MATH378 - Nonlinear Optimization

Based on lectures from Fall 2025 by Prof. Tim Hoheisel.
Notes by Louis Meunier

Contents

I Preliminaries 2
I.1 Terminology 2
I.2 Convex Sets and Functions 3
II Unconstrained Optimization 4
II.1 Theoretical Foundations 4
II.1.1 Quadratic Approximation 6
II.2 Differentiable Convex Functions 7
II.3 Matrix Norms 9
II.4 Descent Methods 11
II.4.1 A General Line-Search Method 11
II.4.1.1 Global Convergence of Algorithm 2.1 12
II.4.2 The Gradient Method 13
II.4.3 Newton-Type Methods 15
II.4.3.1 Convergence Rates and Landau Notation 15
II.4.3.2 Newton's Method for Nonlinear Equations 16
II.4.3.3 Newton's Method for Optimization Problem 18
II.4.4 Quasi-Newton Methods 21
II.4.4.1 Direct Methods 21
II.4.4.2 Inexact Methods 26
II.4.5 Conjugate Gradient Methods for Nonlinear Optimization 27
II.4.5.1 Prelude: Linear Systems 27
II.4.6 The Fletcher-Reeves Method 30
II.5 Least-Squares Problems 33
II.5.1 Linear Least-Squares 33
II.5.2 Gauss-Newton for Nonlinear Least-Squares 33
III Constrained Optimization 34
III.1 Optimality Conditions for Constrained Problems 34
III.1.1 First-Order Optimality Conditions 35
III.1.2 Farkas' Lemma 36

§I Preliminaries

§I.1 Terminology

We consider problems of the form

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

with $X \subset \mathbb{R}^n$ the *feasible region* with x a *feasible point*, and $f : X \to \mathbb{R}$ 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 $y_1, \ldots, y_m \in \mathbb{R}$ measurements taken at m distinct points $x_1, \ldots, x_m \in \mathbb{R}$, the goal is to find a degree $\leq n$ polynomial $q : \mathbb{R} \to \mathbb{R}$, 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

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

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

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

then concisely this problem is equivalent to

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

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 $\overline{x}$ such that $f(\overline{x}) = \inf_X f$, i.e. the value of a point

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

   (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 $f(\overline{x})$.

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 minimimum; 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 $f : \mathbb{R}^n \to \mathbb{R}$. 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 $f : \mathbb{R}^n \to \mathbb{R}$.

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

§I.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, $f : C \to \mathbb{R}$ 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**: $\exp : \mathbb{R} \to \mathbb{R}$, $\log : (0,\infty) \to \mathbb{R}$ are convex. A function $f : \mathbb{R}^n \to \mathbb{R}^m$ 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 $i = 1, \ldots, m$, 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 $f : \mathbb{R}^n \to \mathbb{R}$ be convex and $G : \mathbb{R}^m \to \mathbb{R}^n$ be affine. Then, $f \circ G$ is convex on $\mathbb{R}^m$.

§II Unconstrained Optimization

§II.1 Theoretical Foundations

We focus on the problem
$$\min_{x\in\mathbb{R}^n} f(x),$$
where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable.

↪**Definition 2.1** (Directional derivative): Let $D \subset \mathbb{R}^n$ be open and $f : D \to \mathbb{R}$. We say f *directionally differentiable* at $\bar{x} \in D$ in the direction $d \in \mathbb{R}^n$ if
$$\lim_{t\to 0^+} \frac{f(\bar{x} + td) - f(\bar{x})}{t}$$
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 $f : D \to \mathbb{R}$ 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 $f : \mathbb{R}^n \to \mathbb{R}$ 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

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

using homogeneity of the norm.

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

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

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

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

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

Moreover, since X open, then for t even smaller (if necessary), $\overline{x} + td$ remains in X, thus $\overline{x}$ 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 $\overline{x}$. Then, $\nabla f(\overline{x}) = 0$.

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

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

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

We recall the following from Calculus:

↪**Theorem 2.2** (Taylor's, Second Order): Let $f : D \subset \mathbb{R}^n \to \mathbb{R}$ 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 $f : X \to \mathbb{R}$ 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 $t \to 0^+$, $\nabla^2 f(\eta_t) \to \nabla^2 f(x) < 0$. 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, $f : X \to \mathbb{R}$ 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

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

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.

II.1.1 Quadratic Approximation

Let $f : \mathbb{R}^n \to \mathbb{R}$ 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

$$f : \mathbb{R}^n \to \mathbb{R}, \qquad f(x) = \frac{1}{2}x^T Q x + c^T x + \gamma,$$

a typical quadratic function. Then,

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

We find that f has *no* minimizer if $c \notin \mathrm{rge}(Q)$ 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).

§II.2 Differentiable Convex Functions

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

1. f is convex (on C) iff

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

   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

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

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

Proof. (1., ⇒) 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 $\lambda \to 0^+$, the LHS → 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., ⇐) 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., ⇒) 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 $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable. Then,
a) there exists an *affine function* $g : \mathbb{R}^n \to \mathbb{R}$ such that $g(x) \leq f(x)$ everywhere;
b) if f strongly convex, then it is coercive, i.e. $\lim_{\|x\| \to \infty} f(x) = \infty$.

↪**Corollary 2.2**: Let $f : \mathbb{R}^n \to \mathbb{R}$ 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

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

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 $f : \mathbb{R}^n \to \mathbb{R}$ be convex and continuous, $X \subset \mathbb{R}^n$ convex (and nonempty), and consider the optimization problem

$$\min f(x) \text{ s.t. } x \in X \qquad (\star).$$

Then, the following hold:

1. $\bar{x}$ is a global minimizer of $(\star)$ $\Leftrightarrow$ $\bar{x}$ is a local minimizer of $(\star)$
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 $\lambda \to 1^-$, we see that $\lambda\bar{x} + (1 - \lambda)\hat{x} \to \bar{x}$; 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 $(\star)$ and

$$\min_{x \in \operatorname{lev}_c f \cap X} f(x) \qquad (\star\,\star)$$

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.

§II.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 $\mathbb{R}^n \to \mathbb{R}^m$).

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

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

↪**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

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

PROOF. We have for any m,

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

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 $A_m \to A \in \mathbb{R}^{n \times n}$ (by relabelling). Moreover, observe that

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

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

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

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 $\|B^{-1}\| \leq \frac{\|A\|}{1-\|I-BA\|}$.

§II.4 Descent Methods

II.4.1 A General Line-Search Method

We deal with the unconstrained problem

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

↪**Definition 2.2** (Descent Direction): Let $f : \mathbb{R}^n \to \mathbb{R}$, $x \in \mathbb{R}^n$. $d \in \mathbb{R}^n$ 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 2.5**: If $f : \mathbb{R}^n \to \mathbb{R}$ 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 2.6**: Let $f : \mathbb{R}^n \to \mathbb{R}$ 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 $\nabla f\left(x^k\right)^T d^k < 0$
S4. (Step-size) Determine $t_k > 0$ such that $f\left(x^k + t_k d^k\right) < f\left(x^k\right)$
S5. (Update) Set $x^{k+1} := x^k + t_k d^k$, iterate k, and go back to step 2.

Remark 2.2: a) The generic choice for d^k in 3. is just $d^k := -B_k \nabla f\left(x^k\right)$ for some $B_k > 0$. We focus on:

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

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

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

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* $t_k = \operatorname{argmin}_{t>0} f\left(x_k + td^k\right)$ overemphasizes (i) over (ii).

↪**Definition 2.3** (Step-size rule): Let $f \in C^1(\mathbb{R}^n)$ and

$$\mathcal{A}_f := \left\{(x, d) \mid \nabla f(x)^T d < 0\right\}.$$

A (possible set-valued) map

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

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

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

II.4.1.1 Global Convergence of Algorithm 2.1

↪**Definition 2.4** (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

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

↪**Theorem 2.8**: Let $f \in C^1(\mathbb{R}^n)$. Let $\left\{x^k\right\}, \left\{d^k\right\}, \{t_k\}$ be generated by Algorithm 2.1. Assume the following:

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

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

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

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

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

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

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

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

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

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

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

II.4.2 The Gradient Method

We specialize Algorithm 2.1 here. Specifically, we'll take

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

it's known that

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

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

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

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

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

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

⊛ **Example 2.3** (Illustration of Armijo Rule): For $(x, d) \in \mathcal{A}_f$ and f smooth on $\mathbb{R}^n$, defined $\phi : \mathbb{R} \to \mathbb{R}, \phi(t) := f(x + td)$. The map $t \mapsto \sigma\varphi'(0)t + \varphi(0) = \sigma t \nabla f(x)^T d + \varphi(0)$

↪**Proposition 2.6**: Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable with $\beta, \sigma \in (0,1)$. Then for $(x, d) \in \mathcal{A}_f$, there exists $\ell \in \mathbb{N}_0$ such that

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

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

Proof. Suppose not, i.e.

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

Letting $\ell \to \infty$, 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 $(x, d) \in \mathcal{A}_f$, 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 $\left\Vert \nabla f\left(x^k\right)\right\Vert \leq \varepsilon$, STOP
S2. Set $d^k := -\nabla f\left(x^k\right)$
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 2.4**: Let $f \in C^1(\mathbb{R}^n)$, $x^k \to x$, $d^k \to d$ and $t_k \downarrow 0$. Then

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

Proof. Left as an exercise. ■

↪**Theorem 2.9**: Let $f \in C^1(\mathbb{R}^n)$. Then every cluster point of a sequence $\left\{x^k\right\}$ generated by Algorithm 2.2 is a stationary point of f.

Proof. Let $\overline{x}$ be a cluster point of $\left\{x^k\right\}$ and let $x^k \underset{k\in K}{\to} \overline{x}$, K an infinite subset of $\mathbb{N}$. Assume towards a contradiction $\nabla f(\overline{x}) \neq 0$. As $f\left(x^k\right)$ is monotonically decreasing with cluster point $f(\overline{x})$, it must be that $f\left(x^k\right) \to f(\overline{x})$ along the whole sequence so $f\left(x^{k+1}\right) - f\left(x^k\right) \to 0$. Thus,

$$0 \le t_k \left\|\nabla f\left(x^k\right)\right\|^2 \overset{S2}{=} -t_k \nabla f\left(x^k\right)^T d^k \overset{S3}{\le} \frac{f\left(x^k\right) - f\left(x^{k+1}\right)}{\sigma} \to 0.$$

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

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

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

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

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

$$-\|\nabla f(\overline{x})\|^2 \ge \sigma \|\nabla f(\overline{x})\|^2,$$

which implies σ negative, a contradiction. ■

Remark 2.3: The proof above shows, the following: Let $\left\{x^k\right\}$ such that $x^{k+1} := x^k + t_k d^k$ for $d^k \in \mathbb{R}^n, t_k > 0$, and let $f\left(x^{k+1}\right) \le f\left(x^k\right)$ and $x^k \overset{K}{\to} \overline{x}$ such that $d^k = -\nabla f\left(x^k\right)$, $t_k = T_A\left(x^k, d^k\right)$ for all $k \in K$. Then $\nabla f(\overline{x}) = 0$; 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 $f\left(x^k\right) \to f(\overline{x})$ along the whole sequence.

II.4.3 Newton-Type Methods

II.4.3.1 Convergence Rates and Landau Notation

↪**Definition 2.5**: Let $\left\{x^k \in \mathbb{R}^n\right\}$ converge to $\overline{x}$. Then, $\left\{x^k\right\}$ converges:

1. *linearly* to $\overline{x}$ if there exists $c \in (0,1)$ such that
$$\left\|x^{k+1} - \overline{x}\right\| \le c\left\|x^k - \overline{x}\right\|, \forall k;$$
2. *superlinearly* to $\overline{x}$ if
$$\lim_{k\to\infty} \frac{\left\|x^{k+1} - \overline{x}\right\|}{\left\|x^k - \overline{x}\right\|} = 0;$$
3. *quadratically* to $\overline{x}$ if there exists $C > 0$ such that
$$\left\|x^{k+1} - \overline{x}\right\| \le C\left\|x^k - \overline{x}\right\|^2, \forall k.$$

Remark 2.4: 3. ⇒ 2. ⇒ 1.

Remark 2.5: We needn't assume $x^k \to \overline{x}$ 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 2.6** (Landau Notation): Let $\{a_k\}, \{b_k\}$ be positive sequences $\downarrow 0$. Then,
1. $a_k = o(b_k) \Leftrightarrow \lim_{k\to\infty} \frac{a_k}{b_k} = 0$;
2. $a_k = O(b_k) \Leftrightarrow \exists C > 0 : a_k \leq C b_k$ for all k (sufficiently large).

Remark 2.6: If $x^k \to \bar{x}$, then
1. the convergence is superlinear $\Leftrightarrow \left\|x^{k+1} - \bar{x}\right\| = o\left(\left\|x^k - \bar{x}\right\|\right)$;
2. the convergence is quadratic $\Leftrightarrow \left\|x^{k+1} - \bar{x}\right\| = O\left(\left\|x^k - \bar{x}\right\|^2\right)$.

II.4.3.2 Newton's Method for Nonlinear Equations

We consider the nonlinear equation

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

where $F : \mathbb{R}^n \to \mathbb{R}^n$ 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,

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

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

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

More generally, one solves $F'\left(x^k\right)d = -F\left(x^k\right)$ 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 $\left\|F\left(x^k\right)\right\| < \varepsilon$, STOP.
S2. Compute d^k as a solution of *Newton's equation* $$F'\left(x^k\right)d = -F\left(x^k\right).$$
S3. Set $x^{k+1} := x^k + d^k$, increment k and go to S1.

↪**Lemma 2.5**: Let $F : \mathbb{R}^n \to \mathbb{R}^n$ 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

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

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 $x \in B_\varepsilon(\bar{x})$,

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

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

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

■

Remark 2.7: Observe $F : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\overline{x}$ if and only if $\left\|F\left(x^k\right) - F(\overline{x}) - F'(\overline{x})\left(x^k - \overline{x}\right)\right\| = o\left(\left\|x^k - \overline{x}\right\|\right)$ for every $x^k \to \overline{x}$.

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

↪**Lemma 2.6**: Let $F : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and $x^k \to \overline{x}$, then:
1. $\left\|F\left(x^k\right) - F(\overline{x}) - F'\left(x^k\right)\left(x^k - \overline{x}\right)\right\| = o\left(\left\|x^k - \overline{x}\right\|\right)$;
2. if F' locally Lipschitz at $\overline{x}$, then $\left\|F\left(x^k\right) - F(\overline{x}) - F'\left(x^k\right)\left(x^k - \overline{x}\right)\right\| = O\left(\left\|x^k - \overline{x}\right\|^2\right)$.

Proof.

1. Observe that

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

The left-hand term is $o\left(\left\|x^k - \overline{x}\right\|\right)$ 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 $\overline{x}$. Then,

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

which implies the result.

■

↪**Theorem 2.10**: Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable, $\overline{x} \in \mathbb{R}^n$ such that $F(\overline{x}) = 0$ and $F'(\overline{x})$ is invertible. Then, there exists an $\varepsilon > 0$ such that for every $x^0 \in B_\varepsilon(\overline{x})$, we have:

1. Algorithm 2.3 is well-defined and generates a sequence $\{x^k\}$ which converges to $\overline{x}$;
2. the rate of convergence is (at least) linear;
3. if F' is locally Lipschitz at $\overline{x}$, 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 $\|F(x) - F(\overline{x}) - F'(x)(x - \overline{x})\| \leq \frac{1}{2c}\|x - \overline{x}\|$ for all $x \in B_{\varepsilon_2}(\overline{x})$. Take $\varepsilon = \min\{\varepsilon_1, \varepsilon_2\}$ and pick $x^0 \in B_\varepsilon(\overline{x})$. Then, x^1 is well-defined, since $F'(x^0)$ is invertible, and so

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

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

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

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

2., 3. Analogous to 1.,

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

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

■

II.4.3.3 Newton's Method for Optimization Problem

Consider

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

with $f : \mathbb{R}^n \to \mathbb{R}$ twice continuously differentiable. Recall that if $\overline{x}$ a local minimizer of f, $\nabla f(\overline{x}) = 0$. 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 $\left\|\nabla f\left(x^k\right)\right\| < \varepsilon$, STOP. S2. Compute d^k as a solution of *Newton's equation* $$\nabla^2 f\left(x^k\right) d = -\nabla f\left(x^k\right).$$ 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 2.4**: 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 2.4, S2) reads in this case:

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

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

$$|x_0| < 1 \Rightarrow x_k \to 0, \text{quadratically}$$
$$|x_0| > 1 \Rightarrow x_k \text{ diverges}$$
$$|x_0| = 1 \Rightarrow |x_k| = 1 \forall k,$$

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 $d^k := -\nabla f\left(x^k\right)$;
- 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 $\left\|\nabla f\left(x^k\right)\right\| < \varepsilon$, STOP S2. Determine d^k as a solution of $$\nabla^2 f\left(x^k\right) d = -\nabla f\left(x^k\right).$$ If no solution exists, or if $\nabla f\left(x^k\right)^T d^k \leq -\rho\left\|d^k\right\|^p$, is violated, set $d^k := -\nabla f\left(x^k\right)$ S3. Determine $t_k > 0$ by the Armijo back-tracking rule, i.e. $$t_k := \max_{\ell \in \mathbb{N}_0}\left\{\beta^\ell \mid f\left(x^k + \beta^\ell d^k\right) \leq f\left(x^k\right) + \beta^\ell \sigma \nabla f\left(x^k\right)^T d^k\right\}$$ S4. Set $x^{k+1} := x^k + t_k d^k$, increment k to $k + 1$, and go back to S1.

Remark 2.8: 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. $\left(x^k, d^k\right) \in \mathcal{A}_f$ for every k.

↪**Theorem 2.11** (Global convergence of Algorithm 2.5): Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. Then every cluster point of $\left\{x^k\right\}$ generated by Algorithm 2.5 is a stationary point of f.

Remark 2.9: Note that we didn't impose any invertibility condition on the Hessian of f; indeed, if say the hessian was nowhere invertible, then Algorithm 2.5 just becomes the gradient method with Armijo back-tracking, for which have already established this result.

Proof. Let $\left\{x^k\right\}$ be generated by Algorithm 2.5, with $\left\{x^k\right\}_K \to \bar{x}$. If $d^k := -\nabla f\left(x^k\right)$ for infinitely many $k \in K$ (i.e. along a subsubsequence of $\left\{x^k\right\}$), 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,

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

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

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

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

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

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

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

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

Taking $k \to \infty$ along K, we see that $t_k \left\|d^k\right\|^p \to 0$ 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, $t_k \to 0$. By the Armijo rule, $t_k = \beta^{\ell_k}$, for $\ell_k \in \mathbb{N}_0$, uniquely determined. Since $t_k \to 0$, then $\ell_k \to \infty$. On the other hand, by S3,

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

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

↪**Theorem 2.12** (Fast local convergence of Algorithm 2.5): Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable, $\left\{x^k\right\}$ generated by Algorithm 2.5. If $\overline{x}$ is a cluster point of $\left\{x^k\right\}$ with $\nabla^2 f(\overline{x}) \succ 0$. Then:

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

II.4.4 Quasi-Newton Methods

In Newton's, in general we need to find

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

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

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

II.4.4.1 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)

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

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

Remark 2.10: Suppose x^k a current iterate for an algorithm to minimize $f : \mathbb{R}^n \to \mathbb{R}$ for $f \in C^2$.

1. $\nabla^2 f\left(x^k\right)$ 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,

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

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

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

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

So, the ansatz we take is

$$y^k = H_{k+1} s^k = H_k s^k + \gamma u u^T s^k + \delta v v^T s^k. \qquad (2)$$

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

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

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

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

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,

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

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

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

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

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

↪**Lemma 2.7**: 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 2.11: Note that we didn't need any boundedness restriction for the well-definedness of the Armijo rule.

↪**Lemma 2.8**: 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 2.12: Note that, generally x^k will be in the level set at $f\left(x^0\right)$ 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 2.6.

↪**Lemma 2.9**: Let $y^k, s^k \in \mathbb{R}^n$ such that $\left(y^k\right)^T s^k > 0$ 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

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

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 2.10**: If in the kth iteration of Algorithm 2.6 we have $H_k > 0$ and there exists $t_k \in T_{\text{WP}}\left(x^k, d^k\right)$, then $\left(s^k\right)^T y^k > 0$.

Proof. We have

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

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

↪**Theorem 2.13**: Let $f \in C^1(\mathbb{R}^n)$ and bounded from below. Then, the following hold for the iterates generated by Algorithm 2.6:

1. $\left(s^k\right)^T y^k > 0$;
2. $H_k > 0$;
3. thus, Algorithm 2.6 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 $d^k = -H_k^{-1}\nabla f\left(x^k\right)$ to QNE. Because $\nabla f\left(x^k\right) \neq 0$, $\nabla f\left(x^k\right)^T d^k < 0$ so $\left(x^k, d^k\right) \in \mathcal{A}_f$. By Lem. 2.7, there exists a $t_k \in T_{\text{WP}}\left(x^k, d^k\right)$. Thus, by Lem. 2.10, $\left(y^k\right)^T s^k > 0$ and so by Lem. 2.9 $H_{k+1} > 0$. Since this holds for any k this proves the result. ■

↪**Theorem 2.14**: Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, and $\left\{x^k\right\}, \left\{d^k\right\}, \{t_k\}$ be generated by Algorithm 2.6. 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 $\left\{x^k\right\}$ is a stationary point of f.

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

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

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 $\mathcal{A}_f \cap \mathcal{L} \times \mathbb{R}^n$, in which we always stay) is efficient, so by the previously established global convergence of Algorithm 2.1, we have convergence of this algorithm as well. ■

Remark 2.13: We cited the convergence of Algorithm 2.1, which we couldn't do when proving convergence of the gradient, since the step size in that case was *not* efficient.

Remark 2.14:
1. The assumption that ∇f is Lipschitz on $\operatorname{lev}_{f(x^0)} f$ 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 $f(x) := \frac{1}{2}\operatorname{dist}_C^2(x)$ 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.

II.4.4.2 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 2.15: 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 2.15**: Let $f : \mathbb{R}^n \to \mathbb{R}$ 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 2.7 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 $\eta_k = O\left(\left\|\nabla f\left(x^k\right)\right\|\right)$, then the rate is quadratic.

Remark 2.16: 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
S0. Choose $x^0 \in \mathbb{R}^n, \varepsilon \geq 0, \rho > 0, p > 2, \beta \in (0,1), \sigma \in \left(0, \frac{1}{2}\right)$ and set $k := 0$. S1. If $\left\|\nabla f\left(x^k\right)\right\| \leq \varepsilon$ STOP. S2. Choose $\eta_k \geq 0$ and determine d^k by $$\left\|\nabla^2 f\left(x^k\right) d + \nabla f\left(x^k\right)\right\| \leq \eta_k \left\|\nabla f\left(x^k\right)\right\|.$$ If this is not possible, or not feasible, i.e. $\nabla f\left(x^k\right)^T d^k \leq -\rho \left\|d^k\right\|^p$ is violated, then set $d^k := -\nabla f\left(x^k\right)$. S3. Determine $t_k > 0$ by Armijo, $t_k := \max_{\{\ell \in \mathbb{N}_0\}} \left\{\beta^\ell \mid f\left(x^k + \beta^\ell d^k\right) \leq f\left(x^k\right) + \beta^\ell \sigma \nabla f\left(x^k\right)^T d^k\right\}$. S4. Set $x^{k+1} = x^k + t_k d^k$, increment k and go to S1.

↪**Theorem 2.16**: Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^2 and let $\left\{x^k\right\}$ be generated by Algorithm 2.8 with $\eta_k \downarrow 0$. If $\bar{x}$ is a cluster point of $\left\{x^k\right\}$ with $\nabla^2 f(\bar{x}) > 0$, then the following hold:

1. $\left\{x^k\right\}$ 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.

II.4.5 Conjugate Gradient Methods for Nonlinear Optimization

II.4.5.1 Prelude: Linear Systems

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

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

↪**Definition 2.7** (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 2.11**: Let $A > 0, b \in \mathbb{R}^n$, and $f(x) := \frac{1}{2}x^T Ax - b^T x$. Let $d^0, d^1, ..., d^{n-1}$ 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 $t_k := \operatorname{argmin}_{t>0} f(x^k + t_k d^k)$. Then, after *at most* n iterations, x^n is the (unique) global minimizer $\overline{x}$ $(= A^{-1}b)$ of f. Moreover, with $g^k := \nabla f(x^k)$ $(= Ax^k - b)$, we have

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

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

Proof. The formula for t_k was proven in an exercise. To prove the latter statement, note that

$$\begin{aligned} (g^{k+1})^T d^k &= (Ax^{k+1} - b)^T d^k \\ &= (Ax^k - b + t_k Ad^k)^T d^k \\ &= (g^k)^T d^k + t_k (d^k)^T Ad^k \\ &= (g^k)^T d^k - (g^k)^T d^k = 0. \end{aligned}$$

Moreover, for all $i, j = 0, ..., k$ with $i \neq j$, we have that

$$(g^{i+1} - g^i)^T d^j = (Ax^{i+1} - Ax^i)^T d^j = t_i (d^i)^T Ad^j = 0,$$

hence for all $j = 0, ..., k$,

$$(g^{k+1})^T d^j = (g^{j+1})^T d^j + \sum_{i=j+1}^{k} (g^{i+1} - g^i)^T d^j = 0.$$

Thus, g^n is orthogonal to the n linearly independent vectors $\{d^0, ..., d^{n+1}\}$, which implies $g^n = 0$, thus proving the conclusion. ■

We want to obtain these A-conjugate vectors, while simultaneously ensuring that they are descent directions at each step, i.e. that $\nabla f(x^k)^T d^k < 0$ for all $k = 0, ..., n-1$. We do this algorithmically.

Assume $\nabla f(x^0) \neq 0$ (else we are done), and take $d^0 := -\nabla f(x^0)$. Suppose then we have $l + 1$ A-conjugate vectors $d^0, ..., d^l$ with $\nabla f(x^i)^T d^i < 0$ for each i. Suppose

$$d^{l+1} := -g^{l+1} + \sum_{i=0}^{l} \beta_{il} d^i,$$

where g^{l+1} is "valid" to be used since it is not in the span of $\{d^0, ..., d^l\}$, and $\{\beta_{il}\}$ are scalars to be determined. The condition $(d^{l+1})^T Ad^j = 0$ readily implies that

$$\beta_{jl} := \frac{(g^{l+1})^T Ad^j}{(d^j)^T Ad^j}, \quad j = 0, ..., l.$$

Then, it follows that $\left(g^{l+1}\right)^T d^{l+1} = -\left\|g^{l+1}\right\|^2 < 0$, and since $g^{l+1} = \nabla f\left(x^{l+1}\right)$ by definition, it follows d^{l+1} a descent direction. Thus, it must be that

$$g^{j+1} - g^j = Ax^{j+1} - Ax^j = t_j A d^j,$$

and so with $t_j > 0$,

$$\left(g^{l+1}\right)^T A d^j = \frac{1}{t_j}\left(g^{l+1}\right)^T \left(g^{j+1} - g^j\right),$$

and thus

$$\beta_{jl} = \begin{cases} 0 & j = 0, \dots, l-1 \\ \frac{\left\|g^{j+1}\right\|^2}{\left\|g^l\right\|^2} & j = l \end{cases},$$

and thus our update of d^{l+1} reduces to

$$d^{l+1} := -g^{l+1} + \beta_l d^l, \qquad \beta_l := \beta_{ll}.$$

In summary, this gives the following algorithm.

CG method for linear equations
S0. Choose $x^0 \in \mathbb{R}^n$ and $\varepsilon \geq 0$, set $g^0 := Ax^0 - b$, $d^0 := -g^0$ and initiate $k = 0$. S1. If $\left\|g^k\right\| \leq \varepsilon$, STOP. S2. Put $$t_k := \frac{\left\|g^k\right\|^2}{\left(d^k\right)^T A d^k}.$$ S3. Set $$x^{k+1} = x^k + t_k d^k$$ $$g^{k+1} = g^k + t_k A d^k$$ $$\beta_k = \frac{\left\|g^{k+1}\right\|^2}{\left\|g^k\right\|^2}$$ $$d^{k+1} = -g^{k+1} + \beta_k d^k.$$ S4. Increment k and go to S1.

↪**Theorem 2.17** (Convergence of CG Method): Let $A \in \mathbb{R}^{n \times n}$ be symmetric positive definite, $b \in \mathbb{R}^n$ and $f(x) := \frac{1}{2}x^T A x - b^T x$. Then, Algorithm 2.9 will produce the global miniumum $\bar{x}$ of f after at most n interations. If $m \in \{0, \dots, n\}$ is the smallest number such that $x^m = \bar{x}$, then the following hold as well:

$$\left(d^k\right)^T A d^j = 0, \left(g^k\right)^T g^j = 0, \left(g^k\right)^T d^j = 0, \qquad (k = 1, \dots, m, j = 0, \dots, k-1),$$

$$\left(g^k\right)^T d^k = -\left\|g^k\right\|^2 \qquad (k = 0, \dots, m).$$

II.4.6 The Fletcher-Reeves Method

We want to apply the same method as the previous section for non-quadratic and non-convex functions. The isue we need to resolve, though, is that the step-size rule in S2. of Algorithm 2.9 is no longer appropriate. To resolve, we introduce the *Strong Wolfe-Powell rule.* Choose $\sigma \in (0,1), \rho \in (\sigma, 1)$. The strong WP rule for a differentiable $f : \mathbb{R}^n \to \mathbb{R}$ reads

$$T_{\mathrm{SWP}} : (x,d) \in \mathcal{A}_f \mapsto \left\{ t > 0 \;\middle|\; \begin{matrix} f(x+td) \leq f(x) + \sigma t \nabla f(x)^T d \\ |\nabla f(x+td)^T d| \leq -\rho \nabla f(x)^T d \end{matrix} \right\},$$

noting that clearly $T_{\mathrm{SWP}}(x,d) \subset T_{\mathrm{WP}}(x,d)$.

Fletcher-Reeves
S0. Choose $x^0 \in \mathbb{R}^n$, $\varepsilon \geq 0$, $0 < \sigma < \rho < \frac{1}{2}$, set $d^0 := -\nabla f\left(x^0\right)$ and $k = 0$.
S1. If $\left\Vert \nabla f\left(x^k\right) \right\Vert \leq \varepsilon$, STOP.
S2. Determine $t_k > 0$ such that $$f\left(x^k + t_k d^k\right) \leq f\left(x^k\right) + \sigma t_k \nabla f\left(x^k\right)^T d^k,$$ $$\lvert\nabla f\left(x^k + t_k d^k\right)^T d^k\rvert \leq -\rho \nabla f\left(x^k\right)^T d^k.$$
S3. Set $$x^{k+1} = x^k + t_k d^k$$ $$\beta_k^{\mathrm{FR}} = \frac{\left\Vert \nabla f\left(x^{k+1}\right) \right\Vert^2}{\left\Vert \nabla f\left(x^k\right) \right\Vert^2}$$ $$d^{k+1} = -\nabla f\left(x^{k+1}\right) + \beta_k^{\mathrm{FR}} d^k.$$
S4. Increment k and go to S1.

↪**Lemma 2.12**: Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^1 and bounded from below and $(x,d) \in \mathcal{A}_f$. Then $T_{\mathrm{SWP}}(x,d) \neq \emptyset$.

Proof. Define $\varphi, \psi : \mathbb{R} \to \mathbb{R}$ by

$$\varphi(t) := f(x+td), \qquad \psi(t) = f(x) + \sigma t \nabla f(x)^T d,$$

noting that ψ affine linear with negative slope. We need to show, then, that $\varphi(t) \leq \psi(t)$ and $|\varphi'(t)| \leq -\rho\varphi'(0)$ for some $t > 0$. Now, $\varphi(0) = \psi(0)$, and $\varphi'(0) < \psi'(0)$. By Taylor's theorem, $\varphi(t) < \psi(t)$ for all $t > 0$ sufficiently small. Define

$$t^* = \min\{t > 0 \mid \varphi(t) = \psi(t)\}.$$

This exists, since $\psi(t) \to -\infty$ as $t \to \infty$, and $\varphi(t)$ is bounded from below; for small t, $\varphi(t) < \psi(t)$, so by continuity there must exist $t > 0$ for which $\varphi(t) = \psi(t)$, so t^* well-defined. Moreover, we have then that $\varphi'(t^*) \geq \psi'(t^*)$, which we can see by Taylor/graphically.

Now, we consider two cases. Suppose first that $\varphi'(t^*) < 0$. Then,

$$|\varphi'(t^*)| = -\varphi'(t^*) \leq -\psi'(t^*) = -\sigma \nabla f(x)^T d.$$

We know $\sigma < \rho$, so we're done, so this is further upper bounded by $-\rho \nabla f(x)^T d = -\rho\varphi'(0)$, so we're done in this case with t^*.

Next, suppose $\varphi'(t^*) \geq 0$. t^* won't cut it in this case, but we can see that there exists $t^{**} \in (0, t^*]$, by intermediate value theorem, for which $\varphi'(t^{**}) = 0$. Since t^* the *first* time φ, ψ are equal (being the minimum) and $t^{**} \leq t^*$, it follows that we have $\varphi(t^{**}) < \psi(t^{**})$. Also trivially,

$$|\varphi'(t^{**})| = 0 \leq -\rho\varphi'(0),$$

since $\varphi'(0) < 0$, and thus t^{**} provides the appropriate t value for the claims, so we're done. ■

Remark 2.17: In particular, this immediately gives the well-definedness of Algorithm 2.10, assuming $\left\{x^k\right\} \times \left\{d^k\right\} \in \mathcal{A}_f$.

↪**Lemma 2.13**: Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^1 and bounded from below. Let $\left\{x^k\right\}, \left\{d^k\right\}$ be generated by Algorithm 2.10. Then,

$$-\sum_{j=0}^{k} \rho^j \leq \frac{\nabla f\left(x^k\right)^T d^k}{\left\|\nabla f\left(x^k\right)\right\|^2} \leq -2 + \sum_{j=0}^{k} \rho^j,$$

for all $k \in \mathbb{N}$.

Proof. We induct on k. For $k = 0$, the claim reads

$$-1 \leq -1 \leq -2 + (1) = -1,$$

since $d^0 = -\nabla f\left(x^0\right)$, so it holds trivially.

Suppose the claim for some fixed $k \in \mathbb{N}$. We have

$$\rho \nabla f\left(x^k\right)^T d^k \leq \nabla f\left(x^{k+1}\right)^T d^k \leq -\rho \nabla f\left(x^k\right)^T d^k$$

by (S2), which implies by a little algebraic manipulation

$$-1 + \rho \frac{\nabla f\left(x^k\right)^T d^k}{\left\|\nabla f\left(x^k\right)\right\|^2} \leq -1 + \frac{\nabla f\left(x^{k+1}\right)^T d^k}{\left\|\nabla f\left(x^k\right)\right\|^2} \leq -1 - \rho \frac{\nabla f\left(x^k\right)^T d^k}{\left\|\nabla f\left(x^k\right)\right\|^2}. \qquad (\star)$$

In addition, by (S3), we know

$$\begin{aligned}
\frac{\nabla f\left(x^{k+1}\right)^T d^{k+1}}{\left\|\nabla f\left(x^{k+1}\right)\right\|^2} &= \frac{\nabla f\left(x^{k+1}\right)^T\left(-\nabla f\left(x^{k+1}\right) + \beta_k d^k\right)}{\left\|\nabla f\left(x^{k+1}\right)\right\|^2} \\
&= -\frac{\nabla f\left(x^{k+1}\right)^T \nabla f\left(x^{k+1}\right)}{\left\|\nabla f\left(x^{k+1}\right)\right\|^2} + \beta_k \frac{\nabla f\left(x^{k+1}\right)^T d^k}{\left\|\nabla f\left(x^{k+1}\right)\right\|^2} \\
&= -1 + \frac{\nabla f\left(x^{k+1}\right)^T d^k}{\left\|\nabla f\left(x^k\right)\right\|^2},
\end{aligned}$$

thus

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

thus

$$\begin{aligned}
-\sum_{j=0}^{k+1} \rho^j &= -1 - \sum_{j=1}^{k+1} \rho^j \\
&= -1 + \rho\left(-\sum_{j=0}^{k} \rho^j\right) \\
\text{(induction hypothesis)} \qquad &\leq -1 + \rho \frac{\nabla f\left(x^k\right)^T d^k}{\left\|\nabla f\left(x^k\right)\right\|^2} \\
(\star) \qquad &\leq -1 + \frac{\nabla f\left(x^{k+1}\right)^T d^k}{\left\|\nabla f\left(x^k\right)\right\|^2} \qquad (\dagger) \\
(\star) \qquad &\leq -1 - \rho \frac{\nabla f\left(x^k\right)^T d^k}{\left\|\nabla f\left(x^k\right)\right\|^2} \\
\text{(induction hypothesis)} \qquad &\leq -1 + \rho \sum_{j=0}^{k} \rho^j = -2 + \sum_{j=0}^{k+1} \rho^j.
\end{aligned}$$

But by $(\ast\,\ast)$, the line $(\dagger) = \frac{\nabla f\left(x^{k+1}\right)^T d^{k+1}}{\left\|\nabla f\left(x^{k+1}\right)\right\|^2}$, so we've shown the claim. ∎

↪**Theorem 2.18**: Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^1 and bounded from below, and let $\left\{x^k\right\}, \left\{d^k\right\}$ be generated by Algorithm 2.10. Then,

1. Algorithm 2.10 is well-defined,
2. $\nabla f\left(x^k\right)^T d^k < 0$ for all $k \in \mathbb{N}$ (it is a descent method).

Proof. By the previous lemma,

$$\frac{\nabla f\left(x^k\right)^T d^k}{\left\|\nabla f\left(x^k\right)\right\|^2} \leq -2 + \sum_{j=0}^{k} \rho^j \leq -2 + \sum_{j=0}^{\infty} \rho^j = -2 + \frac{1}{1-\rho} = \frac{2\rho - 1}{1-\rho} < 0,$$

since $\rho < \frac{1}{2}$. Multiplying both sides by $\left\|\nabla f\left(x^k\right)\right\|^2$ gives 2. Combining 2. with the previous previous lemma and the accompanying remarks, 1. follows. ■

↪**Theorem 2.19** (Al-Baali): Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^1 and bounded from below, such that f is Lipschitz on $\mathrm{lev}_{f(x_0)} f$, and let $\left\{x^k\right\}, \left\{d^k\right\}$ be generated by Algorithm 2.10. Then,

$$\lim_{k\to\infty} \left\|\nabla f\left(x^k\right)\right\| = 0.$$

§II.5 Least-Squares Problems

Supposing we wish to find the root of a function $F : \mathbb{R}^n \to \mathbb{R}^m$, we know that when $m = n$, then Newton's method is applicable. More generally, though, for $m \neq n$, such methods are not available. However, we can approach this by equivalently considering the optimization problem

$$\min_x \frac{1}{2} \|F(x)\|^2.$$

Such a problem, i.e. "minimizing the square of the norm", will be considered here. Naturally, since this is now a real-valued objective function, we could just apply Newton's method to it, but we'll do things a little more interesting.

II.5.1 Linear Least-Squares

Suppose $F(x) = Ax - b$ an affine linear function for $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$; the least-squares problem just becomes

$$\min_x \frac{1}{2}\|Ax - b\|^2. \qquad (\dagger)$$

↪**Theorem 2.20**:

1. $\bar{x}$ solves $(\dagger)$ $\Leftrightarrow$ $\bar{x}$ solves $A^TAx = A^Tb$.
2. $(\dagger)$ always has a solution.
3. $(\dagger)$ has a unique solution $\Leftrightarrow \mathrm{rank}(A) = n$.

Proof.

1. With $f(x) := \frac{1}{2}\|Ax - b\|^2$ the function of interest, one readily checks $\nabla f(x) = A^TAx - A^Tb$ (by chain rule, or by expanding f as a "proper" quadratic) and $\nabla^2 f(x) = A^TA$. Thus, since $A^TA \geq 0$ always, f is convex so stationary points are equivalent to minimization points, and thus we need $\nabla f(x) = 0 \Leftrightarrow A^TAx = A^Tb$.
2. By 1., we have a solution $\Leftrightarrow A^Tb$ in the image of A^TA; but this is equal to the image of A^T, and obviously A^Tb in the image of A^T.
3. Similarly to the previous, we will have a unique solution to $A^TAx = A^Tb$ iff A^TA has full rank $\Leftrightarrow A$ has full rank.

■

II.5.2 Gauss-Newton for Nonlinear Least-Squares

Suppose $F \in C^1$. Inspired by Newton's method, we will, instead of linearizing $f(x) := \frac{1}{2}\|F(x)\|^2$, we will linearize $F(x)$; plugging this linearization back into the norm squared, we expect a quadratic function. Indeed, suppose we have an iterate $x^k \in \mathbb{R}^n$; then, the linearization of F about x^k is given by

$$F_k(x) = F\left(x^k\right) + F'\left(x^k\right)\left(x - x^k\right).$$

Then,

$$q(x) := \frac{1}{2}\|F_k(x)\|^2 = \cdots = \frac{1}{2}x^T\left(F'\left(x^k\right)^T F'\left(x^k\right)\right)x + \left[F'\left(x^k\right)^T F\left(x^k\right) - F'\left(x^k\right)^T F'\left(x^k\right)x^k\right]^T x + \text{const},$$

where const independent of x. Assume $F'\left(x^k\right)$ of full rank n. Then, $F'\left(x^k\right)^T F'\left(x^k\right) > 0$, and so by the previous section,

$$\begin{aligned} x^+ \in \operatorname{argmin}(q) &\Leftrightarrow \nabla q(x^+) = 0 \\ &\Leftrightarrow F'\left(x^k\right)^T F'\left(x^k\right)x^+ + F'\left(x^k\right)^T F\left(x^k\right) - F'\left(x^k\right)^T F'\left(x^k\right)x^k = 0 \\ &\Leftrightarrow x^+ = x^k \underbrace{- \left(F'\left(x^k\right)^T F'\left(x^k\right)\right)^{-1} F'\left(x^k\right)^T F\left(x^k\right)}_{:=d^k}. \end{aligned}$$

Thus, this inspires the Gauss-Newton Method; supposing we can find d as a solution to the *Gauss-Newton Equations* (GNE),

$$F'\left(x^k\right)^T F\left(x^k\right)d = -F'\left(x^k\right)^T F\left(x^k\right),$$

then we update $x^{k+1} = x^k + d^k$. In particular, with this choice,

$$\nabla f(x)^T d^k = -\left(\underbrace{F'\left(x^k\right)^T F(x)^k}_{=u}\right)^T \underbrace{\left(F'\left(x^k\right)^T F'\left(x^k\right)\right)^{-1}}_{\geq 0}\left(\underbrace{F'\left(x^k\right)^T F\left(x^k\right)}_{=u}\right) < 0,$$

i.e., if $\nabla f\left(x^k\right) \neq 0$ and $F'\left(x^k\right)$ of rank n, then d^k a descent direction.

The Newton's Equation for the same function f would read

$$\left(F'\left(x^k\right)^T F'\left(x^k\right) + S\left(x^k\right)\right)d = -F'\left(x^k\right)^T F\left(x^k\right),$$

where

$$S\left(x^k\right) = \sum_{i=1}^{m} F_i\left(x^k\right)\nabla^2 F_i\left(x^k\right);$$

if S were zero, then this the same as the GNE (though of course, this will not hold in general).

§III Constrained Optimization

§III.1 Optimality Conditions for Constrained Problems

Consider

$$\min f(x) \text{ s.t.} \begin{matrix} g_i(x) \leq 0 \forall i = 1, \dots, m, \\ h_j(x) = 0 \forall j = 1, \dots, p \end{matrix},$$

where we will assume $f, g_i, h_j : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable. We call such a problem a *nonlinear program*. We put as before the *feasible set*

$$X := \left\{x \mid g_i(x) \le 0 \forall_{i=1}^m, h_j(x) = 0 \forall_{j=1}^p\right\}.$$

We'll also define the index sets

$$I := \{1, ..., m\}, \qquad J := \{1, ..., p\},$$

and the *active indices* for a point $\overline{x}$ by

$$I(\overline{x}) := \{i \in I \mid g_i(\overline{x}) = 0\} \subset I.$$

III.1.1 First-Order Optimality Conditions

Consider the slightly more abstract problem

$$\min_x f(x) \text{ s.t. } x \in S \qquad (\dagger),$$

with $f : \mathbb{R}^n \to \mathbb{R}$ in C^1 and $S \subset \mathbb{R}^n$ closed and nonempty.

↪**Definition 3.1** (Cones): A nonempty set $K \subset \mathbb{R}^n$ is said to be a *cone* if

$$\lambda v \in K \quad \forall v \in K, \lambda \ge 0,$$

i.e. K is closed under positive scalings of vectors in K.

Remark 3.1: We can in fact replace $\mathbb{R}^n$ with any real vector space V, for a cone living in V.

We have that

- any vector space;
- the nonnegative reals;
- $\Lambda := \left\{(x, y)^T \mid x, y \in K, x^T y = 0\right\}$, where K a given cone;
- and $S_+^n := \{A \in \mathbb{R}^{n \times n} \mid A \ge 0\}$ (embedded in an appropriate space of matrices)

are all cones, for instance.

↪**Definition 3.2**: Let $S \subset \mathbb{R}^n, \overline{x} \in S$. Then

$$T_s(\overline{x}) := \left\{d \in R^n \mid \exists\left\{x^k \in S\right\} \to \overline{x}, \{t_k\} \downarrow 0 \text{ s.t. } \frac{x^k - \overline{x}}{t_k} \to d\right\}$$

is called the *tangent cone of S at* $\overline{x}$.

↪**Proposition 3.1**: Let $S \subset \mathbb{R}^n, x \in S$. Then $T_S(x)$ is a closed cone.

↪**Theorem 3.1** (Basic First-Order Optimality Conditions): Let $\overline{x}$ be a local minimizer of $(\dagger)$. Then,

1. $\nabla f(\overline{x})^T d \ge 0$ for all $d \in T_S(\overline{x})$;
2. if S is convex, then $\nabla f(\overline{x})^T (x - \overline{x}) \ge 0$ for all $x \in S$.

PROOF.

1. Let $d \in T_S(\bar{x})$. By definition, there exists $\left\{x^k\right\} \subset S$ and $\{t_k\} \downarrow 0$ for which $\frac{x^k - \bar{x}}{t_k} \to d$. As x^k feasible and $\bar{x}$ a local minimizer of f over S,

$$f\left(x^k\right) - f(\bar{x}) \geq 0$$

for all k sufficiently large. By the mean value theorem, there is for each k sufficiently large a θ_k on the line between x^k and $\bar{x}$ for which

$$f\left(x^k\right) - f(\bar{x}) = \nabla f(\theta_k)^T \left(x^k - \bar{x}\right),$$

so

$$0 \leq \frac{f\left(x^k\right) - f(\bar{x})}{t_k} = \frac{\nabla f(\theta_k)^T \left(x^k - \bar{x}\right)}{t_k} \underset{k}{\to} \nabla f(\bar{x})^T d.$$

2. Suppose not. Then, there exists a $\hat{x} \in S$ such that $\nabla f(\bar{x})^T(\hat{x} - \bar{x}) < 0$. By convexity, $\bar{x} + \lambda(\hat{x} - \bar{x}) \in S$ for $\lambda \in (0, 1)$. By mean value theorem, for every such λ there exists a θ_λ on the line between $\bar{x} + \lambda(\hat{x} - \bar{x})$ and $\bar{x}$ for which

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

By supposition, for λ sufficiently close to 0, the right-hand side will remain negative (since $\nabla f(\theta_\lambda) \underset{\lambda \to 0}{\to} \nabla f(\bar{x})$), so for sufficiently small λ,

$$f(\bar{x} + \lambda(\hat{x} - \bar{x})) < f(\bar{x}),$$

and since $\bar{x} + \lambda(\hat{x} - \bar{x})$ remains feasible for all λ by covexity, this contradicts minimality.

■

Remark 3.2: Computationally, this isn't very helpful - in practice, i.e. in trying to compute local minimizers, we'd need to compute $\nabla f(\bar{x})^T d$ for every d in the tangent cone of a given S at a given point $\bar{x}$. In general, we don't know what this set looks like, and even if we did, this isn't a feasible condition to check for every such point, since it isn't easy to interpret computationally.

> You can never tell the computer what the fucking set looks like
>
> — Tim H

III.1.2 Farkas' Lemma

↪**Definition 3.3** (Projection): Let $S \subset \mathbb{R}^n$ be nonempty, $x \in \mathbb{R}^n$. The *projection* of x onto S is given by

$$P_S(x) := \operatorname{argmin}_{y \in S} \frac{1}{2}\|x - y\|^2.$$

Remark 3.3: This is, in general, a set-valued function; it could even be empty (for instance, if $S = [0, 1)$ and $x = 2$.)

↪**Proposition 3.2**: Let $x \in \mathbb{R}^n, S \subset \mathbb{R}^n$ nonempty, closed and convex. Then,

1. $P_S(x)$ has exactly one element, so P_S can be viewed $\mathbb{R}^n \to S$;
2. $P_S(x) = x \Leftrightarrow x \in S$;
3. (The Projection Theorem) $(P_S(x) - x)^T(y - P_S(x)) \geq 0$ for all $y \in S$.

Proof.

1. This follows from the fact that $S \ni y \mapsto \|x - y\|_2^2$ a strongly convex function.
2. Clear.
3. Take $f(y) = \frac{1}{2}\|x - y\|^2$ in the last theorem.

■