MATH378 - Nonlinear Optimization

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

Contents

§1 Preliminaries

§1.1 Terminology

We consider problems of the form

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

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

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

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

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

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

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

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

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

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

then concisely this problem is equivalent to

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

a so-called *least-squares problem*.

We have two related tasks:

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

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

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

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

↪**Definition 1.1** (Minimizers): Let $X \subset \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$. Then $\bar{x} \in X$ is called a

- *global minimizer (of f over X)* if $f(\bar{x}) \leq f(x) \forall x \in X$, or equivalently if $\bar{x} \in \operatorname{argmin}_X f$;
- *local minimizer (of f over X)* if $f(\bar{x}) \leq f(x) \forall x \in X \cap B_\varepsilon(\bar{x})$ for some $\varepsilon > 0$.

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

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

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

Remark 1.1:

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

We recall the following result from calculus/analysis:

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

From, we immediately have the following:

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

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

§1.2 Convex Sets and Functions

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

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

1. *convex (on C)* if
$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y),$$
for every $x, y \in C$ and $\lambda \in (0,1)$;
2. *strictly convex (on C)* if the inequality $\leq$ is replaced with $<$;
3. *strongly convex (on C)* if there exists a $\mu > 0$ such that
$$f(\lambda x + (1-\lambda)y) + \mu\lambda(1-\lambda)\|x-y\|^2 \leq \lambda f(x) + (1-\lambda)f(y),$$
for every $x, y \in C$ and $\lambda \in (0,1)$; we call μ the *modulus of strong convexity*.

Remark 1.2: 3. ⇒ 2. ⇒ 1.

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

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

↪**Proposition 1.2**:

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

§2 Unconstrained Optimization

§2.1 Theoretical Foundations

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

↪**Definition 2.1** (Directional derivative): Let $D \subset \mathbb{R}^n$ be open and $f : D \to \mathbb{R}$. We say f *directionally differentiable* at $\bar{x} \in D$ in the direction $d \in \mathbb{R}^n$ if
$$\lim_{t\to 0^+} \frac{f(\bar{x}+td) - f(\bar{x})}{t}$$
exists, in which case we denote the limit by $f'(\bar{x}; d)$.

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

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

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

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

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

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

using homogeneity of the norm.

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

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

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

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

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

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

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

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

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

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

We recall the following from Calculus:

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

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

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

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

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

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

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

for sufficiently small t, a contradiction. ■

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

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

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

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

2.1.1 Quadratic Approximation

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

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

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

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

a typical quadratic function. Then,

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

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

§2.2 Differentiable Convex Functions

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

1. f is convex (on C) iff

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

   for every $x, \overline{x} \in C$;
2. f is *strictly* convex iff same inequality as 1. with strict inequality;
3. f is *strongly* convex with modulus $\sigma > 0$ iff

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

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

Proof. (1., ⇒) Let $x, \overline{x} \in C$ and $\lambda \in (0, 1)$. Then,

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

which implies

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

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

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

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

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

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

thus proving convexity.

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

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

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

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

and the result follows by dividing both sides by λ.

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

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

↪**Corollary 2.1**: Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable. Then,
a) there exists an *affine function* $g : \mathbb{R}^n \to \mathbb{R}$ such that $g(x) \leq f(x)$ everywhere;
b) if f strongly convex, then it is coercive, i.e. $\lim_{\|x\| \to \infty} f(x) = \infty$.

↪**Corollary 2.2**: Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable, then TFAE:
1. $\bar{x}$ is a global minimizer of f;
2. $\bar{x}$ is a local minimizer of f;
3. $\bar{x}$ is a stationary point of f.

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

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

for any $x \in \mathbb{R}^n$. ■

↪**Corollary 2.3**: (2.2.4)

↪**Theorem 2.6** (Twice Differentiable Convex Functions): Let $\Omega \subset \mathbb{R}^n$ open and convex and $f \in C^2(\Omega)$. Then,
1. f is convex on Ω iff $\nabla^2 f \geq 0$;
2. f is strictly convex on $\Omega \Leftarrow \nabla^2 f > 0$;
2. f is σ-strongly convex on $\Omega \Leftrightarrow \sigma \leq \lambda_{\min}(\nabla^2 f(x))$ for all $x \in \Omega$.

↪**Corollary 2.4**: Let $A \in \mathbb{R}^{n \times n}$ be symmetric, $b \in \mathbb{R}^n$ and $f(x) := \frac{1}{2}x^T A x + b^T x$. Then,
1. f convex $\Leftrightarrow A \geq 0$;
2. f strongly convex $\Leftrightarrow A > 0$.

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

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

Then, the following hold:

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

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

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

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

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

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

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

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

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

§2.3 Matrix Norms

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

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

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

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

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

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

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

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

Supping both sides over all such x gives the result.

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

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

■

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

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

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

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

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

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

Proof. We have for any m,

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

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

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

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

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

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

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

§3 Descent Methods

§3.1 A General Line-Search Method

We deal with the unconstrained problem

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

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

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

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

Proof. $\nabla f(x)^T(-B\nabla f(x)) = -\nabla f(x)^T B \nabla f(x) < 0$. ∎

A generic method/strategy for solving $(\star)$:
S1. (Initialization) Choose $x^0 \in \mathbb{R}^n$ and set $k := 0$
S2. (Termination) If x^k satisfies a "termination criterion", STOP
S3. (Search direction) Determine d^k such that $\nabla f\left(x^k\right)^T d^k < 0$
S4. (Step-size) Determine $t_k > 0$ such that $f\left(x^k + t_k d^k\right) < f\left(x^k\right)$
S5. (Update) Set $x^{k+1} := x^k + t_k d^k$, iterate k, and go back to step 2.

Remark 3.1: a) The generic choice for d^k in 3. is just $d^k := -B_k \nabla f(x^k)$ for some $B_k > 0$. We focus on:
- $B_k = I$ (*gradient-descent*)
- $B_k = \nabla^2 f(x^k)^{-1}$ (*Newton's method*)
- $B_k \approx \nabla^2 f(x^k)^{-1}$ (*quasi Newton's method*)

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

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

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

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

(i) decreasing f along $x^k + td^k$ as much as possible;

(ii) keeping computational efforts low.

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

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

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

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.

3.1.1 Global Convergence of Algorithm 3.1

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

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

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

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

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

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

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

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

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

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

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

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

§3.2 The Gradient Method

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

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

it's know that

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

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

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

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

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

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

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

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

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

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

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

Proof. Suppose not, i.e.

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

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

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

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

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

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

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

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

Proof. Left as an exercise. ∎

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

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

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

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

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

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

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

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

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

which implies σ negative, a contradiction. ■

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