

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.1.2 Differentiable Convex Functions	7
II.1.3 Matrix Norms	9
II.1.4 Descent Methods	11
II.1.4.1 A General Line-Search Method	11
II.1.4.1.1 Global Convergence of Algorithm 2.1	12
II.1.4.2 The Gradient Method	13
II.1.4.3 Newton-Type Methods	15
II.1.4.3.1 Convergence Rates and Landau Notation	15
II.1.4.3.2 Newton's Method for Nonlinear Equations	16
II.1.4.3.3 Newton's Method for Optimization Problem	18
II.1.4.4 Quasi-Newton Methods	21
II.1.4.4.1 Direct Methods	21
II.1.4.4.2 Inexact Methods	26
II.1.4.5 Conjugate Gradient Methods for Nonlinear Optimization	27
II.1.4.5.1 Prelude: Linear Systems	27
II.1.4.6 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
III.1.3 Karush-Kuhn-Tucker Conditions	38
III.1.4 Constraint Qualifications	39
III.1.5 Affine constraints	42
III.1.6 Convex Problems	42
III.2 Lagrangian Duality	43
III.2.1 The Dual Problem	43
III.2.2 Weak and Strong Duality	44
III.2.3 The Saddle Point Theorem	45
III.3 Penalty Methods	46
III.3.1 The Quadratic Penalty Function	46

§I PRELIMINARIES

§I.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

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

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 (off 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 (off 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. ■

§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, $f : C \rightarrow \mathbb{R}$ 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 .

§II UNCONSTRAINED OPTIMIZATION

§II.1 Theoretical Foundations

We focus on the problem

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

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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, $\frac{t^2}{d^T} d^T \nabla^2 f(\eta_t) d < 0$ 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 .

II.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

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

㊂ **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).

§II.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

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

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 $\Omega \Leftrightarrow \nabla^2 f > 0$;
3. 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 $f(x) := \frac{1}{2}x^T Ax + b^T x$. 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

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

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 $\lambda \rightarrow 1^-$, 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 (\star) and

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

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.

III.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 $A = (a_{ij})_{i=1,\dots,m, j=1,\dots,n} \in \mathbb{R}^{m \times n}$, 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

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

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

§II.4 Descent Methods

II.4.1 A General Line-Search Method

We deal with the unconstrained problem

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

→ **Definition 2.2** (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 2.5:** 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 2.6:** 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 2.2: 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 2.3** (Step-size rule): Let $f \in C^1(\mathbb{R}^n)$ and

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

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

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

→ **Theorem 2.8:** Let $f \in C^1(\mathbb{R}^n)$. Let $\{x^k\}, \{d^k\}, \{t_k\}$ be generated by Algorithm 2.1. 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 . ■

II.4.2 The Gradient Method

We specialize Algorithm 2.1 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

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

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.

④ **Example 2.3** (Illustration of Armijo Rule): For $(x, d) \in \mathcal{A}_f$ and f smooth on \mathbb{R}^n , defined $\phi : \mathbb{R} \rightarrow \mathbb{R}$, $\phi(t) := f(x + td)$. The map $t \mapsto \sigma\varphi'(0)t + \varphi(0) = \sigma t \nabla f(x)^T d + \varphi(0)$

→ **Proposition 2.6:** Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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

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

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 $(x, d) \in \mathcal{A}_f$, 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 $\ \nabla f(x^k)\ \leq \varepsilon$, 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 2.4:** 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 2.9:** Let $f \in C^1(\mathbb{R}^n)$. Then every cluster point of a sequence $\{x^k\}$ generated by Algorithm 2.2 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 2.3: 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.

II.4.3 Newton-Type Methods

II.4.3.1 Convergence Rates and Landau Notation

→ **Definition 2.5:** 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 2.4: 3. \Rightarrow 2. \Rightarrow 1.

Remark 2.5: 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 2.6** (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. $a_k = O(b_k) \Leftrightarrow \exists C > 0 : a_k \leq C b_k$ for all k (sufficiently large).

Remark 2.6: 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 $\Leftrightarrow \|x^{k+1} - \bar{x}\| = O(\|x^k - \bar{x}\|^2)$.

II.4.3.2 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 2.5:** 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 2.7: 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 2.6:** Let $F : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable and $x^k \rightarrow \bar{x}$, then:

1. $\|F(x^k) - F(\bar{x}) - F'(x^k)(x^k - \bar{x})\| = o(\|x^k - \bar{x}\|);$
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

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

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,

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

which implies the result.

■

→ **Theorem 2.10:** 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 2.3 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

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

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.,

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

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..

■

II.4.3.3 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 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:

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

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 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. $(x^k, d^k) \in \mathcal{A}_f$ for every k .

→ **Theorem 2.11** (Global convergence of Algorithm 2.5): Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be twice continuously differentiable. Then every cluster point of $\{x^k\}$ 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 $\{x^k\}$ be generated by Algorithm 2.5, 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 2.12** (Fast local convergence of Algorithm 2.5): Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be twice continuously differentiable, $\{x^k\}$ generated by Algorithm 2.5. 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.

II.4.4 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.

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)

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

3. H_{k+1} can be achieved from H_k “efficiently”

4. The result method has strong local convergence properties

Remark 2.10: 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

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

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

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 $k := 0$.

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

S2. Determine d^k as a solution to the QNE,

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

S3. Determine $t_k > 0$ such that

$$f(x^k + t_k d^k) \leq f(x^k) + \sigma t_k \nabla f(x^k)^T d^k,$$

(this is just the Armijo condition), AND

$$\nabla f(x^k + t_k d^k)^T d^k \geq \rho \nabla f(x^k)^T d^k,$$

call the *Wolfe-Powell rule*.

S4. Set

$$\begin{aligned} x^{k+1} &:= x^k + t_k d^k, \\ s^k &:= x^{k+1} - x^k, \\ y^k &:= \nabla f(x^{k+1}) - \nabla f(x^k), \\ H_{k+1} &:= H_{k+1}^{\text{BFGS}}. \end{aligned}$$

S5. Increment k and go to S1.

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

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

↪**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 $f(x^0)$ 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 2.6.

↪**Lemma 2.9:** 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

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

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 2.10:** If in the k th iteration of Algorithm 2.6 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

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

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

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

↪ **Theorem 2.14:** Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable, and $\{x^k\}, \{d^k\}, \{t_k\}$ 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 $\{x^k\}$ is a stationary point of f .

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

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

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 $\mathcal{A}_f \cap \mathcal{L} \times \mathbb{R}^n$, 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 $\text{lev}_{f(x^0)} f$ 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.

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 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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 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 $\eta_k = O(\|\nabla f(x^k)\|)$, 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 $x^0 \in \mathbb{R}^n, \varepsilon \geq 0, \rho > 0, p > 2, \beta \in (0, 1), \sigma \in (0, \frac{1}{2})$ and set $k := 0$.

S1. If $\|\nabla f(x^k)\| \leq \varepsilon$ STOP.

S2. Choose $\eta_k \geq 0$ and determine d^k by

$$\|\nabla^2 f(x^k) d + \nabla f(x^k)\| \leq \eta_k \|\nabla f(x^k)\|.$$

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

S3. Determine $t_k > 0$ by Armijo, $t_k := \max_{\{\ell \in \mathbb{N}_0\}} \left\{ \beta^\ell \mid f(x^k + \beta^\ell d^k) \leq f(x^k) + \beta^\ell \sigma \nabla f(x^k)^T d^k \right\}$.

S4. Set $x^{k+1} = x^k + t_k d^k$, increment k and go to S1.

→ **Theorem 2.16:** Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be C^2 and let $\{x^k\}$ be generated by Algorithm 2.8 with $\eta_k \downarrow 0$.

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.

II.4.5 Conjugate Gradient Methods for Nonlinear Optimization

II.4.5.1 Prelude: Linear Systems

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

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

→ **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, \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 $t_k := \operatorname{argmin}_{t>0} f(x^k + t_k d^k)$. 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

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

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

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 $i, j = 0, \dots, k$ 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 $j = 0, \dots, k$,

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

Thus, g^n is orthogonal to the n linearly independent vectors $\{d^0, \dots, d^{n-1}\}$, 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 $k = 0, \dots, n-1$. 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, \dots, 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 $\{d^0, \dots, d^l\}$, and $\{\beta_{il}\}$ are scalars to be determined. The condition $(d^{l+1})^T Ad^j = 0$ readily implies that

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

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

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

and so with $t_j > 0$,

$$(g^{l+1})^T Ad^j = \frac{1}{t_j} (g^{l+1})^T (g^{j+1} - g^j),$$

and thus

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

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

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

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 $\ g^k\ \leq \varepsilon$, STOP.
S2. Put
$t_k := \frac{\ g^k\ ^2}{(d^k)^T Ad^k}.$
S3. Set
$x^{k+1} = x^k + t_k d^k$
$g^{k+1} = g^k + t_k Ad^k$
$\beta_k = \frac{\ g^{k+1}\ ^2}{\ g^k\ ^2}$
$d^{k+1} = -g^{k+1} + \beta_k d^k.$
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 $f(x) := \frac{1}{2}x^T Ax - b^T x$. Then, Algorithm 2.9 will produce the global minimum \bar{x} of f after at most n iterations. If $m \in \{0, \dots, n\}$ is the smallest number such that $x^m = \bar{x}$, then the following hold as well:

$$(d^k)^T Ad^j = 0, (g^k)^T g^j = 0, (g^k)^T d^j = 0, \quad (k = 1, \dots, m, j = 0, \dots, k-1),$$

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

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 issue 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 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ reads

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

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

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

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

PROOF. Define $\varphi, \psi : \mathbb{R} \rightarrow \mathbb{R}$ by

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

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 $\psi(t) \rightarrow -\infty$ as $t \rightarrow \infty$, 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 $\{x^k\} \times \{d^k\} \in \mathcal{A}_f$.

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

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

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

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

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

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

$$-1 + \rho \frac{\nabla f(x^k)^T d^k}{\|\nabla f(x^k)\|^2} \leq -1 + \frac{\nabla f(x^{k+1})^T d^k}{\|\nabla f(x^k)\|^2} \leq -1 - \rho \frac{\nabla f(x^k)^T d^k}{\|\nabla f(x^k)\|^2}. \quad (*)$$

In addition, by (S3), we know

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

thus

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

thus

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

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

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

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

PROOF. By the previous lemma,

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

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

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

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

§II.5 Least-Squares Problems

Supposing we wish to find the root of a function $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$, 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

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

→ **Theorem 2.20:**

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

PROOF.

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

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

Then,

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

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

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

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

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

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

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

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

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

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

where

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

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

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

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

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

We'll also define the index sets

$$I := \{1, \dots, m\}, \quad J := \{1, \dots, p\},$$

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

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

III.1.1 First-Order Optimality Conditions

Consider the slightly more abstract problem

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

with $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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

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

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

are all cones, for instance.

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

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

is called the *tangent cone of S at \bar{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 \bar{x} be a local minimizer of (\dagger) .

Then,

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

PROOF.

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

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

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

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

so

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

- 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 $\nabla f(\theta_\lambda) \xrightarrow{\lambda \rightarrow 0} \nabla f(\bar{x})$), 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 convexity, 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 $\mathbb{R}^n \rightarrow S$;
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.

■

→ **Lemma 3.1:** Let $B \in \mathbb{R}^{\ell \times n}$. Then, $\{Bx \mid x \in \mathbb{R}_+^n\}$ is a nonempty, closed, convex cone.

PROOF. Convexity and cone properties are clear. Closed? ■

→ **Theorem 3.2 (Farkas' Lemma):** Let $B \in \mathbb{R}^{\ell \times n}$, $h \in \mathbb{R}^n$. Then, the following are equivalent:

1. The system

$$B^T x = h, x \geq 0$$

has a solution.

2. $h^T d \geq 0$ for all d such that $Bd \geq 0$.

Remark 3.4: $x \geq 0$ should be understood component-wise, i.e. each component of x is nonnegative.

PROOF. (1. \Rightarrow 2.) Let $x \geq 0$ such that $B^T x = h$. Then, if d such that $Bd \geq 0$,

$$h^T d = (B^T x)^T d = x^T Bd \geq 0.$$

(2. \Rightarrow 1.) Suppose 1. doesn't hold, i.e.

$$h \notin K = \{B^T x \mid x \geq 0\},$$

where K a closed, convex cone as the previous lemma. Set $\bar{s} = P_K(h) \in K$, which is well-defined since K closed and convex. Set $\bar{d} = \bar{s} - h \neq 0$. Thus, by the rojection theorem,

$$\bar{d}^T(s - \bar{s}) \geq 0$$

for all $s \in K$.

By taking $s = 2\bar{s} \in K$, we see then that $\bar{d}^T \bar{s} \geq 0$. Also, taking $\bar{s} = 0$, this implies $-\bar{d}^T \bar{s} \geq 0$, by which we must have $\bar{d}^T \bar{s} = 0$ and thus $\bar{d}^T s \geq 0$ for all $s \in K$. By definition of K , $(B\bar{d})^T x = \bar{d}^T B^T x \geq 0$ for all $x \geq 0$. This implies $B\bar{d} \geq 0$, by taking x to be each standard unit vector e_i .

OTOH,

$$h^T \bar{d} = (\bar{s} - \bar{d})^T \bar{d} = \underbrace{\bar{s}^T \bar{d}}_{=0} - \|\bar{d}\|^2 < 0,$$

since $\bar{d} \neq 0$. This contradicts 2. ■

III.1.3 Karush-Kuhn-Tucker Conditions

→ **Definition 3.4** (KKT Conditions): Consider the standard nonlinear program

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

1. The function $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$ defined by

$$\begin{aligned} L(x, \lambda, \mu) &= f(x) + \sum_{i=1}^m \lambda_i g_i(x) + \sum_{j=1}^p \mu_j h_j(x) \\ &= f(x) + \lambda^T g(x) + \mu^T h(x), \end{aligned}$$

where $\lambda := (\lambda_1, \dots, \lambda_m)$, $g = (g_1, \dots, g_m)$, $\mu = (\mu_1, \dots, \mu_p)$, $h = (h_1, \dots, h_p)$, is called the *Lagrangian* of the problem (*).

2. The set of conditions

$$\begin{aligned} \nabla L_x(x, \lambda, \mu) &= 0, \\ h(x) &= 0, \\ \lambda &\geq 0, g(x) \leq 0, \lambda^T g(x) &= 0 \end{aligned}$$

are called the *KKT Condition* of (*).

3. A triple $(\bar{x}, \bar{\lambda}, \bar{\mu})$ that satisfies the KKT conditions is called a *KKT point* of (*).
4. Given \bar{x} feasible for (*), define

$$M(\bar{x}) = \{(\lambda, \mu) \mid (\bar{x}, \lambda, \mu) \text{ is a KKT point of } (*)\}.$$

→ **Definition 3.5** (Linearized Cone): Let X be the feasible set of (*) and $\bar{x} \in X$. The *linearized cone* (of X) at \bar{x} is given by the set

$$\mathcal{L}_X(\bar{x}) := \left\{ d \mid \begin{array}{l} \nabla g_i(\bar{x})^T d \leq 0 \forall i \in I(\bar{x}) \\ \nabla h_j(\bar{x})^T d = 0 \forall j \in J \end{array} \right\}.$$

→ **Definition 3.6** (Abadie Constraint Qualification): Let $\bar{x} \in X$. We say that the *Abadie constraint qualification (ACQ)* holds at \bar{x} if $T_X(\bar{x}) = \mathcal{L}_X(\bar{x})$.

Remark 3.5: We may represent the constraints that lead to X in different ways. These different representations may lead to different linearized cones $\mathcal{L}_X(\bar{x})$, but will NOT change $T_X(\bar{x})$. So, the ACQ may hold/not hold depending on how we represent X for a fixed problem.

→ **Theorem 3.3** (KKT Conditions Under ACQ): Let \bar{x} be a local minimizer of (\star) such that ACQ holds at \bar{x} . Then, there exist $(\bar{\lambda}, \bar{\mu}) \in \mathbb{R}^m \times \mathbb{R}^p$ such that $(\bar{x}, \bar{\lambda}, \bar{\mu})$ a KKT point.

PROOF. \bar{x} a local minimizer implies by the basic first-order optimality conditions for (\star) that $\nabla f(\bar{x})^T d \geq 0$ for all $d \in T_X(\bar{x})$. Set

$$B = \begin{pmatrix} -\nabla g_i(x)^T & (i \in I(\bar{x})) \\ -\nabla h_j(x)^T & (j \in J) \\ \nabla h_j(x)^T & (j \in J) \end{pmatrix} \in \mathbb{R}^{(|I(\bar{x})|+2p) \times n}.$$

Note that $d \in \mathcal{L}_X(\bar{x})$ iff $Bd \geq 0$. By the ACQ, $\nabla f(\bar{x})^T d \geq 0$ for all $d \in \mathcal{L}_X(\bar{x})$, hence $\nabla f(\bar{x})^T d \geq 0$ for all d such that $Bd \geq 0$. By Farkas' Lemma (taking B as defined, $h = \nabla f(\bar{x})$), there exists a $y = (y^1, y^2, y^3) \in \mathbb{R}^{|I(\bar{x})|} \times \mathbb{R}^p \times \mathbb{R}^p$ such that $B^T y = \nabla f(\bar{x})$ and $y \geq 0$. Define

$$\bar{\lambda} := \begin{cases} y_i^1 & i \in I(\bar{x}), \\ 0 & \text{else} \end{cases}, \quad \bar{\mu} := y^2 - y^3.$$

Then, $(\bar{x}, \bar{\lambda}, \bar{\mu})$ is a KKT point. ■

⊕ **Example 3.1:** Consider

$$\min x_1^2 + x_2^2 \quad \text{s.t. } x_1, x_2 \geq 0, x_1 x_2 = 0,$$

with $X = \{x \in \mathbb{R}^2 \mid x_1, x_2 \geq 0, x_1 x_2 = 0\}$. Let $\bar{x} = (0, 0)^T \in X$. We find that

$$T_X(\bar{x}) = X, \quad \mathcal{L}_X(\bar{x}) = \mathbb{R}_+^2.$$

So, ACQ does not hold. However, with $\bar{\lambda} = 0$ and $\bar{\mu} = 1$, we find $\nabla f(\bar{x}) + \bar{\lambda}_1 \nabla g_1(\bar{x}) + \bar{\lambda}_2 \nabla g_2(\bar{x}) + \bar{\mu} \nabla h(\bar{x}) = 0$, and we find $(\bar{x}, \bar{\lambda}, \bar{\mu})$ is a KKT point.

III.1.4 Constraint Qualifications

→ **Definition 3.7:** Let \bar{x} be a feasible point of the generic nonlinear program. We say that:

1. the *linear independence constraint qualification (LICQ)* holds at \bar{x} if the vectors

$$\nabla g_i(\bar{x}), i \in I(\bar{x}), \nabla h_j(\bar{x}), j \in J,$$

are linearly independent;

2. the *Mangasarian-Fromovitz CQ (MFCQ)* holds at \bar{x} if the gradients

$$\nabla h_j(\bar{x}), j \in J$$

are linearly independent, and there exists a $d \in \mathbb{R}^n$ such that

$$\nabla g_i(\bar{x})^T d < 0, i \in I(\bar{x}),$$

$$\nabla h_j(\bar{x})^T d = 0, j \in J.$$

→ **Proposition 3.3** ($\text{LICQ} \Rightarrow \text{MFCQ}$): Let \bar{x} be feasible and such that LICQ holds at \bar{x} . Then MFCQ holds at \bar{x} .

PROOF. Left as an exercise. ■

↪ **Theorem 3.4** (Implicit Function Theorem): Let $U \subset \mathbb{R}^p, V \subset \mathbb{R}^m$ open and $F : U \times V \rightarrow \mathbb{R}^b$ continuous differentiable such that for $(\bar{x}, \bar{y}) \in U \times V, F(\bar{y}, \bar{z}) = 0$ and $D_y F(\bar{y}, \bar{z}) \in \mathbb{R}^{p \times b}$ invertible. Then, there exist neighborhoods $\hat{U} \subset U$ and $\hat{V} \subset V$ and $g : \hat{V} \rightarrow \hat{U}$ continuously differentiable such that

$$F(g(z), z) = 0, \forall z \in \hat{V},$$

and

$$Dg(z) = -[D_y F(g(z), z)]^{-1} D_z F(g(z), z),$$

and for all $(y, z) \in \hat{U} \times \hat{V}$ such that $F(y, z) = 0$, then $y = g(z)$.

↪ **Lemma 3.2:** Let \bar{x} be feasible such that $\nabla h_j(\bar{x}), j \in J$ are linearly independent, and there exists d such that $\nabla g_i(\bar{x})^T d < 0$ for $i \in I(\bar{x})$ and $\nabla h_j(\bar{x})^T d = 0$ for $j \in J$. Then, there exists an $\varepsilon > 0$ and a C^1 -curve $x : (-\varepsilon, \varepsilon) \rightarrow \mathbb{R}^n$ such that $x(t)$ feasible for all $t \in [0, \varepsilon]$, $x(0) = \bar{x}$, and $x'(0) = d$.

PROOF. Define $H : \mathbb{R}^p \times \mathbb{R} \rightarrow \mathbb{R}^p$ by

$$H_j(y, t) := h_j(\bar{x} + td + h'(\bar{x})^T y), \quad j = 1, \dots, p.$$

The equation $H(y, t) = 0$ has as solution $(\bar{y}, \bar{t}) = (0, 0)$, as $h_j(\bar{x}) = 0$ by feasibility. Moreover,

$$D_y H(\bar{y}, \bar{t}) = h'(\bar{x})h'(\bar{x})^T \in \mathbb{R}^{p \times p},$$

with $h'(\bar{x})^T$ having linearly independent columns, hence being invertible, so all of $D_y H(\bar{y}, \bar{t})$ is invertible. By the implicit function theorem, there exists an $\varepsilon > 0$ and a C^1 -curve $y : (-\varepsilon, \varepsilon) \rightarrow \mathbb{R}^p$ for which $y(0) = 0, H(y(t), t) = 0$ for all $t \in (-\varepsilon, \varepsilon)$, and $y'(t) = -[D_y H(y(t), t)]^{-1} D_t H(y(t), t)$ for all $t \in (-\varepsilon, \varepsilon)$. In particular,

$$y'(0) = -[D_y H(0, 0)]^{-1} h'(\bar{x})d = 0,$$

since $h'(\bar{x}) = [\nabla h_j(\bar{x})^T] = 0$ by assumption. Set now $x(t) = \bar{x} + td + h'(\bar{x})^T y(t)$. Making ε smaller if necessary, we see that $x : (-\varepsilon, \varepsilon) \rightarrow \mathbb{R}^n$ has all the desired properties:

- $x \in C^1$ since $y \in C^1$
- $x(0) = \bar{x}$
- $x'(0) = d$
- $x(t)$ is feasible: by construction, for $t \in [0, \varepsilon], h_j(x(t)) = H_j(y(t), t) = 0$; for $i \in I \setminus I(\bar{x})$, then $g_i(\bar{x}) = g_i(x(0)) < 0$, then by continuity $g_i(x(t))$ will remain negative for all sufficiently small t ; for $i \in I(\bar{x}), g_i(\bar{x}) = g_i(x(0)) = 0$, and $\frac{d}{dt} g_i(x(0)) = \nabla g_i(\bar{x})^T d < 0$, which implies $g_i(x(t)) < 0$ for all t sufficiently small.

→ **Proposition 3.4** (MFCQ \Rightarrow ACQ): Let \bar{x} be a feasible point such that MFCQ holds at \bar{x} ; then ACQ holds.

PROOF. We only need to show $\mathcal{L}_X(\bar{x}) \subset T_X(\bar{x})$. Let $d \in \mathcal{L}_X(\bar{x})$. By MFCQ there exists \hat{d} such that $\nabla g_i(\bar{x})^T \hat{d} < 0$ for $i \in I(\bar{x})$ and $\nabla h_j(\bar{x})^T \hat{d} = 0$ for all $j \in J$. Set $d(\delta) = d + \delta \hat{d}$ for all $\delta > 0$. Then,

$$\nabla g_i(\bar{x})^T d(\delta) < 0, \forall i \in I(\bar{x}), \quad \nabla h_j(\bar{x})^T d(\delta) = 0 \forall j \in J.$$

Applying the previous lemma to $d(\delta)$ yields a C^1 -curve $x : (-\varepsilon, \varepsilon) \rightarrow \mathbb{R}^n$ such that $x(0) = \bar{x}$, $x'(0) = d(\delta)$, and $x(t)$ feasible for all $t \in [0, \varepsilon]$. Let $t_k \downarrow 0$ and set $x^k := x(t_k)$, then x^k feasible and $x^k \rightarrow \bar{x}$. We see that

$$d(\delta) = x'(0) = \lim_{k \rightarrow \infty} \frac{x(t_k) - x(0)}{t_k} = \lim_{k \rightarrow \infty} \frac{x^k - \bar{x}}{t_k} \in T_X(\bar{x}),$$

hence for all $\delta > 0$, $d(\delta) \in T_X(\bar{x})$. Letting $\delta \rightarrow 0$, $d(\delta) \rightarrow d$ and thus $d \in T_X(\bar{x})$ as well, since $T_X(\bar{x})$ closed. ■

→ **Corollary 3.1:** Let \bar{x} be a local min of the generic nonlinear program such that MFCQ holds at \bar{x} . Then the following hold:

1. $M(\bar{x}) \neq \emptyset$;
2. $M(\bar{x})$ is bounded.

PROOF. 1. follows from the previous proposition and the related theorem for ACQ. For 2., suppose otherwise, then there exists $\{(\lambda^k, \mu^k)\} \subset M(\bar{x})$ for which $\|(\lambda^k, \mu^k)\| \rightarrow \infty$. Without loss of generality, we may assume $\frac{\lambda^k, \mu^k}{\|(\lambda^k, \mu^k)\|} \rightarrow (\tilde{\lambda}, \tilde{\mu}) \neq 0$. Since $(\bar{x}, \lambda^k, \mu^k)$ is a KKT point, we have

$$0 = \frac{\nabla f(\bar{x}) + \sum_{i \in I(\bar{x})} \lambda_i^k \nabla g_i(\bar{x}) + \sum_{j \in J} \mu_j^k \nabla h_j(\bar{x})}{\|(\lambda^k, \mu^k)\|} \rightarrow \sum_{i \in I(\bar{x})} \tilde{\lambda}_i \nabla g_i(\bar{x}) + \sum_{j \in J} \tilde{\mu}_j \nabla h_j(\bar{x}).$$

If $\tilde{\lambda} = 0$, $\tilde{\mu} \neq 0$ and thus $0 = \sum_j \tilde{\mu}_j \nabla h_j(\bar{x})$ contradicts the linear independence of $\{\nabla h_j(\bar{x})\}$ of MFCQ.

If $\tilde{\lambda} \neq 0$, then there exists $i_0 \in I(\bar{x})$ such that $\tilde{\lambda}_{i_0} > 0$. Multiplying the above expression by d from MFCQ yields

$$0 = \sum_{i \in I(\bar{x})} \underbrace{\tilde{\lambda}_i \nabla g_i(\bar{x})^T d}_{< 0} + \sum_{i \in J} \underbrace{\tilde{\mu}_j \nabla h_j(\bar{x})^T d}_{= 0} \leq \tilde{\lambda}_{i_0} \nabla g_{i_0}(\bar{x})^T d < 0,$$

a contradiction. ■

→ **Corollary 3.2:** Let \bar{x} be a local minimimum of the generic nonlinear program such that LICQ holds. Then, $M(\bar{x})$ is a singleton.

In summary, we have

$$\begin{array}{ccc}
\text{LICQ} \Rightarrow & \text{MFCQ} \Rightarrow & \text{ACQ} \\
\Downarrow & \Downarrow & \Downarrow \\
M(\bar{x}) \text{ is: } & \begin{array}{l} \text{a singleton} \\ \text{nonempty, nonempty} \\ \text{bounded} \end{array} &
\end{array}$$

III.1.5 Affine constraints

→ **Definition 3.8** (Affine CQ): We say that the *affine CQ* holds if all constraints are affine, i.e. there exists $a_i \in \mathbb{R}^n, \alpha_i \in \mathbb{R}$ ($i \in I$), $b_j \in \mathbb{R}^n, \beta_j \in \mathbb{R}$ ($j \in J$) such that

$$g_i(x) = a_i^T x - \alpha_i, \quad h_j(x) = b_j^T - \beta_j.$$

→ **Proposition 3.5:** Let the affine CQ hold. Then, ACQ holds at every feasible point.

PROOF. Let $\bar{x} \in X$. As before, we only need to show $\mathcal{L}_X(\bar{x}) \subset T_X(\bar{x})$, then apply the previous lemma. Pick $d \in \mathcal{L}_X(\bar{x})$, i.e. $a_i^T d \leq 0, b_j^T d = 0$ for $i \in I(\bar{x}), j \in J$. Let now $t_k \downarrow 0$ and $x^k := \bar{x} + t_k d$. Then, $x^k \rightarrow \bar{x}$ as $t_k \rightarrow 0$ and $\frac{x^k - \bar{x}}{t_k} \rightarrow d$. It remains to show that $x^k \in X$ for all k sufficiently large. We check the conditions for x^k to be sufficient.

If $i \notin I(\bar{x})$, $a_i^T \bar{x} < \alpha_i$, so by continuity this strict inequality remains true replacing \bar{x} by x^k for k sufficiently large.

If $i \in I(\bar{x})$, $a_i^T \bar{x} = \alpha_i$ thus

$$a_i^T x^k = a_i^T \bar{x} + t_k a_i^T d \leq \alpha_i.$$

Similar work follows for $j \in J$, i.e. $b_j^T x^k = b_j^T \bar{x} + t_k b_j^T d = \beta_j + 0$. ■

III.1.6 Convex Problems

We consider

$$\begin{array}{ll}
\min f(x) & g_i(x) \leq 0, i \in I \\
\text{s.t.} & b_j^T x = \beta_j, j \in J'
\end{array}$$

where now we assume $f, g_i : \mathbb{R}^n \rightarrow \mathbb{R}$ are C^1 and convex. The feasible set

$$X = \{x \mid g_i(x) \leq 0 (i \in I), b_j^T x = \beta_j (j \in J)\}$$

is then convex (as per a midterm question).

→ **Theorem 3.5** (KKT for Convex NLP): Let \bar{x} be feasible for the convex NLP above and consider the following statements:

1. there exists $(\bar{\lambda}, \bar{\mu}) \in M(\bar{x})$ (i.e., $(\bar{x}, \bar{\lambda}, \bar{\mu})$ is a KKT point);
2. \bar{x} is a (global) minimizer;

then, always 1. \Rightarrow 2., and hence if a constraint qualification holds at \bar{x} , then 1. and 2. are equivalent.

PROOF. Let $(\bar{x}, \bar{\lambda}, \bar{\mu})$ be a KKT point. Then for any $x \in X$,

$$\begin{aligned}
f(x) &\stackrel{\text{convexity}}{\geq} f(\bar{x}) + \nabla f(\bar{x})^T(x - \bar{x}) \\
&\stackrel{\text{KKT}}{=} f(\bar{x}) + \left(- \sum_{i \in I(\bar{x})} \bar{\lambda}_i \nabla g_i(\bar{x}) - \sum_{j \in J} \bar{\mu}_j \nabla h_j(\bar{x}) \right)^T (x - \bar{x}) \\
&= f(\bar{x}) - \underbrace{\sum_{i \in I(\bar{x})} \bar{\lambda}_i \nabla g_i(\bar{x})^T (x - \bar{x})}_{\leq g_i(x) - g_i(\bar{x}) = g_i(x)} - \underbrace{\sum_{j \in J} \bar{\mu}_j \nabla h_j(\bar{x})^T (x - \bar{x})}_{= 0, \text{ feasibility}} \\
&\geq f(\bar{x}) - \underbrace{\sum_{i \in I(\bar{x})} \bar{\lambda}_i g_i(x)}_{\leq 0} \geq f(\bar{x}),
\end{aligned}$$

which gives \bar{x} indeed a local min, so 2. holds. The reverse implication holding under a CQ is the very definition of a CQ. ■

→ **Definition 3.9** (Slater CQ): We say the *Slater CQ* (*SCQ*) holds for the convex NLP if there exists a \hat{x} such that

$$g_i(\hat{x}) < 0, (i \in I), h_j(\hat{x}) = 0 (j \in J).$$

We call such a point a *Slater point*.

→ **Proposition 3.6** (*SCQ* \Rightarrow *ACQ*): Let *SCQ* hold for the convex NLP. Then, *ACQ* holds at every feasible point.

PROOF. Let $\bar{x} \in X$ and set

$$F(\bar{x}) := \{d \mid \nabla g_i(\bar{x})^T d < 0, i \in I, b_j^T d = 0, j \in J\}.$$

Then, $F(\bar{x}) \subset T_X(\bar{x})$ (exercise). As $T_X(\bar{x})$ is closed, it follows that $\overline{F(\bar{x})} \subset T_X(\bar{x})$. We know show that $\mathcal{L}_X(\bar{x}) \subset \overline{F(\bar{x})}$, from whence we'll be done. So, let $d \in \mathcal{L}_X(\bar{x})$ and let \hat{x} be a Slater point. Put $\hat{d} = \hat{x} - \bar{x}$. Then,

$$\nabla g_i(\bar{x})^T \hat{d} \leq g_i(\hat{x}) - g_i(\bar{x}) < 0$$

for $i \in I(\bar{x})$, and moreover, $\nabla h_j(\bar{x})^T \hat{d} = 0$. Define for $\delta > 0$, $d(\delta) := d + \delta \hat{d}$. Then, $d(\delta) \in F(\bar{x})$ for all $\delta > 0$ (CHECK THIS). Let $\delta \rightarrow 0$; then $d(\delta) \rightarrow d \in \overline{F(\bar{x})}$. ■

§III.2 Lagrangian Duality

Consider as before the standard NLP

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

with $f, g_i, h_j : \mathbb{R}^n \rightarrow \mathbb{R}$ with $X = \{x \mid g(x) \leq 0, h(x) = 0\} \neq \emptyset$ (we don't make smoothness assumptions just yet). The *Lagrangian* of the problem (P) is given by

$$L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}, \quad L(x, \lambda, \mu) = f(x) + \lambda^T g(x) + \mu^T h(x).$$

III.2.1 The Dual Problem

Observe that

$$\sup_{\lambda, \mu} L(x, \lambda, \mu) = \begin{cases} f(x) & \text{if } x \in X \\ \infty & \text{else} \end{cases}$$

Then, the problem (P) is really equivalent to the problem

$$\min_{x \in \mathbb{R}^n} \sup_{\substack{\lambda \in \mathbb{R}_+^m, \\ \mu \in \mathbb{R}^p}} L(x, \lambda, \mu) = \inf_x p(x),$$

where we call $p(x)$ the *primal objective*. A question of interest is when can we switch the order of the min and the sup?

↪ **Definition 3.10** (Lagrangian Dual): The *(Lagrangian) dual* of (P) is given by

$$\max d(\lambda, \mu) \text{ s.t. } \lambda \geq 0 \quad (D)$$

where

$$d : \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R} \cup \{\infty\}, \quad d(\lambda, \mu) := \inf_x L(x, \lambda, \mu).$$

The latter is called the *dual objective*.

⊗ **Example 3.2** (LP): Consider the standard linear program (LP)

$$\min_x c^T x \mid Ax = b, x \geq 0,$$

with $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, c \in \mathbb{R}^n$. The Lagrangian for this problem reads

$$\begin{aligned} L(x, \lambda, \mu) &= c^T x - \lambda^T x + \mu^T (b - Ax) \\ &= (c - \lambda - A^T \mu)^T x + b^T \mu. \end{aligned}$$

We see that $\nabla_x L(x, \lambda, \mu) = c - \lambda - A^T \mu$. Note that the function $x \mapsto L(x, \lambda, \mu)$ is an affine function, which takes its minimum iff $c - \lambda - A^T \mu = 0$, in which case

$$\inf_{x \in \mathbb{R}^n} L(x, \lambda, \mu) = b^T \mu.$$

Otherwise,

$$\inf_{x \in \mathbb{R}^n} L(x, \lambda, \mu) = -\infty.$$

Thus, the dual objective in this case just reads

$$d(\lambda, \mu) = \begin{cases} b^T \mu & \text{if } c = \lambda A^T \mu \\ -\infty & \text{else} \end{cases}$$

Maximizing this function is thusly equivalent to

$$\max_x b^T \mu \text{ s.t. } A^T \mu + \lambda = c, \lambda \geq 0.$$

We can incorporate λ directly into the first constraint, and get the one-variable problem

$$\max_\mu b^T \mu \text{ s.t. } A^T \mu \leq c.$$

III.2.2 Weak and Strong Duality

→ **Theorem 3.6** (Weak Duality): Let \hat{x} be feasible for (P) and $(\hat{\lambda}, \hat{\mu})$ be feasible for (D) . Then, $p(\hat{x}) \geq d(\hat{\lambda}, \hat{\mu})$.

PROOF. We have

$$\begin{aligned} p(\hat{x}) &= f(\hat{x}) \\ &\geq f(\hat{x}) + \underbrace{\hat{\lambda}^T g(\hat{x})}_{\leq 0} + \underbrace{\mu^T h(\hat{x})}_{=0} \\ &= L(\hat{x}, \hat{\lambda}, \hat{\mu}) \geq \inf_{x \in \mathbb{R}^n} L(x, \hat{\lambda}, \hat{\mu}) = d(\hat{\lambda}, \hat{\mu}). \end{aligned}$$

■

From weak duality, $\bar{p} := \inf_{x \in X} p(x) \geq \sup_{\lambda \geq 0} d(\lambda, \mu) =: \bar{d}$, we have the *duality gap* $\bar{p} - \bar{d} \geq 0$ (taking by convention $+\infty - \infty := +\infty$.)

⊕ **Example 3.3** (Nonzero Duality Gap): Consider

$$\min f(x) := \begin{cases} x^2 - 2x & x \geq 0, \\ x & \text{else} \end{cases}$$

such that $g(x) := -x \leq 0$. The Lagrangian reads

$$L(x, \lambda) := \begin{cases} x^2 - (2 + \lambda)x & x \geq 0, \\ (1 - \lambda)x & \text{else} \end{cases}$$

A short computation gives

$$d(\lambda) = \begin{cases} -\frac{(2+\lambda)^2}{4} & \lambda \geq 1, \\ -\infty & \text{else} \end{cases}$$

and thus $\bar{d} = d(1) = -\frac{9}{4}$. On the other hand, the optimal value of the primal is $f(1) = -1$.

→ **Theorem 3.7** (Strong Duality): Consider the convex NLP

$$\min f(x) \text{ s.t. } g_i(x) \leq 0 \forall_{i=1}^m, h_j(x) = 0 \forall_{j=1}^p,$$

with $f, g_i : \mathbb{R}^n \rightarrow \mathbb{R}$ convex and $h_j : \mathbb{R}^n \rightarrow \mathbb{R}$ affine. If Slater CQ holds, then $\bar{p} = \bar{d}$.

PROOF. This one's TOO hard :-(. ■

III.2.3 The Saddle Point Theorem

→ **Definition 3.11** (Saddle Point of a Lagrangian): Let $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$ be the Lagrangian of (P) . Then, $(\bar{x}, \bar{\lambda}, \bar{\mu})$ is called a *saddle-point* of L if the following inequalities hold:

$$L(\bar{x}, \lambda, \mu) \leq L(\bar{x}, \bar{\lambda}, \bar{\mu}) \leq L(x, \bar{\lambda}, \bar{\mu}), \quad \forall (x, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}_+^m \times \mathbb{R}^p.$$

→ **Theorem 3.8** (Saddle Point Theorem): Let $(\bar{x}, \bar{\lambda}, \bar{\mu}) \in \mathbb{R}^n \times \mathbb{R}_+^m \times \mathbb{R}^p$. TFAE:

1. $(\bar{x}, \bar{\lambda}, \bar{\mu})$ a saddle point of L for (P) ;
2. \bar{x} solves (P) , $(\bar{\lambda}, \bar{\mu})$ solves (D) and $f(\bar{x}) = d(\bar{\lambda}, \bar{\mu})$, i.e. the duality gap is zero.

PROOF. (1. \Rightarrow 2.) We see that

$$\begin{aligned}
L(\bar{x}, \bar{\lambda}, \bar{\mu}) &= \underbrace{\inf_{x \in \mathbb{R}^n} L(x, \bar{\lambda}, \bar{\mu})}_{=d(\bar{\lambda}, \bar{\mu})} \leq \sup_{\substack{\lambda \in \mathbb{R}_+^m \\ \mu \in \mathbb{R}^p}} \inf_{x \in \mathbb{R}^n} L(x, \bar{\lambda}, \bar{\mu}) \\
&\leq \inf_{x \in \mathbb{R}^n} \sup_{\substack{\lambda \in \mathbb{R}_+^m \\ \mu \in \mathbb{R}^p}} L(x, \lambda, \mu) \\
&\leq \underbrace{\sup_{\substack{\lambda \in \mathbb{R}_+^m \\ \mu \in \mathbb{R}^p}} L(\bar{x}, \lambda, \mu)}_{=p(\bar{x})} = L(\bar{x}, \bar{\lambda}, \bar{\mu}),
\end{aligned}$$

using the saddle point properties in the first, last equalities. Thus equality must hold throughout, so in particular, $d(\bar{\lambda}, \bar{\mu}) = p(\bar{x})$; moreover $p(\bar{x}) < \infty$ by the same chain of equalities so \bar{x} feasible for (P) . By definition $(\bar{\lambda} \geq 0)$ too, $(\bar{\lambda}, \bar{\mu})$ feasible for (D) . Moreover, by weak duality, \bar{x} solves (P) and $(\bar{\lambda}, \bar{\mu})$ solves (D) .

(2. \Rightarrow 1.) Observe

$$\begin{aligned}
L(\bar{x}, \bar{\lambda}, \bar{\mu}) &\leq f(\bar{x}) (= d(\bar{\lambda}, \bar{\mu})) \\
&= p(\bar{x}) \\
&= \overbrace{\sup_{\lambda \geq 0, \mu} L(\bar{x}, \lambda, \mu)}^{(c)} \\
&= d(\bar{\lambda}, \bar{\mu}) = \underbrace{\inf_x L(x, \bar{\lambda}, \bar{\mu})}_{(b)} \leq \overbrace{L(\bar{x}, \bar{\lambda}, \bar{\mu})}^{(d)}.
\end{aligned}$$

The inequalities $(b) \leq (d) = (a) \leq (c)$ gives the desired result. ■

§III.3 Penalty Methods

Consider

$$\min f(x) \text{ s.t. } x \in X \quad (1).$$

Replace (1) with a sequence of unconstrained problems

$$\min f(x) + \alpha_k r(x), \quad \forall k \in \mathbb{N},$$

where $\alpha_k > 0$ a *penalty parameter*, $r(x) : \mathbb{R}^n \rightarrow \mathbb{R}$ such that $r(x) \geq 0$ and $r(x) = 0 \Leftrightarrow x \in X$. The sum $f + \alpha_k r$ is called the *penalty function*.

III.3.1 The Quadratic Penalty Function

Consider a special instance of (1) given by

$$\min f(x) \text{ s.t. } h_j(x) = 0 \forall_{j=1}^p,$$

for $f, h_j : \mathbb{R}^n \rightarrow \mathbb{R}$ at least continuous. The *quadratic penalty function* for the problem is given by

$$P_\alpha(x) = f(x) + \frac{\alpha}{2} \|h(x)\|^2.$$

This readily gives rise to the (abstract) algorithm Algorithm 3.11.

Quadratic Penalty Method
S0. Choose $\alpha_0 > 0$, set $k = 0$
S1. Determine $x^k \in \mathbb{R}^n$ as a solution of $\min P_{\alpha_k}(x)$
S2. If $h(x^k) = 0$, STOP
S3. Choose $\alpha_{k+1} > \alpha_k$, increment k and go to S.

→ **Theorem 3.9:** Suppose f, h continuous such that the feasible set $X \neq \emptyset$. Let $\alpha_k \uparrow \infty$ and $\{x^k\}$ generated by Algorithm 3.11. Then, the following hold:

1. $\{P_{\alpha_k}(x^k)\}$ is monotonically increasing;
2. $\{\|h(x^k)\|\}$ is monotonically decreasing;
3. $\{f(x^k)\}$ is monotonically increasing;
4. $\lim_k h(x^k) = 0$;
5. Every cluster point of $\{x^k\}$ is a solution of the constrained optimization problem.

PROOF.

1. $P_{\alpha_k}(x^k) \leq P_{\alpha_k}(x^{k+1})$ since x^k minimizes P_{α_k} , and $P_{\alpha_k}(x^{k+1}) \leq P_{\alpha_{k+1}}(x^{k+1})$ since $\alpha_{k+1} > \alpha_k$.
2. We have similarly that $P_{\alpha_k}(x^k) \leq P_{\alpha_k}(x^{k+1})$ and $P_{\alpha_{k+1}}(x^{k+1}) \leq P_{\alpha_{k+1}}(x^k)$, so adding these gives us

$$P_{\alpha_k}(x^k) + P_{\alpha_{k+1}}(x^{k+1}) \leq P_{\alpha_k}(x^{k+1}) + P_{\alpha_{k+1}}(x^k).$$

Both sides of this inequality contain $f(x^k) + f(x^{k+1})$; cancelling these and multiplying by 2, we get

$$\alpha_k \|h(x^k)\|^2 + \alpha_{k+1} \|h(x^{k+1})\|^2 \leq \alpha_k \|h(x^{k+1})\|^2 + \alpha_{k+1} \|h(x^k)\|^2.$$

Rearranging, we get

$$(\alpha_k - \alpha_{k+1}) (\|h(x^k)\|^2 - \|h(x^{k+1})\|^2) \leq 0.$$

Since $\alpha_k - \alpha_{k+1} < 0$, we get the result.

3. From $P_{\alpha_k}(x^k) \leq P_{\alpha_k}(x^{k+1})$, we get

$$f(x^k) + \frac{\alpha_k}{2} \|h(x^k)\|^2 \leq f(x^{k+1}) + \frac{\alpha_k}{2} \|h(x^{k+1})\|^2.$$

Using $\|h(x^{k+1})\|^2 \leq \|h(x^k)\|^2$ (from 2.) and cancelling gives $f(x^k) \leq f(x^{k+1})$, as needed.

4. Observe that $P_{\alpha_k}(x^k) = \inf_{x \in \mathbb{R}^n} P_{\alpha_k}(x) \leq \inf_{x \in X} P_{\alpha_k}(x)$

You can't get worse from picking from a bigger set; that's what the Quebec government doesn't understand.

But if $x \in X$, $P_{\alpha_k}(x) = f(x)$ so further $P_{\alpha_k}(x^k) \leq \inf_X f(x) < \infty$ (since $X \neq \emptyset$). Thus, we have that

$$f(x^0) + \frac{\alpha_k}{2} \|h(x^k)\|^2 \leq P_{\alpha_k}(x^k) < \infty.$$

Letting $k \rightarrow \infty$, $\alpha_k \rightarrow \infty$; but the whole sequence must be bounded, and thus $\|h(x^k)\|^2$ must go to zero (so in particular $h(x^k) \rightarrow 0$).

5. Let \bar{x} a cluster point of x^k , supposing it converges along the whole sequence wlog.

Thus, $h(\bar{x}) = 0$ by continuity and 4., so in particular $\bar{x} \in X$. Moreover, $f(\bar{x}) = \lim_k f(x^k) \leq \lim_k P_{\alpha_k}(x^k) \leq \inf_X f$ (using the bound we used in 4.). Thus, \bar{x} indeed a solution of the corresponding optimization problem. ■

Remark 3.6:

1. We used often the fact $f(x^k) \leq P_{\alpha_k}(x^k) \leq \inf_X f$; in particular if $x^k \in X$, then x^k solves the problem (hence why our stopping criterion is appropriate).
2. Algorithm 3.11, and the corresponding theorem, also apply to the more general problem $\min f$ with both $h(x) = 0$ and $g(x) \leq 0$ (with the components of h, g continuous) because this problem is equivalent to the equality-constraint problem

$$\min f(x) \text{ s.t. } h_j(x) = 0, \max\{g_i(x), 0\} = 0.$$

The corresponding (quadratic) penalty function would then be

$$P_\alpha(x) = f(x) + \frac{\alpha}{2} \|h(x)\|^2 + \frac{\alpha}{2} \sum_{i=1}^p \max\{g_i(x), 0\}^2.$$

Then, P_α remains continuous. In fact, if $f, g_i, h_j \in C^1$, so is P_α ; the only stick point is in whether the function $\tau \mapsto \max\{\tau, 0\}^2$ is C^1 (indeed, without the square this is not true).

Our goal now is, assuming $f, h_j \in C^1$ and given $\{x^k\}$ generated by Algorithm 3.11, to find $\{\mu^k\}$ such that $\{(x^k, \mu^k)\}$ converges to a KKT point.

Observe that $x^k \in \operatorname{argmin}_X P_\alpha \Rightarrow 0 = \nabla P_{\alpha_k}(x^k) = \nabla f(x^k) + \sum_{j=1}^p \alpha_j h_j(x^k) \nabla h_j(x^k)$. Put $\mu_j^k := \alpha_j h_j(x^k)$; these needn't be a KKT point, since we don't know if x^k feasible hence we don't know if $h_j(x^k) = 0$. But it seems like μ_j^k seems like a good choice.

↪ **Theorem 3.10:** With the hypotheses of the previous remarks, let $\{x^k\}$ be generated by Algorithm 3.11 and assume $x^k \rightarrow \bar{x}$, and that LICQ holds at \bar{x} . For $\mu_j^k = \alpha_j h_j(x^k)$, we have μ^k converges to some $\bar{\mu}$ and $\bar{\mu} \in M(\bar{x})$.

PROOF. Define $A_k = h'(x^k) \in \mathbb{R}^{p \times n}$. We see that $A_k \rightarrow h'(\bar{x}) =: \bar{A}$ by continuity. By LICQ, $\bar{A} \cdot \bar{A}^T \in \mathbb{R}^{p \times p}$ is invertible. Thus, $A_k \cdot A_k^T$ is also invertible for k sufficiently large, so $(A_k \cdot A_k^T)^{-1} \rightarrow (\bar{A} \cdot \bar{A}^T)^{-1}$ as well. Thus,

$$\nabla f(x^k) + \sum_{j=1}^p \mu_j^k \nabla h_j(x^k) = 0 \Leftrightarrow A_k^T \mu_k = -\nabla f(x^k),$$

and thus

$$\mu_k = - (A_k A_k^T)^{-1} A_k \nabla f(x^k) \rightarrow - (\overline{A} \overline{A}^T)^{-1} \overline{A} \nabla f(\bar{x}) =: \bar{\mu}.$$

This completes the proof, since \bar{x} feasible as well. ■