13 |
| 3.3 Newton-Type Methods | 15 |
| 3.3.1 Convergence Rates and Landau Notation | 15 |
| 3.3.2 Newton's Method for Nonlinear Equations | 16 |
| 3.3.3 Newton's Method for Optimization Problem | 18 |
| 3.4 Quasi-Newton Methods | 21 |
| 3.4.1 Direct Methods | 22 |
| 3.4.2 Inexact Methods | 26 |
| 3.5 Conjugate Gradient Methods for Nonlinear Optimization | 27 |
| 3.5.1 Prelude: Linear Systems | 27 |

§1 PRELIMINARIES

§1.1 Terminology

We consider problems of the form

$$\text{minimize } f(x) \text{ subject to } x \in X, \quad (\dagger)$$

with $X \subset \mathbb{R}^n$ the *feasible region* with x a *feasible point*, and $f : X \rightarrow \mathbb{R}$ the *objective (function)*; more concisely we simply write

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

When $X = \mathbb{R}^n$, we say the problem (\dagger) is *unconstrained*, and conversely *constrained* when $X \subsetneq \mathbb{R}^n$.

⊗ **Example 1.1** (Polynomial Fit): Given $y_1, \dots, y_m \in \mathbb{R}$ measurements taken at m distinct points $x_1, \dots, x_m \in \mathbb{R}$, the goal is to find a degree $\leq n$ polynomial $q : \mathbb{R} \rightarrow \mathbb{R}$, of the form

$$q(x) = \sum_{k=0}^n \beta_k x^k,$$

“fitting” the data $\{(x_i, y_i)\}_i$, in the sense that $q(x_i) \approx y_i$ for each i . In the form of (\dagger) , we can write this precisely as

$$\min_{\beta \in \mathbb{R}^{n+1}} \frac{1}{2} \sum_{i=0}^n \left(\underbrace{\beta_n x_i^n + \dots + \beta_1 x_i + \beta_0}_{q(x_i)} - y_i \right)^2;$$

namely, we seek to minimize the ℓ^2 -distance between $(q(x_i))$ and (y_i) . If we write

$$X := \begin{pmatrix} 1 & x_1 & \dots & x_1^n \\ \vdots & \dots & \dots & \vdots \\ 1 & x_m & \dots & x_m^n \end{pmatrix} \in \mathbb{R}^{m \times (n+1)}, \quad y := \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} \in \mathbb{R}^m,$$

then concisely this problem is equivalent to

$$\min_{\beta \in \mathbb{R}^{n+1}} \frac{1}{2} \|X \cdot \beta - y\|_2^2,$$

a so-called *least-squares problem*.

We have two related tasks:

1. Find the optimal value asked for by (\dagger) , that is what $\inf_X f$ is;
2. Find a specific point \bar{x} such that $f(\bar{x}) = \inf_X f$, i.e. the value of a point

$$\bar{x} \in \operatorname{argmin}_X f := \left\{ x \in X \mid f(x) = \inf_X f \right\}.$$

(noting that argmin should be viewed as a set-valued function, as there may be multiple admissible minimizers) Notice that if we can accomplish 2., we’ve accomplished 1. by computing $f(\bar{x})$.

Note that $\bar{x} \in \operatorname{argmin}_X f \Rightarrow f(\bar{x}) = \inf_X f$, but $\inf_X f \in \mathbb{R}$ does *not* necessarily imply $\operatorname{argmin}_X f \neq \emptyset$, that is, there needn't be a feasible minimum; for instance $\inf_{x \in \mathbb{R}} e^x = 0$, but $\operatorname{argmin}_{\mathbb{R}} f = \emptyset$ (there is no x for which $e^x = 0$).

- ↪ **Definition 1.1** (Minimizers): Let $X \subset \mathbb{R}^n$ and $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Then $\bar{x} \in X$ is called a
- *global minimizer* (of f over X) if $f(\bar{x}) \leq f(x) \forall x \in X$, or equivalently if $\bar{x} \in \operatorname{argmin}_X f$;
 - *local minimizer* (of f over X) if $f(\bar{x}) \leq f(x) \forall x \in X \cap B_\varepsilon(\bar{x})$ for some $\varepsilon > 0$.

In addition, we have *strict* versions of each by replacing " \leq " with " $<$ ".

↪ **Definition 1.2** (Some Geometric Tools): Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$.

- $\operatorname{gph} f := \{(x, f(x)) \mid x \in \mathbb{R}^n\} \subseteq \mathbb{R}^n \times \mathbb{R}$
- $f^{-1}(\{c\}) := \{x \mid f(x) = c\} \equiv \text{contour/level set at } c$
- $\operatorname{lev}_c f := f^{-1}((-\infty, c]) = \{x \mid f(x) \leq c\} \equiv \text{lower level/sublevel set at } c$

Remark 1.1:

- $\operatorname{lev}_{\inf f} f = \operatorname{argmin} f$
- assume f continuous; then all (sub)level sets are closed (possibly empty)

We recall the following result from calculus/analysis:

↪ **Theorem 1.1** (Weierstrass): Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuous and $X \subset \mathbb{R}^n$ compact. Then, $\operatorname{argmin}_X f \neq \emptyset$.

From, we immediately have the following:

↪ **Proposition 1.1:** Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ continuous. If there exists a $c \in \mathbb{R}$ such that $\operatorname{lev}_c f$ is nonempty and bounded, then $\operatorname{argmin}_{\mathbb{R}^n} f \neq \emptyset$.

PROOF. Since f continuous, $\operatorname{lev}_c f$ is closed (being the inverse image of a closed set), thus $\operatorname{lev}_c f$ is compact (and in particular nonempty). By Weierstrass, f takes a minimum over $\operatorname{lev}_c f$, namely there is $\bar{x} \in \operatorname{lev}_c f$ with $f(\bar{x}) \leq f(x) \leq c$ for each $x \in \operatorname{lev}_c f$. Also, $f(x) > c$ for each $x \notin \operatorname{lev}_c f$ (by virtue of being a level set), and thus $f(\bar{x}) \leq f(x)$ for each $x \in \mathbb{R}^n$. Thus, \bar{x} is a global minimizer and so the theorem follows. ■

§1.2 Convex Sets and Functions

↪ **Definition 1.3** (Convex Sets): $C \subset \mathbb{R}^n$ is *convex* if for any $x, y \in C$ and $\lambda \in (0, 1)$, $\lambda x + (1 - \lambda)y \in C$; that is, the entire line between x and y remains in C .

↪ **Definition 1.4** (Convex Functions): Let $C \subset \mathbb{R}^n$ be convex. Then, $f : C \rightarrow \mathbb{R}$ is called

1. *convex (on C)* if

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y),$$

for every $x, y \in C$ and $\lambda \in (0, 1)$;

2. *strictly convex (on C)* if the inequality \leq is replaced with $<$;

3. *strongly convex (on C)* if there exists a $\mu > 0$ such that

$$f(\lambda x + (1 - \lambda)y) + \mu\lambda(1 - \lambda)\|x - y\|^2 \leq \lambda f(x) + (1 - \lambda)f(y),$$

for every $x, y \in C$ and $\lambda \in (0, 1)$; we call μ the *modulus of strong convexity*.

Remark 1.2: 3. \Rightarrow 2. \Rightarrow 1.

Remark 1.3: A function is convex iff its epigraph is a convex set.

⊗ **Example 1.2:** $\exp : \mathbb{R} \rightarrow \mathbb{R}$, $\log : (0, \infty) \rightarrow \mathbb{R}$ are convex. A function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ of the form $f(x) = Ax - b$ for $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ is called *affine linear*. For $m = 1$, every affine linear function is convex. All norms on \mathbb{R}^n are convex.

↪ **Proposition 1.2:**

1. (*Positive combinations*) Let f_i be convex on \mathbb{R}^n and $\lambda_i > 0$ scalars for $i = 1, \dots, m$, then $\sum_{i=1}^m \lambda_i f_i$ is convex; as long as one is strictly (resp. strongly) convex, the sum is strictly (resp. strongly) convex as well.
2. (*Composition with affine mappings*) Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be convex and $G : \mathbb{R}^m \rightarrow \mathbb{R}^n$ be affine. Then, $f \circ G$ is convex on \mathbb{R}^m .

§2 UNCONSTRAINED OPTIMIZATION

§2.1 Theoretical Foundations

We focus on the problem

$$\min_{x \in \mathbb{R}^n} f(x),$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable.

↪ **Definition 2.1** (Directional derivative): Let $D \subset \mathbb{R}^n$ be open and $f : D \rightarrow \mathbb{R}$. We say f *directionally differentiable* at $\bar{x} \in D$ in the direction $d \in \mathbb{R}^n$ if

$$\lim_{t \rightarrow 0^+} \frac{f(\bar{x} + td) - f(\bar{x})}{t}$$

exists, in which case we denote the limit by $f'(\bar{x}; d)$.

↪ **Lemma 2.1:** Let $D \subset \mathbb{R}^n$ be open and $f : D \rightarrow \mathbb{R}$ differentiable at $x \in D$. Then, f is directionally differentiable at x in every direction d , with

$$f'(x; d) = \nabla f(x)^T d = \langle \nabla f(x), d \rangle.$$

⊗ **Example 2.1** (Directional derivatives of the Euclidean norm): Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ by $f(x) = \|x\|$ the usual Euclidean norm. Then, we claim

$$f'(x; d) = \begin{cases} \frac{x^T d}{\|x\|} & x \neq 0 \\ \|d\| & x = 0 \end{cases}.$$

For $x \neq 0$, this follows from the previous lemma and the calculation $\nabla f(x) = \frac{x}{\|x\|}$. For $x = 0$, we look at the limit

$$\lim_{t \rightarrow 0^+} \frac{f(0 + td) - f(0)}{t} = \lim_{t \rightarrow 0^+} \frac{t\|d\| - 0}{t} = \|d\|,$$

using homogeneity of the norm.

↪ **Lemma 2.2** (Basic Optimality Condition): Let $X \subset \mathbb{R}^n$ be open and $f : X \rightarrow \mathbb{R}$. If \bar{x} is a *local minimizer* of f over X and f is directionally differentiable at \bar{x} , then $f'(\bar{x}; d) \geq 0$ for all $d \in \mathbb{R}^n$.

PROOF. Assume otherwise, that there is a direction $d \in \mathbb{R}^n$ for which the $f'(\bar{x}; d) < 0$, i.e.

$$\lim_{t \rightarrow 0^+} \frac{f(\bar{x} + td) - f(\bar{x})}{t} < 0.$$

Then, for all sufficiently small $t > 0$, we must have

$$f(\bar{x} + td) < f(\bar{x}).$$

Moreover, since X open, then for t even smaller (if necessary), $\bar{x} + td$ remains in X , thus \bar{x} cannot be a local minimizer. ■

↪ **Theorem 2.1** (Fermat's Rule): In addition to the assumptions of the previous lemma, assume further that f is differentiable at \bar{x} . Then, $\nabla f(\bar{x}) = 0$.

PROOF. From the previous, we know $0 \leq f'(\bar{x}; d)$ for any d . Take $d = -\nabla f(\bar{x})$, then using the representation of a directional derivative for a differentiable function, and the fact that norms are nonnegative,

$$0 \leq -\|\nabla f(\bar{x})\|^2 \leq 0,$$

which can only hold if $\|\nabla f(\bar{x})\| = 0$ hence $\nabla f(\bar{x}) = 0$ ■

We recall the following from Calculus:

↪ **Theorem 2.2** (Taylor's, Second Order): Let $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ be twice continuously differentiable, then for each $x, y \in D$, there is an η lying on the line between x and y such that

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} (y - x)^T \nabla^2 f(\eta) (y - x).$$

↪ **Theorem 2.3** (2nd-order Optimality Conditions): Let $X \subseteq \mathbb{R}^n$ open and $f : X \rightarrow \mathbb{R}$ twice continuously differentiable. Then, if x a local minimizer of f over X , then the Hessian matrix $\nabla^2 f(x)$ is positive semi-definite.

PROOF. Suppose not, then there exists a d such that $d^T \nabla^2 f(x) d < 0$. By Taylor's, for every $t > 0$, there is an η_t on the line between x and $x + td$ such that

$$\begin{aligned} f(x + td) &= f(x) + \underbrace{t \nabla f(x)^T d}_{=0} + \frac{1}{2} t^2 d^T \nabla^2 f(\eta_t) d \\ &= f(x) + \frac{t^2}{d^T} \nabla^2 f(\eta_t) d. \end{aligned}$$

As $t \rightarrow 0^+$, $\nabla^2 f(\eta_t) \rightarrow \nabla^2 f(x) < 0$. By continuity, for t sufficiently small, $\frac{t^2}{2} d^T \nabla^2 f(\eta_t) d < 0$ for t sufficiently small, whence we find

$$f(x + td) < f(x),$$

for sufficiently small t , a contradiction. ■

↪ **Lemma 2.3**: Let $X \subset \mathbb{R}^n$ open, $f : X \rightarrow \mathbb{R}$ in C^2 . If $\bar{x} \in \mathbb{R}^n$ is such that $\nabla^2 f(\bar{x}) > 0$ (i.e. is positive definite), then there exists $\varepsilon, \mu > 0$ such that $B_\varepsilon(\bar{x}) \subset X$ and

$$d^T \nabla^2 f(x) d \geq \mu \|d\|^2, \quad \forall d \in \mathbb{R}^n, x \in B_\varepsilon(\bar{x}).$$

Combining this and Taylor's Theorem, we can deduce the following (our first "sufficient" result of this section):

↪ **Theorem 2.4** (Sufficient Optimality Condition): Let $X \subset \mathbb{R}^n$ open and $f \in C^2(X)$. Let \bar{x} be a stationary point of f such that $\nabla^2 f(\bar{x}) > 0$. Then, \bar{x} is a *strict* local minimizer of f .

2.1.1 Quadratic Approximation

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be C^2 and $\bar{x} \in \mathbb{R}^n$. By Taylor's, we can approximate

$$f(y) \approx g(y) := f(\bar{x}) + \nabla f(\bar{x})^T (y - \bar{x}) + \frac{1}{2} (y - \bar{x})^T \nabla^2 f(\bar{x}) (y - \bar{x}).$$

⊗ **Example 2.2** (Quadratic Functions): For $Q \in \mathbb{R}^{n \times n}$ symmetric, $c \in \mathbb{R}^n$ and $\gamma \in \mathbb{R}$, let

$$f : \mathbb{R}^n \rightarrow \mathbb{R}, \quad f(x) = \frac{1}{2}x^T Qx + c^T x + \gamma,$$

a typical quadratic function. Then,

$$\nabla f(x) = \frac{1}{2}(Q + Q^T)x + c = Qx + c, \quad \nabla^2 f(x) = Q.$$

We find that f has *no* minimizer if $c \notin \text{rge}(Q)$ or Q is not positive semi-definite, combining our previous two results. In turn, if Q is positive definite (and thus invertible), there is a unique local minimizer $\bar{x} = -Q^{-1}c$ (and global minimizer, as we'll see).

§2.2 Differentiable Convex Functions

↪ **Theorem 2.5:** Let $C \subset \mathbb{R}^n$ be open and convex and $f : C \rightarrow \mathbb{R}$ differentiable on C . Then:

1. f is convex (on C) iff

$$f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^T(x - \bar{x}) \quad \star_1$$

for every $x, \bar{x} \in C$;

2. f is *strictly* convex iff same inequality as 1. with strict inequality;

3. f is *strongly* convex with modulus $\sigma > 0$ iff

$$f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^T(x - \bar{x}) + \frac{\sigma}{2}\|x - \bar{x}\|^2 \quad \star_2$$

for every $x, \bar{x} \in C$.

PROOF. (1., \Rightarrow) Let $x, \bar{x} \in C$ and $\lambda \in (0, 1)$. Then,

$$f(\lambda x + (1 - \lambda)\bar{x}) - f(\bar{x}) \leq \lambda(f(x) - f(\bar{x})),$$

which implies

$$\frac{f(\bar{x} + \lambda(x - \bar{x})) - f(\bar{x})}{\lambda} \leq f(x) - f(\bar{x}).$$

Letting $\lambda \rightarrow 0^+$, the LHS \rightarrow the directional derivative of f at \bar{x} in the direction $x - \bar{x}$, which is equal to, by differentiability of f , $\nabla f(\bar{x})^T(x - \bar{x})$, thus the result.

(1., \Leftarrow) Let $x_1, x_2 \in C$ and $\lambda \in (0, 1)$. Let $\bar{x} := \lambda x_1 + (1 - \lambda)x_2$. \star_1 implies

$$f(x_i) \geq f(\bar{x}) + \nabla f(\bar{x})^T(x_i - \bar{x}),$$

for each of $i = 1, 2$. Taking “a convex combination of these inequalities”, i.e. multiplying them by $\lambda, 1 - \lambda$ resp. and adding, we find

$$\lambda f(x_1) + (1 - \lambda)f(x_2) \geq f(\bar{x}) + \nabla f(\bar{x})^T(\lambda x_1 + (1 - \lambda)x_2 - \bar{x}) = f(\lambda x_1 + (1 - \lambda)x_2),$$

thus proving convexity.

(2., \Rightarrow) Let $x \neq \bar{x} \in C$ and $\lambda \in (0, 1)$. Then, by 1., as we've just proven,

$$\lambda \nabla f(\bar{x})^T(x - \bar{x}) \leq f(\bar{x} + \lambda(x - \bar{x})) - f(\bar{x}).$$

But $f(\bar{x} + \lambda(x - \bar{x})) < \lambda f(x) + (1 - \lambda)f(\bar{x})$ by strict convexity, so we have

$$\lambda \nabla f(\bar{x})^T (x - \bar{x}) < \lambda(f(x) - f(\bar{x})),$$

and the result follows by dividing both sides by λ .

(2., \Leftarrow) Same as (1., \Leftarrow) replacing " \leq " with " $<$ ".

(3.) Apply 1. to $f - \frac{\sigma}{2}\|\cdot\|^2$, which is still convex if f σ -strongly convex, as one can check. ■

↪ **Corollary 2.1:** Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be convex and differentiable. Then,

- a) there exists an *affine function* $g : \mathbb{R}^n \rightarrow \mathbb{R}$ such that $g(x) \leq f(x)$ everywhere;
- b) if f strongly convex, then it is coercive, i.e. $\lim_{\|x\| \rightarrow \infty} f(x) = \infty$.

↪ **Corollary 2.2:** Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be convex and differentiable, then TFAE:

- 1. \bar{x} is a global minimizer of f ;
- 2. \bar{x} is a local minimizer of f ;
- 3. \bar{x} is a stationary point of f .

PROOF. 1. \Rightarrow 2. is trivial and 2. \Rightarrow 3. was already proven and 3. \Rightarrow 1. follows from the fact that differentiability gives

$$f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x})$$

for any $x \in \mathbb{R}^n$. ■

↪ **Corollary 2.3:** (2.2.4)

↪ **Theorem 2.6** (Twice Differentiable Convex Functions): Let $\Omega \subset \mathbb{R}^n$ open and convex and $f \in C^2(\Omega)$. Then,

- 1. f is convex on Ω iff $\nabla^2 f \geq 0$;
- 2. f is strictly convex on $\Omega \Leftarrow \nabla^2 f > 0$;
- 2. f is σ -strongly convex on $\Omega \Leftrightarrow \sigma \leq \lambda_{\min}(\nabla^2 f(x))$ for all $x \in \Omega$.

↪ **Corollary 2.4:** Let $A \in \mathbb{R}^{n \times n}$ be symmetric, $b \in \mathbb{R}^n$ and $f(x) := \frac{1}{2}x^T A x + b^T x$. Then,

- 1. f convex $\Leftrightarrow A \geq 0$;
- 2. f strongly convex $\Leftrightarrow A > 0$.

↪ **Theorem 2.7** (Convex Optimization): Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be convex and continuous, $X \subset \mathbb{R}^n$ convex (and nonempty), and consider the optimization problem

$$\min f(x) \text{ s.t. } x \in X \quad (*)$$

Then, the following hold:

1. \bar{x} is a global minimizer of $(*) \Leftrightarrow \bar{x}$ is a local minimizer of $(*)$
2. $\operatorname{argmin}_X f$ is convex (possibly empty)
3. f is strictly convex $\Rightarrow \operatorname{argmin}_X f$ has at *most* one element
4. f is strongly convex and differentiable, and X closed, $\Rightarrow \operatorname{argmin}_X f$ has *exactly* one element

PROOF. (1., \Rightarrow) Trivial. (1., \Leftarrow) Let \bar{x} be a local minimizer of f over X , and suppose towards a contradiction that there exists some $\hat{x} \in X$ such that $f(\hat{x}) < f(\bar{x})$. By convexity of f, X , we know for $\lambda \in (0, 1)$, $\lambda\bar{x} + (1 - \lambda)\hat{x} \in X$ and

$$f(\lambda\bar{x} + (1 - \lambda)\hat{x}) \leq \lambda f(\bar{x}) + (1 - \lambda)f(\hat{x}) < f(\bar{x}).$$

Letting $\lambda \rightarrow 1^-$, we see that $\lambda\bar{x} + (1 - \lambda)\hat{x} \rightarrow \bar{x}$; in particular, for any neighborhood of \bar{x} we can construct a point which strictly lower bounds $f(\bar{x})$, which contradicts the assumption that \bar{x} a local minimizer.

(2.) and (3.) are left as an exercise.

(4.) We know that f is strictly convex and level-bounded. By (3.) we know there is at most one minimizer, so we just need to show there exists one. Take $c \in \mathbb{R}$ such that $\operatorname{lev}_c(f) \cap X \neq \emptyset$ (which certainly exists by taking, say, $f(x)$ for some $x \in X$). Then, notice that $(*)$ and

$$\min_{x \in \operatorname{lev}_c f \cap X} f(x) \quad (**)$$

have the same solutions i.e. the same set of global minimizers (noting that this remains a convex problem). Since f continuous and $\operatorname{lev}_c f \cap X$ compact and nonempty, f attains a minimum on $\operatorname{lev}_c f \cap X$, as we needed to show. ■

Remark 2.1: Note that level sets of convex functions are convex, this is left as an exercise.

§2.3 Matrix Norms

We denote by $\mathbb{R}^{m \times n}$ the space of real-valued $m \times n$ matrices (i.e. of linear operators from $\mathbb{R}^n \rightarrow \mathbb{R}^m$).

↪ **Proposition 2.1** (Operator Norms): Let $\|\cdot\|_*$ be a norm on \mathbb{R}^m and \mathbb{R}^n , resp. Then, the map

$$\mathbb{R}^{m \times n} \ni A \mapsto \|A\|_* := \sup_{\substack{x \in \mathbb{R}^n, \\ \|x\|_* \neq 0}} \frac{\|Ax\|_*}{\|x\|_*} \in \mathbb{R}$$

is a norm on $\mathbb{R}^{m \times n}$. In addition,

$$\|A\|_* = \sup_{\|x\|_* = 1} \|Ax\|_* = \sup_{\|x\|_* \leq 1} \|Ax\|_*.$$

PROOF. We first note that all of these sup's are truly max's since they are maximizing continuous functions over compact sets.

Let $A \in \mathbb{R}^{m \times n}$. The first "In addition" equality follows from positive homogeneity, since $\frac{x}{\|x\|_*}$ a unit vector. For the second, note that " \leq " is trivial, since we are supping over a larger (super)set. For " \geq ", we have for any x with $\|x\|_* \leq 1$,

$$\|Ax\|_* = \|x\|_* \left\| A \frac{x}{\|x\|_*} \right\|_* \leq \left\| A \frac{x}{\|x\|_*} \right\|_*.$$

Supping both sides over all such x gives the result.

We now check that $\|\cdot\|_*$ actually a norm on $\mathbb{R}^{m \times n}$.

1. $\|A\|_* = 0 \Leftrightarrow \sup_{\|x\|_* = 1} \|Ax\|_* = 0 \Leftrightarrow \|Ax\|_* = 0 \forall \|x\|_* = 1 \Leftrightarrow Ax = 0 \forall \|x\|_* = 1 \Leftrightarrow A = 0$
2. For $\lambda \in \mathbb{R}, A \in \mathbb{R}^{m \times n}$, $\|\lambda A\|_* = \sup \|\lambda Ax\|_* = |\lambda| \cdot \sup \|Ax\|_* = |\lambda| \|A\|_*$
3. For $A, B \in \mathbb{R}^{m \times n}$, $\|A + B\|_* \leq \|A\|_* + \|B\|_*$ using properties of sups of sums

■

↪ **Proposition 2.2:** Let $A = (a_{ij})_{\substack{i=1, \dots, m \\ j=1, \dots, n}} \in \mathbb{R}^{m \times n}$, then:

1. $\|A\|_1 = \max_{j=1}^n \sum_{i=1}^m |a_{ij}|$
2. $\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)} = \sigma_{\max}(A)$
3. $\|A\|_\infty = \max_{i=1}^m \sum_{j=1}^n |a_{ij}|$

↪ **Proposition 2.3:** Let $\|\cdot\|_*$ be a norm on $\mathbb{R}^n, \mathbb{R}^m$, and \mathbb{R}^p . For $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$,

1. $\|Ax\|_* \leq \|A\|_* \cdot \|x\|_*$
2. $\|AB\|_* \leq \|A\|_* \cdot \|B\|_*$

↪ **Proposition 2.4** (Banach Lemma): Let $C \in \mathbb{R}^{n \times n}$ with $\|C\| < 1$, where $\|\cdot\|$ submultiplicative. Then, $I + C$ is invertible, and

$$\|(1 + C)^{-1}\| \leq \frac{1}{1 - \|C\|}.$$

PROOF. We have for any m ,

$$\left\| \sum_{i=1}^m (-C)^i \right\| \leq \sum_{i=1}^m \|C\|^i \xrightarrow{m \rightarrow \infty} \frac{1}{1 - \|C\|}.$$

Hence, $A_m := \sum_{i=1}^m (-C)^i$ a sequence of matrices with bounded norm uniformly in m , and thus has a converging subsequence, so wlog $A_m \rightarrow A \in \mathbb{R}^{n \times n}$ (by relabelling).

Moreover, observe that

$$A_m \cdot (I + C) = \sum_{i=0}^m (-C)^i (I + C) = \sum_{i=0}^m [(-C)^i - (-C)^{i+1}] = (-C)^0 - (-C)^{m+1} = I - (-C)^{m+1}.$$

Now, $\|C^{m+1}\| \leq \|C\|^{m+1} \rightarrow 0$, since $\|C\| < 1$, thus $C \rightarrow 0$. Hence, taking limits in the line above implies

$$A(I + C) = \lim_{m \rightarrow \infty} A_m(I + C) = I,$$

implying A the inverse of $(I + C)$, proving the proposition. ■

↪ **Corollary 2.5:** Let $A, B \in \mathbb{R}^{n \times n}$ with $\|I - BA\| < 1$ for $\|\cdot\|$ submultiplicative. Then, A and B are invertible, and $\|B^{-1}\| \leq \frac{\|A\|}{1 - \|I - BA\|}$.

§3 DESCENT METHODS

§3.1 A General Line-Search Method

We deal with the unconstrained problem

$$\min_{x \in \mathbb{R}^n} f(x) \quad (\star).$$

↪ **Definition 3.1** (Descent Direction): Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $x \in \mathbb{R}^n$. $d \in \mathbb{R}^n$ is a *descent direction* of f at x if there exists a $\bar{t} > 0$ such that $f(x + td) < f(x)$ for all $t \in (0, \bar{t})$.

↪ **Proposition 3.1:** If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is directionally differentiable at $x \in \mathbb{R}^n$ in the direction d with $f'(x; d) < 0$, then d a descent direction of f at x ; in particular if f differentiable at x , then true for d if $\nabla f(x)^T d < 0$.

↪ **Corollary 3.1:** Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ differentiable, $B \in \mathbb{R}^{n \times n}$ positive definite, and $x \in \mathbb{R}^n$. Then $\nabla f(x) \neq 0 \Rightarrow -B\nabla f(x)$ is a descent direction of f at x .

PROOF. $\nabla f(x)^T (-B\nabla f(x)) = -\nabla f(x)^T B \nabla f(x) < 0$. ■

A generic method/strategy for solving (\star) :

S1. (Initialization) Choose $x^0 \in \mathbb{R}^n$ and set $k := 0$

S2. (Termination) If x^k satisfies a “termination criterion”, STOP

S3. (Search direction) Determine d^k such that $\nabla f(x^k)^T d^k < 0$

S4. (Step-size) Determine $t_k > 0$ such that $f(x^k + t_k d^k) < f(x^k)$

S5. (Update) Set $x^{k+1} := x^k + t_k d^k$, iterate k , and go back to step 2.

Remark 3.1: a) The generic choice for d^k in 3. is just $d^k := -B_k \nabla f(x^k)$ for some $B_k > 0$. We focus on:

- $B_k = I$ (gradient-descent)
- $B_k = \nabla^2 f(x^k)^{-1}$ (Newton's method)
- $B_k \approx \nabla^2 f(x^k)^{-1}$ (quasi Newton's method)

b) Step 4. is called *line-search*, since $t_k > 0$ determined by looking at

$$0 < t \mapsto f(x^k + td^k),$$

i.e. along the (half)line $t > 0$.

c) Executing Step 4. is a trade-off between

- (i) decreasing f along $x^k + td^k$ as much as possible;
- (ii) keeping computational efforts low.

For instance, the *exact minimization rule* $t_k = \operatorname{argmin}_{t>0} f(x_k + td^k)$ overemphasizes (i) over (ii).

↪ **Definition 3.2** (Step-size rule): Let $f \in C^1(\mathbb{R}^n)$ and

$$A_f := \{(x, d) \mid \nabla f(x)^T d < 0\}.$$

A (possible set-valued) map

$$T : (x, d) \in A_f \mapsto T(x, d) \in \mathbb{R}_+$$

is called a *step-size rule* for f .

If T is well-defined for all C^1 -functions, we say T well-defined.

3.1.1 Global Convergence of Algorithm 3.1

↪ **Definition 3.3** (Efficient step-size): Let $f \in C^1(\mathbb{R}^n)$. The step-size rule T is called *efficient* for f if there exists $\theta > 0$ such that

$$f(x + td) \leq f(x) - \theta \left(\frac{\nabla f(x)^T d}{\|d\|} \right)^2, \quad \forall t \in T(x, d), (x, d) \in A_f.$$

↪ **Theorem 3.1:** Let $f \in C^1(\mathbb{R}^n)$. Let $\{x^k\}, \{d^k\}, \{t_k\}$ be generated by Algorithm 3.1. Assume the following:

1. $\exists c > 0$ such that $-\left(\nabla f(x^k)^T d^k\right) / (\|\nabla f(x^k)\| \cdot \|d^k\|) \geq c$ for all k (this is called the *angle condition*), and
2. there exists $\theta > 0$ such that $f(x^k + t_k d^k) \leq f(x^k) - \theta \cdot \left(\nabla f(x^k)^T d^k / \|d^k\|\right)^2$ for all k (which is satisfied if $t_k \in T(x^k, d^k)$ for an efficient T).

Then, every cluster point of $\{x^k\}$ is a stationary point of f .

PROOF. By condition 2., there is $\theta > 0$ such that

$$f(x^{k+1}) \leq f(x^k) - \theta \left(\frac{\nabla f(x^k)^T d^k}{\|d^k\|} \right)^2,$$

for all $k \in \mathbb{N}$. By 1., we know

$$\left(\frac{\nabla f(x^k)^T d^k}{\|d^k\|} \right)^2 \geq c^2 \|\nabla f(x^k)\|^2.$$

Put $\kappa := \theta c^2$, then these two inequalities imply

$$f(x^{k+1}) \leq f(x^k) - \kappa \cdot \|\nabla f(x^k)\|^2. \quad (*)$$

Let \bar{x} be a cluster point of $\{x^k\}$. As $\{f(x^k)\}$ is monotonically decreasing (by construction in the algorithm), and has cluster point $f(\bar{x})$ by continuity, it follows that $f(x_k) \rightarrow f(\bar{x})$ along the whole sequence. In particular, $f(x^{k+1}) - f(x^k) \rightarrow 0$; thus, from (*),

$$0 \leq \kappa \|\nabla f(x^k)\|^2 \leq f(x^k) - f(x^{k+1}) \rightarrow 0,$$

and thus $\nabla f(x^k) \rightarrow \nabla f(\bar{x}) = 0$, so indeed \bar{x} a stationary point of f . ■

§3.2 The Gradient Method

We specialize Algorithm 3.1 here. Specifically, we'll take

$$d^k := -\nabla f(x^k);$$

it's known that

$$\frac{-\nabla f(x^k)}{\|\nabla f(x^k)\|} = \operatorname{argmin}_{d: \|d\| \leq 1} \nabla f(x^k)^T d,$$

with $\|\cdot\|$ the 2 norm.

We use a step-size rule called "Armijo rule". Choose parameters $\beta, \sigma \in (0, 1)$. For $(x, d) \in \mathcal{A}_f$, we define our step-size rule by

$$T_A(x, d) := \max_{\ell \in \mathbb{N}_0} \left\{ \beta^\ell \mid \underbrace{f(x + \beta^\ell d) \leq f(x) + \beta^\ell \sigma \nabla f(x)^T d}_{\text{"Armijo condition"}} \right\}.$$

For instance, consider $f(x) = (x - 1)^2 - 1$. The minimum of this function is $f^* = -1$. Choose $x^k := \frac{1}{k}$, then

$$f(x^k) = \frac{2k + 1}{k^2} \rightarrow 0 \neq f^*,$$

even though $f(x^{k+1}) - f(x^k) < 0$; we don't actually reach the right stationary point with our chosen step size.

⊗ **Example 3.1** (Illustration of Armijo Rule): For $(x, d) \in A_f$ and f smooth on \mathbb{R}^n , defined $\phi : \mathbb{R} \rightarrow \mathbb{R}, \phi(t) := f(x + td)$. The map $t \mapsto \sigma \phi'(0)t + \phi(0) = \sigma t \nabla f(x)^T d + \phi(0)$

↪ **Proposition 3.2:** Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be differentiable with $\beta, \sigma \in (0, 1)$. Then for $(x, d) \in A_f$, there exists $\ell \in \mathbb{N}_0$ such that

$$f(x + \beta^\ell d) \leq f(x) + \beta^\ell \sigma \nabla f(x)^T d,$$

i.e. $T_A(x, d) \neq \emptyset$.

PROOF. Suppose not, i.e.

$$\frac{f(x + \beta^\ell d) - f(x)}{\beta^\ell} > \sigma \nabla f(x)^T d, \forall \ell \in \mathbb{N}_0.$$

Letting $\ell \rightarrow \infty$, the left-hand side converges to $\nabla f(x)^T d$, so

$$\nabla f(x)^T d \geq \sigma \nabla f(x)^T d.$$

But $(x, d) \in A_f$, so $\nabla f(x)^T d < 0$ so dividing both sides of this inequality by this quantity, this implies $\sigma \leq 0$, which is a contradiction. ■

We now prove convergence of an algorithm based on the Armijo Rule:

| Gradient Descent with Armijo Rule |
|--|
| S0. Choose $x^0 \in \mathbb{R}^n, \sigma, \beta \in (0, 1), \varepsilon \geq 0$, and set $k := 0$ |
| S1. If $\ \nabla f(x^k)\ \leq \varepsilon$, STOP |
| S2. Set $d^k := -\nabla f(x^k)$ |
| S3. Determine $t_k > 0$ by |
| $t_k = T_A(x, d)$ |
| as defined above. |
| S4. Set $x^{k+1} = x^k + t_k d^k$, iterate k and go to S1. |

↪ **Lemma 3.1:** Let $f \in C^1(\mathbb{R}^n), x^k \rightarrow x, d^k \rightarrow d$ and $t_k \downarrow 0$. Then

$$\lim_{k \rightarrow \infty} \frac{f(x^k + t_k d^k) - f(x^k)}{t_k} = \nabla f(x)^T d.$$

PROOF. Left as an exercise. ■

↪ **Theorem 3.2:** Let $f \in C^1(\mathbb{R}^n)$. Then every cluster point of a sequence $\{x^k\}$ generated by Algorithm 3.2 is a stationary point of f .

PROOF. Let \bar{x} be a cluster point of $\{x^k\}$ and let $x^k \xrightarrow{k \in K} \bar{x}$, K an infinite subset of \mathbb{N} .

Assume towards a contradiction $\nabla f(\bar{x}) \neq 0$. As $f(x^k)$ is monotonically decreasing with cluster point $f(\bar{x})$, it must be that $f(x^k) \rightarrow f(\bar{x})$ along the whole sequence so $f(x^{k+1}) - f(x^k) \rightarrow 0$. Thus,

$$0 \leq t_k \|\nabla f(x^k)\|^2 \stackrel{S2}{=} -t_k \nabla f(x^k)^T d^k \stackrel{S3}{\leq} \frac{f(x^k) - f(x^{k+1})}{\sigma} \rightarrow 0.$$

Thus, $0 = \lim_{k \in K} t_k \|\nabla f(x^k)\| = \|\nabla f(\bar{x})\| \lim_{k \in K} t_k$. We assumed \bar{x} not a stationary point, so it follows that $t_k \xrightarrow{k \in K} 0$. By S3, for $\beta^{\ell_k} = t_k$,

$$\frac{f(x^k + \beta^{\ell_k-1} d^k) - f(x^k)}{\beta^{\ell_k-1}} > \sigma \nabla f(x^k)^T d^k.$$

Letting $k \rightarrow \infty$ along K , the LHS converges to, by the previous lemma, to

$$\nabla f(\bar{x})^T d = -\nabla f(\bar{x})^T \nabla f(\bar{x}) = -\|\nabla f(\bar{x})\|^2,$$

and the RHS converges to $\sigma \|\nabla f(\bar{x})\|^2$, which implies

$$-\|\nabla f(\bar{x})\|^2 \geq \sigma \|\nabla f(\bar{x})\|^2,$$

which implies σ negative, a contradiction. ■

Remark 3.2: The proof above shows, the following: Let $\{x^k\}$ such that $x^{k+1} := x^k + t_k d^k$ for $d^k \in \mathbb{R}^n$, $t_k > 0$, and let $f(x^{k+1}) \leq f(x^k)$ and $x^k \xrightarrow{K} \bar{x}$ such that $d^k = -\nabla f(x^k)$, $t_k = T_A(x^k, d^k)$ for all $k \in K$. Then $\nabla f(\bar{x}) = 0$; i.e., all of the “focus” is on the subsequence along K . The only time we needed the whole sequence was to use the fact that $f(x^k) \rightarrow f(\bar{x})$ along the whole sequence.

§3.3 Newton-Type Methods

3.3.1 Convergence Rates and Landau Notation

Definition 3.4: Let $\{x^k \in \mathbb{R}^n\}$ converge to \bar{x} . Then, $\{x^k\}$ converges:

1. *linearly* to \bar{x} if there exists $c \in (0, 1)$ such that

$$\|x^{k+1} - \bar{x}\| \leq c \|x^k - \bar{x}\|, \forall k;$$

2. *superlinearly* to \bar{x} if

$$\lim_{k \rightarrow \infty} \frac{\|x^{k+1} - \bar{x}\|}{\|x^k - \bar{x}\|} = 0;$$

3. *quadratically* to \bar{x} if there exists $C > 0$ such that

$$\|x^{k+1} - \bar{x}\| \leq C \|x^k - \bar{x}\|^2, \forall k.$$

Remark 3.3: 3. \Rightarrow 2. \Rightarrow 1.

Remark 3.4: We needn't assume $x^k \rightarrow \bar{x}$ for the first two definitions; their statements alone imply convergence. However, the last does not; there exists sequences with this property that do not converge.

↪ **Definition 3.5** (Landau Notation): Let $\{a_k\}, \{b_k\}$ be positive sequences $\downarrow 0$. Then,

1. $a_k = o(b_k) \Leftrightarrow \lim_{k \rightarrow \infty} \frac{a_k}{b_k} = 0$;
2. $a_k = O(b_k) \Leftrightarrow \exists C > 0 : a_k \leq Cb_k$ for all k (sufficiently large).

Remark 3.5: If $x^k \rightarrow \bar{x}$, then

1. the convergence is superlinear $\Leftrightarrow \|x^{k+1} - \bar{x}\| = o(\|x^k - \bar{x}\|)$;
2. the convergence is quadratic $\Leftrightarrow \|x^{k+1} - \bar{x}\| = O(\|x^k - \bar{x}\|^2)$.

3.3.2 Newton's Method for Nonlinear Equations

We consider the nonlinear equation

$$F(x) = 0, \quad (*)$$

where $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is smooth (continuously differentiable). Our goal is to find a numerical scheme that can determine approximate zeros of F , i.e. solutions to $(*)$. The idea of Newton's method for such a problem, is, given $x^k \in \mathbb{R}^n$, to consider the (affine) linear approximation of F about x^k ,

$$F_k : x \mapsto F(x^k) + F'(x^k)(x - x^k),$$

where F' the Jacobian of F . Then, we compute x^{k+1} as a solution of $F_k(x) = 0$. Namely, if $F'(x^k)$ invertible, then solving for $F_k(x^{k+1}) = 0$, we find

$$x^{k+1} = x^k - F'(x^k)^{-1}F(x^k).$$

More generally, one solves $F'(x^k)d = -F(x^k)$ and sets $x^{k+1} := x^k + d^k$.

Specifically, we have the following algorithm:

| Newton's Method (Local Version) |
|---|
| S0. Choose $x^0 \in \mathbb{R}^n, \varepsilon > 0$, and set $k := 0$. |
| S1. If $\ F(x^k)\ < \varepsilon$, STOP. |
| S2. Compute d^k as a solution of <i>Newton's equation</i> |
| $F'(x^k)d = -F(x^k).$ |
| S3. Set $x^{k+1} := x^k + d^k$, increment k and go to S1. |

↪ **Lemma 3.2:** Let $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be C^1 , and $\bar{x} \in \mathbb{R}^n$ such that $F'(\bar{x})$ is invertible. Then, there exists $\varepsilon > 0$ such that $F'(x)$ remains invertible for all $x \in B_\varepsilon(\bar{x})$, and there exists $C > 0$ such that

$$\|F'(x)^{-1}\| \leq C, \quad \forall x \in B_\varepsilon(\bar{x}).$$

PROOF. Since F' continuous at \bar{x} , there exists $\varepsilon > 0$ such that $\|F'(\bar{x}) - F'(x)\| \leq \frac{1}{2\|F'(\bar{x})^{-1}\|}$ for all $x \in B_\varepsilon(\bar{x})$. Then, for all $x \in B_\varepsilon(\bar{x})$,

$$\begin{aligned}\|I - F'(x)F'(\bar{x})^{-1}\| &= \|(F'(\bar{x}) - F'(x))F'(\bar{x})^{-1}\| \\ &\leq \|F'(\bar{x}) - F'(x)\| \|F'(\bar{x})^{-1}\| \leq \frac{1}{2} < 1.\end{aligned}$$

By a corollary of the Banach lemma, $F'(x)$ invertible over $B_\varepsilon(\bar{x})$, and

$$\|F'(x)^{-1}\| \leq \frac{\|F'(\bar{x})^{-1}\|}{1 - \|I - F'(x)F'(\bar{x})^{-1}\|} \leq 2\|F'(\bar{x})^{-1}\| =: C.$$

■

Remark 3.6: Observe $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is differentiable at \bar{x} if and only if $\|F(x^k) - F(\bar{x}) - F'(\bar{x})(x^k - \bar{x})\| = o(\|x^k - \bar{x}\|)$ for every $x^k \rightarrow \bar{x}$.

This can be sharpened if F' is continuous or even locally Lipschitz.

↪ **Lemma 3.3:** Let $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be continuously differentiable and $x^k \rightarrow \bar{x}$, then:

1. $\|F(x^k) - F(\bar{x}) - F'(\bar{x})(x^k - \bar{x})\| = o(\|x^k - \bar{x}\|)$;
2. if F' locally Lipschitz at \bar{x} , then $\|F(x^k) - F(\bar{x}) - F'(x^k)(x^k - \bar{x})\| = O(\|x^k - \bar{x}\|^2)$.

PROOF.

1. Observe that

$$\begin{aligned}&\|F(x^k) - F(\bar{x}) - F'(x^k)(x^k - \bar{x})\| \\ &\leq \|F(x^k) - F(\bar{x}) - F'(\bar{x})(x^k - \bar{x})\| + \|F'(\bar{x})(x^k - \bar{x}) - F'(x^k)(x^k - \bar{x})\| \\ &\leq \|F(x^k) - F(\bar{x}) - F'(\bar{x})(x^k - \bar{x})\| + \|F'(\bar{x}) - F'(x^k)\| \|x^k - \bar{x}\|.\end{aligned}$$

The left-hand term is $o(\|x^k - \bar{x}\|)$ by our observations previously, and the right-hand term is as well by continuity of F' , thus so is the sum.

2. Let $L > 0$ be a local Lipschitz constant of F' at \bar{x} . Then,

$$\begin{aligned}\|F(x^k) - F(\bar{x}) - F'(x^k)(x^k - \bar{x})\| &= \left\| \int_0^1 F'(\bar{x} + t(x^k - \bar{x})) dt (x^k - \bar{x}) - F'(x^k)(x^k - \bar{x}) \right\| \\ &\leq \int_0^1 \|F'(\bar{x} + t(x^k - \bar{x})) - F'(x^k)\| dt \cdot \|x^k - \bar{x}\| \\ &\leq L \int_0^1 |1 - t| \|x^k - \bar{x}\| dt \cdot \|x^k - \bar{x}\| \\ &= L \|x^k - \bar{x}\|^2 \int_0^1 (1 - t) dt = \frac{L}{2} \|x^k - \bar{x}\|^2,\end{aligned}$$

which implies the result.

■

↪ **Theorem 3.3:** Let $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be continuously differentiable, $\bar{x} \in \mathbb{R}^n$ such that $F(\bar{x}) = 0$ and $F'(\bar{x})$ is invertible. Then, there exists an $\varepsilon > 0$ such that for every $x^0 \in B_\varepsilon(\bar{x})$, we have:

1. Algorithm 3.3 is well-defined and generates a sequence $\{x^k\}$ which converges to \bar{x} ;
2. the rate of convergence is (at least) linear;
3. if F' is locally Lipschitz at \bar{x} , then the rate is quadratic.

PROOF.

1. By the previous lemma, we know there is $\varepsilon_1, c > 0$ such that $\|F'(x)^{-1}\| \leq c$ for all $x \in B_{\varepsilon_1}(x)$. Further, there exists an $\varepsilon_2 > 0$ such that $\|F(x) - F(\bar{x}) - F'(x)(x - \bar{x})\| \leq \frac{1}{2c}\|x - \bar{x}\|$ for all $x \in B_{\varepsilon_2}(\bar{x})$. Take $\varepsilon = \min\{\varepsilon_1, \varepsilon_2\}$ and pick $x^0 \in B_\varepsilon(\bar{x})$. Then, x^1 is well-defined, since $F'(x^0)$ is invertible, and so

$$\begin{aligned}
\|x^1 - \bar{x}\| &= \|x^0 - F'(x^0)^{-1}F(x^0) - \bar{x}\| \\
&= \left\| F'(x^0)^{-1} \left(F(x^0) - \underbrace{F(\bar{x})}_{=0} - F'(x^0)(x^0 - \bar{x}) \right) \right\| \\
&\leq \|F'(x^0)^{-1}\| \|F(x^0) - F(\bar{x}) - F'(x^0)(x^0 - \bar{x})\| \\
&\leq c \cdot \frac{1}{2c} \|x^0 - \bar{x}\| \\
&= \frac{1}{2} \|x^0 - \bar{x}\| < \frac{\varepsilon}{2},
\end{aligned}$$

so in particular, $x^1 \in B_{\varepsilon/2}(\bar{x}) \subset B_\varepsilon(\bar{x})$. Inductively,

$$\|x^k - \bar{x}\| \leq \left(\frac{1}{2}\right)^k \|x^0 - \bar{x}\|,$$

for every $k \in \mathbb{N}$. Thus, x^k well-defined and converges to \bar{x} .

2., 3. Analogous to 1.,

$$\begin{aligned}
\|x^{k+1} - \bar{x}\| &= \|x^k - F'(x^k)^{-1}F(x^k) - \bar{x}\| \\
&= \left\| x^k - F'(x^k)^{-1}F(x^k) - \bar{x} \right\| \\
&\leq \|F'(x^k)^{-1}\| \|F(x^k) - F(\bar{x}) - F'(x^k)(x^k - \bar{x})\| \\
&\leq c \|F(x^k) - F(\bar{x}) - F'(x^k)(x^k - \bar{x})\|.
\end{aligned}$$

This final line is little o of $\|x^k - \bar{x}\|$ or this quantity squared by the previous lemma, which proves the result depending on the assumptions of 2., 3. ■

3.3.3 Newton's Method for Optimization Problem

Consider

$$\min_{x \in \mathbb{R}^n} f(x),$$

with $f : \mathbb{R}^n \rightarrow \mathbb{R}$ twice continuously differentiable. Recall that if \bar{x} a local minimizer of f , $\nabla f(\bar{x}) = 0$. We'll now specialize Newton's to $F := \nabla f$:

| Newton's Method for Optimization (Local Version) |
|--|
|--|

| |
|---|
| S0. Choose $x^0 \in \mathbb{R}^n, \varepsilon > 0$, and set $k := 0$. |
|---|

| |
|--|
| S1. If $\ \nabla f(x^k)\ < \varepsilon$, STOP. |
|--|

| |
|---|
| S2. Compute d^k as a solution of <i>Newton's equation</i> |
|---|

| |
|--------------------------------------|
| $\nabla^2 f(x^k)d = -\nabla f(x^k).$ |
|--------------------------------------|

| |
|--|
| S3. Set $x^{k+1} := x^k + d^k$, increment k and go to S1. |
|--|

We then have an analogous convergence result to the previous theorem by simply applying $F := \nabla f$; in particular, if f thrice continuously differentiable, we have quadratic convergence.

⊗ **Example 3.2:** Let $f(x) := \sqrt{x^2 + 1}$. Then $f'(x) = \frac{x}{\sqrt{x^2+1}}, f''(x) = \frac{1}{(x^2+1)^{3/2}}$. Newton's equation (i.e. Algorithm 3.4, S2) reads in this case:

$$\frac{1}{(x_k^2 + 1)^{3/2}}d = -\frac{x_k}{\sqrt{x_k^2 + 1}}.$$

This gives solution $d_k = -(x_k^2 + 1)x_k$, so $x_{k+1} = -x_k^3$. Then, notice that if:

$$|x_0| < 1 \Rightarrow x_k \rightarrow 0, \text{ quadratically}$$

$$|x_0| > 1 \Rightarrow x_k \text{ diverges}$$

$$|x_0| = 1 \Rightarrow |x_k| = 1 \forall k,$$

so the convergence is truly local; if we start too far from 0, we'll never have convergence.

We can see from this example that this truly a local algorithm. A general globalization strategy is to:

- if Newton's equation has no solution, or doesn't provide sufficient decay, set $d^k := -\nabla f(x^k)$;
- introduce a step-size.

| Newton's Method (Global Version) |
|---|
| S0. Choose $x^0 \in \mathbb{R}^n, \varepsilon > 0, \rho > 0, p > 2, \beta \in (0, 1), \sigma \in (0, 1/2)$ and set $k := 0$ |
| S1. If $\ \nabla f(x^k)\ < \varepsilon$, STOP |
| S2. Determine d^k as a solution of $\nabla^2 f(x^k) d = -\nabla f(x^k).$ |
| If no solution exists, or if $\nabla f(x^k)^T d^k \leq -\rho \ d^k\ ^p$, is violated, set $d^k := -\nabla f(x^k)$ |
| S3. Determine $t_k > 0$ by the Armijo back-tracking rule, i.e. $t_k := \max_{\ell \in \mathbb{N}_0} \left\{ \beta^\ell \mid f(x^k + \beta^\ell d^k) \leq f(x^k) + \beta^\ell \sigma \nabla f(x^k)^T d^k \right\}$ |
| S4. Set $x^{k+1} := x^k + t_k d^k$, increment k to $k + 1$, and go back to S1. |

Remark 3.7: S3. well-defined since in either choice of d^k in S2., we will have a descent direction so the choice of t_k in S3. is valid; i.e. $(x^k, d^k) \in A_f$ for every k .

→ **Theorem 3.4** (Global convergence of Algorithm 3.5): Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be twice continuously differentiable. Then every cluster point of $\{x^k\}$ generated by Algorithm 3.5 is a stationary point of f .

Remark 3.8: Note that we didn't impose any invertibility condition on the Hessian of f ; indeed, if say the hessian was nowhere invertible, then Algorithm 3.5 just becomes the gradient method with Armijo back-tracking, for which have already established this result.

PROOF. Let $\{x^k\}$ be generated by Algorithm 3.5, with $\{x^k\}_K \rightarrow \bar{x}$. If $d^k := -\nabla f(x^k)$ for infinitely many $k \in K$ (i.e. along a subsubsequence of $\{x^k\}$), then we have nothing to prove by the previous remark.

Otherwise, assume wlog (by passing to a subsubsequence again if necessary) that d^k is determined by the Newton equation for all $k \in K$. Suppose towards a contradiction that $\nabla f(\bar{x}) \neq 0$. By Newton's equation,

$$\|\nabla f(x^k)\| = \|\nabla^2 f(x^k) d^k\| \leq \|\nabla^2 f(x^k)\| \|d^k\|, \quad \forall k \in K.$$

By assumption $\|\nabla^2 f(x^k)\| \neq 0$; if it were, then by assumption $\nabla f(x^k) = 0$, i.e. we'd have already reached our stationary point, which we assumed doesn't happen. So, we may write $\frac{\|\nabla f(x^k)\|}{\|\nabla^2 f(x^k)\|} \leq \|d^k\|$ for all $k \in K$. We claim that there exists $c_1, c_2 > 0$ such that

$$0 < c_1 \leq \|d^k\| \leq c_2, \quad \forall k \in K.$$

We have existence of c_1 since, if it didn't, we could find a subsequence of the d^k 's such that $d^k \rightarrow 0$ along this subsequence; but by our bound above and the fact that $\|\nabla^2 f(x^k)\|$ uniformly bounded (by continuity), then $\|\nabla f(x^k)\|$ would converge to zero along the subsequence too, a contradiction.

The existence of c_2 follows from the sufficient decrease condition. Indeed, suppose such a c_2 didn't exist; by the condition

$$\nabla f(x^k)^T \frac{d^k}{\|d^k\|} \leq -\rho \|d^k\|^{p-1};$$

the left-hand side is bounded (since $\nabla f(x^k) \rightarrow \nabla f(\bar{x})$ and $\frac{d^k}{\|d^k\|}$ lives on the unit sphere). Since c_2 assumed not to exist, there is a subsequence $\|d^k\| \rightarrow \infty$, but then $-\rho \|d^k\|^{p-1} \rightarrow -\infty$, contradicting the fact that the LHS is bounded. Hence, there also exists such a c_2 as claimed.

As $\{f(x^k)\}$ is monotonically decreasing (by construction in S3) and converges along a subsequence K to $f(\bar{x})$, then $f(x^k)$ converges along the whole sequence to $f(\bar{x})$. In particular, $f(x^{k+1}) - f(x^k) \rightarrow 0$. Then,

$$\frac{f(x^{k+1}) - f(x^k)}{\sigma} \leq t_k \nabla f(x^k)^T d^k \leq -\rho t_k \|d^k\|^p \leq 0.$$

Taking $k \rightarrow \infty$ along K , we see that $t_k \|d^k\|^p \rightarrow 0$ along K as well. We show now that $\{t_k\}_K$ actually uniformly bounded away from zero. Suppose not. Then, along a sub(sub)sequence, $t_k \rightarrow 0$. By the Armijo rule, $t_k = \beta^{\ell_k}$, for $\ell_k \in \mathbb{N}_0$, uniquely determined. Since $t_k \rightarrow 0$, then $\ell_k \rightarrow \infty$. On the other hand, by S3,

$$\frac{f(x^k + \beta^{\ell_k-1} d^k) - f(x^k)}{\beta^{\ell_k-1}} > \sigma \nabla f(x^k)^T d^k.$$

Suppose $d^k \rightarrow \bar{d} \neq 0$ (by again passing to a subsequence if necessary), which we may assume by boundedness. Taking $k \rightarrow \infty$, the LHS converges to $\nabla f(\bar{x})^T \bar{d}$ and the RHS converges to $\sigma \nabla f(\bar{x})^T \bar{d}$ so $\nabla f(\bar{x})^T \bar{d} \geq \sigma \nabla f(\bar{x})^T \bar{d}$, which implies since $\sigma \in (0, \frac{1}{2})$ that $\nabla f(\bar{x})^T \bar{d} \geq 0$. Taking $k \rightarrow \infty$ in the sufficient decrease condition statement shows that this is a contradiction. Hence, t_k uniformly bounded away from 0. Hence, there exists a $\bar{t} > 0$ such that $t_k \geq \bar{t}$ for all $k \in K$. But we had that $t^k \nabla f(x^k)^T d^k \rightarrow 0$, so by boundedness of t_k it must be that $\nabla f(x^k)^T d^k \rightarrow 0$ along the subsequence; by the sufficient decrease condition again, it must be that $d^k \rightarrow 0$, which it can't, as we showed it was uniformly bounded away, and thus we have a contradiction. ■

↪ **Theorem 3.5** (Fast local convergence of Algorithm 3.5): Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be twice continuously differentiable, $\{x^k\}$ generated by Algorithm 3.5. If \bar{x} is a cluster point of $\{x^k\}$ with $\nabla^2 f(\bar{x}) > 0$. Then:

1. $\{x^k\} \rightarrow \bar{x}$ along the *whole* sequence, so \bar{x} is a strict local minimizer of f ;
2. for $k \in \mathbb{N}$ sufficiently large, d^k will be determined by the Newton equation in S2;
3. $\{x^k\} \rightarrow \bar{x}$ at least superlinearly;
4. if $\nabla^2 f$ locally Lipschitz, $\{x^k\} \rightarrow \bar{x}$ quadratically.

§3.4 Quasi-Newton Methods

In Newton's, in general we need to find

$$d^k \text{ solving } \nabla^2 f(x^k) d = -\nabla f(x^k).$$

Advantages/disadvantages:

(+) Global convergence with fast local convergence

(-) Evaluating $\nabla^2 f$ can be expensive/impossible.

Dealing with the second, there are two general approaches:

- *Direct Methods*: replace $\nabla^2 f(x^k)$ with some matrix H_k approximating it;
- *Indirect Methods*: replace $\nabla^2 f(x^k)^{-1}$ by B_k approximating it;

where H_k, B_k reasonably computational, and other convergence results are preserved.

3.4.1 Direct Methods

The typical conditions we put on H_{k+1} as described above are:

1. H_{k+1} symmetric
2. H_{k+1} satisfies the *Quasi-Newton equation* (QNE)

$$H_{k+1}s^k = y^k, \quad s^k := x^{k+1} - x^k, \quad y^k := \nabla f(x^{k+1}) - \nabla f(x^k)$$

3. H_{k+1} can be achieved from H_k “efficiently”
4. The result method has strong local convergence properties

Remark 3.9: Suppose x^k a current iterate for an algorithm to minimize $f : \mathbb{R}^n \rightarrow \mathbb{R}$ for $f \in C^2$.

1. $\nabla^2 f(x^k)$ does not generally satisfy QNE;
2. condition 1 above is motivated by the fact that Hessians are symmetric;
3. the QNE is motivated by the mean-value theorem for vector-valued functions,

$$\nabla f(x^{k+1}) - \nabla f(x^k) = \int_0^1 \nabla^2 f(x^k + t(x^{k+1} - x^k)) dt \cdot (x^{k+1} - x^k);$$

we can think of the integrated term as an averaging of the Hessian along the line between x^k, x^{k+1} .

We follow a so-called *symmetric rank-2 approach*; given H_k , we update

$$H_{k+1} = H_k + \gamma uu^T + \delta vv^T, \quad \gamma, \delta \in \mathbb{R}; u, v \in \mathbb{R}^n. \quad (1)$$

Note that if we put $S := uu^T$ for $u \neq 0$, $\text{rank}(S) = 1$ and $S^T = S$.

So, the ansatz we take is

$$y^k = H_{k+1}s^k = H_k s^k + \gamma uu^T s^k + \delta vv^T s^k. \quad (2)$$

If $H_k > 0$ and $(y^k)^T s^k \neq 0$, then taking $u := y^k, v := H_k s^k, \gamma := \frac{1}{(y^k)^T s^k}$ and $\delta := -\frac{1}{(s^k)^T H_k s^k}$ will solve (2), and gives the formula

$$H_{k+1}^{\text{BFGS}} := H_k - \frac{(H_k s^k)(H_k s^k)^T}{(s^k)^T H_k s^k} + \frac{y^k (y^k)^T}{(y^k)^T s^k} \quad (3),$$

the so-called “BFGS” formula. Another update formula that can be obtained that solves (2) is

$$H_{k+1}^{\text{DFP}} := H_k + \frac{(y^k - H_k s^k)(y^k)^T + y^k (y^k - H_k s^k)^T}{(y^k)^T s^k} - \frac{(y^k - H_k s^k)^T s^k}{[(h^k)^T s^k]^2} y^k (y^k)^T.$$

Note that any convex combination of formulas that satisfy (2) also satisfy (2); thus, we define the so-called *Broyden class* by the family of convex combinations of the above two formula,

$$H_{k+1}^\lambda := (1 - \lambda)H_{k+1}^{\text{DFP}} + \lambda H_{k+1}^{\text{BFGS}}, \quad \forall \lambda \in [0, 1].$$

Algorithmically, for $f \in C^1$;

| Globalized BFGS Method |
|--|
| <p>S0. Choose $x^0 \in \mathbb{R}^n$, $H_0 \in \mathbb{R}^{n \times n}$ symmetric positive definite, $\sigma \in (0, \frac{1}{2})$, $\rho \in (\sigma, 1)$, $\varepsilon \geq 0$ and set $k := 0$.</p> <p>S1. If $\ \nabla f(x^k)\ \leq \varepsilon$, STOP.</p> <p>S2. Determine d^k as a solution to the QNE,</p> $H_k d = -\nabla f(x^k).$ <p>S3. Determine $t_k > 0$ such that</p> $f(x^k + t_k d^k) \leq f(x^k) + \sigma t_k \nabla f(x^k)^T d^k,$ <p>(this is just the Armijo condition), AND</p> $\nabla f(x^k + t_k d^k)^T d^k \geq \rho \nabla f(x^k)^T d^k,$ <p>call the <i>Wolfe-Powell rule</i>.</p> <p>S4. Set</p> $\begin{aligned} x^{k+1} &:= x^k + t_k d^k, \\ s^k &:= x^{k+1} - x^k, \\ y^k &:= \nabla f(x^{k+1}) - \nabla f(x^k), \\ H_{k+1} &:= H_{k+1}^{\text{BFGS}}. \end{aligned}$ <p>S5. Increment k and go to S1.</p> |

We use the *Wolfe-Powell rule*; i.e., for $f : \mathbb{R}^n \rightarrow \mathbb{R}$ differentiable, $\sigma \in (0, \frac{1}{2})$, $\rho \in (\sigma, 1)$,

$$T_{\text{WP}} : \mathcal{A}_f \ni (x, d) \mapsto \left\{ t > 0 \mid \begin{aligned} f(x + td) &\leq f(x) + \sigma t \nabla f(x)^T d \\ \nabla f(x + td)^T d &\geq \rho \nabla f(x)^T d \end{aligned} \right\} \subset \mathbb{R}_+.$$

↪ **Lemma 3.4:** For $f \in C^1$ and $(x, d) \in \mathcal{A}_f$, assume that f is bounded from below on $\{x + td \mid t > 0\}$. Then, $T_{\text{WP}}(x, d) \neq \emptyset$.

Remark 3.10: Note that we didn't need any boundedness restriction for the well-definedness of the Armijo rule.

↪ **Lemma 3.5:** For $f \in C^1$, bounded from below with ∇f Lipschitz continuous on $\mathcal{L} := \text{lev}_{f(x^0)} f$. Then, T_{WP} restricted to $\mathcal{A}_f \cap (\mathcal{L} \times \mathbb{R}^n)$ is *efficient*, i.e. there exists a $\theta > 0$ such that $f(x + td) \leq f(x) - \theta \left(\frac{\nabla f(x)^T d}{\|\nabla f(x)\| \|d\|} \right)^2$ for every $(x, d) \in \mathcal{A}_f \cap (\mathcal{L} \times \mathbb{R}^n)$ and $t \in T_{\text{WP}}(x, d)$.

Remark 3.11: Note that, generally x^k will be in the level set at $f(x^0)$ for every $k \geq 0$ when x^k defined by a descent method. So in the context of this lemma, we will have the efficient bound at every iterate.

We turn to analyze Algorithm 3.6.

↪ **Lemma 3.6:** Let $y^k, s^k \in \mathbb{R}^n$ such that $(y^k)^T s^k > 0$ and $H_k > 0$. Then,

$$H_{k+1}^{\text{BFGS}} > 0.$$

PROOF. For fixed k , set $H_+ := H_{k+1}$, $H := H_k$, $s := s^k$ and $y := y^k$ for notational convenience. As $H > 0$, there exists a $W > 0$ such that $W^2 = H$. Let $d \in \mathbb{R}^n - \{0\}$ and set $z := Ws$, $v := Wd$. Then

$$\begin{aligned} d^T H_+ d &= d^T \left(H + \frac{yy^T}{y^T s} - \frac{Hss^T H}{s^T Hs} \right) d \\ &= d^T H d + d^T \frac{yy^T}{y^T s} d - d^T \frac{Hss^T H}{s^T Hs} d \\ &= d^T H d + \frac{(y^T d)^2}{y^T s} - \frac{(d^T Hs)^2}{s^T Hs} \\ &= \|v\|^2 + \frac{(y^T d)^2}{y^T s} - \frac{(v^T z)^2}{\|z\|^2} \\ &\geq \|v\|^2 + \frac{(y^T d)^2}{y^T s} - \|v\|^2 \\ &= \frac{(y^T d)^2}{y^T s} \geq 0, \end{aligned}$$

using Cauchy-Schwarz. In particular, equality (to zero) holds throughout iff v and z are linearly dependent and $y^T d = 0$. Suppose this is the case. In particular, there is an $\alpha \in \mathbb{R}$ for which $v = \alpha z$. Then, $d = W^{-1}v = \alpha W^{-1}z = \alpha s$, thus $0 = d^T y = \alpha s^T y$, hence α must equal zero, since we assumed $y^T s > 0$. Thus, $d = 0$, which we also assumed wasn't the case. Thus, we can never have equality here, and thus $d^T H_+ d > 0$, and so $H_+ > 0$. ■

↪ **Lemma 3.7:** If in the k th iteration of Algorithm 3.6 we have $H_k > 0$ and there exists $t_k \in T_{\text{WP}}(x^k, d^k)$, then $(s^k)^T y^k > 0$.

PROOF. We have

$$\begin{aligned}
(s^k)^T y^k &= (x^{k+1} - x^k)^T (\nabla f(x^{k+1}) - \nabla f(x^k)) \\
&= t_k (d^k)^T (\nabla f(x^{k+1}) - \nabla f(x^k)) \\
&\stackrel{\text{WP}}{\geq} t_k (\rho - 1) \nabla f(x^k)^T d^k \\
&= \underbrace{t_k(1 - \rho)}_{>0} \underbrace{\left(\frac{\nabla f(x^k)}{\neq 0} \right)^T}_{>0} H_k^{-1} \nabla f(x^k) \\
&> 0,
\end{aligned}$$

since $H_k^{-1} > 0$ and $t_k > 0$ and $0 < \rho < 1$. ■

↪ **Theorem 3.6:** Let $f \in C^1(\mathbb{R}^n)$ and bounded from below. Then, the following hold for the iterates generated by Algorithm 3.6:

1. $(s^k)^T y^k > 0$;
2. $H_k > 0$;
3. thus, Algorithm 3.6 is well-defined, i.e. at each iteration, each step generates a valid value.

PROOF. We prove inductively on k , with the fact that $H_0 > 0$ already establishing 2. for the base step. $H_k > 0$ implies the existence of a unique solution $d^k = -H_k^{-1} \nabla f(x^k)$ to QNE. Because $\nabla f(x^k) \neq 0$, $\nabla f(x^k)^T d^k < 0$ so $(x^k, d^k) \in A_f$. By [Lem. 3.4](#), there exists a $t_k \in T_{\text{WP}}(x^k, d^k)$. Thus, by [Lem. 3.7](#), $(y^k)^T s^k > 0$ and so by [Lem. 3.6](#) $H_{k+1} > 0$. Since this holds for any k this proves the result. ■

↪ **Theorem 3.7:** Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable, and $\{x^k\}, \{d^k\}, \{t_k\}$ be generated by Algorithm 3.6. assume that ∇f is Lipschitz on $\mathcal{L} := \text{lev}_f(x^0)f$, and that there exists a $c > 0$ such that $\text{cond}(H_k) := \frac{\lambda_{\max}(H_k)}{\lambda_{\min}(H_k)} \leq \frac{1}{c}$ for all $k \in \mathbb{N}$. Then every cluster point of $\{x^k\}$ is a stationary point of f .

PROOF. For all $k \in \mathbb{N}$,

$$\begin{aligned}
-\nabla f(x^k)^T d^k &= (d^k)^T H_k d^k \geq \lambda_{\min}(H_k) \|d^k\|^2 \\
&= \lambda_{\min}(H_k) \|H_k^{-1} \nabla f(x^k)\| \|d^k\| \\
&= \frac{\lambda_{\min}(H_k)}{\|H_k\|} \|H_k^{-1} \nabla f(x^k)\| \|d^k\| \\
&\geq \frac{\lambda_{\min}(H_k)}{\lambda_{\max}(H_k)} \|\nabla f(x^k)\| \|d^k\| \\
&= \frac{1}{\text{cond}(H_k)} \|\nabla f(x^k)\| \|d^k\| \\
&\geq c \|\nabla f(x^k)\| \|d^k\|,
\end{aligned}$$

and thus $-\frac{\nabla f(x^k)^T d^k}{\|\nabla f(x^k)\| \|d^k\|} \geq c$ for all $k \in \mathbb{N}$ (this is the so-called “angle condition”).

Moreover, under the assumptions on f , the Wolfe-Powell rule (restricted to $A_f \cap \mathcal{L} \times \mathbb{R}^n$, in which we always stay) is efficient, so by the previously established global convergence of Algorithm 3.1, we have convergence of this algorithm as well. ■

Remark 3.12: We cited the convergence of Algorithm 3.1, which we couldn’t do when proving convergence of the gradient, since the step size in that case was *not* efficient.

Remark 3.13:

1. The assumption that ∇f is Lipschitz on $\text{lev}_{f(x^0)} f$ is satisfied under either of the following conditions,

- (i) $f \in C^2$ and $\|\nabla^2 f(x)\|$ bounded on a convex superset of \mathcal{L} ;
- (ii) $f \in C^2$ and \mathcal{L} is bounded (hence compact).

An example of a C^1 function that is not C^2 but still globally Lipschitz is $f(x) := \frac{1}{2} \text{dist}_C^2(x)$ where C a convex set, and $\nabla f(x) = x - P_C(x)$ where P_C the projection onto C .

2. The BFGS method is regarded as one of the most robust methods for smooth, unconstrained optimization up to medium scale. For large-scale, there is a method called “limited memory BFGS”. Surprisingly, BFGS can be modified to work well for nonsmooth functions with a special line search method.

3.4.2 Inexact Methods

The local Newton’s method involves finding d^k such that $\nabla^2 f(x^k) d^k = -\nabla f(x^k)$. Quasi-Newton methods replace the Hessian with an approximation, and indirect methods further allow the flexibility to let d^k approximately solve this equation (since solving this equation exactly can be costly). The goal is to find d^k such that

$$\frac{\|\nabla^2 f(x^k) d + \nabla f(x^k)\|}{\|\nabla f(x^k)\|} \leq \eta_k$$

for a prescribed tolerance η_k . This is called the *inexact Newton’s equation*.

Remark 3.14: Dividing by $\|\nabla f(x^k)\|$ here enforces the idea that the closer x^k is to a stationary point, the higher accuracy we require.

| Local Inexact Newton's Method |
|--|
| S0. Choose $x^0 \in \mathbb{R}^n, \varepsilon \geq 0$ and set $k := 0$. |
| S1. If $\ \nabla f(x^k)\ \leq \varepsilon$, STOP. |
| S2. Choose $\eta_k \geq 0$ and determine d^k such that |
| $\frac{\ \nabla^2 f(x^k) d + \nabla f(x^k)\ }{\ \nabla f(x^k)\ } \leq \eta_k.$ |
| S3. Set $x^{k+1} = x^k + d^k$, increment k and go to S1. |

↪ **Theorem 3.8:** Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be C^2 , let \bar{x} be a stationary point of f such that $\nabla^2 f(\bar{x})$ is invertible. Then there exists $\varepsilon > 0$ such that for all $x^0 \in B_\varepsilon(\bar{x})$:

1. If $\eta_k \leq \bar{\eta}$ for all $k \in \mathbb{N}$ for some $\bar{\eta} > 0$ sufficiently small, then Algorithm 3.7 is well-defined and generates a sequence that converges at least linearly to \bar{x} .
2. If $\eta_k \downarrow 0$, the rate of convergence is superlinear.
3. If $\nabla^2 f$ is locally Lipschitz (for instance, if $f \in C^3$) and $\eta_k = O(\|\nabla f(x^k)\|)$, then the rate is quadratic.

Remark 3.15: For $\eta_k = 0$, we just recover the local Newton's method. 2. and 3. strongly point their fingers at how to choose η_k . 1. is theoretically important, but practically useless since $\bar{\eta}$ is generally unknown.

| Globalized Inexact Newton's Method |
|--|
| <p>S0. Choose $x^0 \in \mathbb{R}^n, \varepsilon \geq 0, \rho > 0, p > 2, \beta \in (0, 1), \sigma \in (0, \frac{1}{2})$ and set $k := 0$.</p> <p>S1. If $\ \nabla f(x^k)\ \leq \varepsilon$ STOP.</p> <p>S2. Choose $\eta_k \geq 0$ and determine d^k by</p> $\ \nabla^2 f(x^k)d + \nabla f(x^k)\ \leq \eta_k \ \nabla f(x^k)\ .$ <p>If this is not possible, or not feasible, i.e. $\nabla f(x^k)^T d^k \leq -\rho \ d^k\ ^p$ is violated, then set $d^k := -\nabla f(x^k)$.</p> <p>S3. Determine $t_k > 0$ by Armijo, $t_k := \max_{\{\ell \in \mathbb{N}_0\}} \left\{ \beta^\ell \mid f(x^k + \beta^\ell d^k) \leq f(x^k) + \beta^\ell \sigma \nabla f(x^k)^T d^k \right\}$.</p> <p>S4. Set $x^{k+1} = x^k + t_k d^k$, increment k and go to S1.</p> |

↪ **Theorem 3.9:** Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be C^2 and let $\{x^k\}$ be generated by Algorithm 3.8 with $\eta_k \downarrow 0$. If \bar{x} is a cluster point of $\{x^k\}$ with $\nabla^2 f(\bar{x}) > 0$, then the following hold:

1. $\{x^k\}$ converges along the whole sequence to \bar{x} , which is a strict locally minimizer of f .
2. For all k sufficiently large, d^k will be given by the inexact Newton equation.
3. For all k sufficiently large, the full step-size $t_k = 1$ will be accepted.
4. The convergence is at least superlinear.

§3.5 Conjugate Gradient Methods for Nonlinear Optimization

3.5.1 Prelude: Linear Systems

Remark that, for $A > 0$ and $b \in \mathbb{R}^n$,

$$Ax = b \quad \Leftrightarrow \quad x \text{ minimizes } f(x) := \frac{1}{2} x^T A x - b^T x.$$

↪ **Definition 3.6** (A -conjugate vectors): Let $A > 0$ and $x, y \in \mathbb{R}^n \setminus \{0\}$ are called A -conjugate if

$$x^T A y = 0$$

(i.e. x, y are orthogonal in the inner product induced by A , denoted $\langle \cdot, \cdot \rangle_A$).

↪ **Lemma 3.8:** Let $A > 0, b \in \mathbb{R}^n$, and $f(x) := \frac{1}{2}x^T Ax - b^T x$. Let d^0, d^1, \dots, d^{n-1} be (pairwise) A -conjugate. Let $\{x^k\}$ be generated by $x^{k+1} = x^k + t_k d^k, x^0 \in \mathbb{R}^n$, where $t_k := \operatorname{argmin}_{t \geq 0} f(x^k + t d^k)$. Then, after *at most* n iterations, x^n is the (unique) global minimizer $\bar{x} (= A^{-1}b)$ of f . Moreover, with $g^k := \nabla f(x^k) (= Ax^k - b)$, we have

$$t_k = -\frac{(g^k)^T d^k}{(d^k)^T A d^k} > 0,$$

and $(g^{k+1})^T d^j = 0$ for all $j = 0, \dots, k$.