Optimization and Machine Learning M - Theorems and Definitions

Dante Piotto

spring semester 2024

Contents

Chapter 1

Non Linear Programming

1.1 Unconstrained Optimization

The problem to be solved is defined as:

$$\min f(x) \quad x \in \mathbb{R}^n$$

1.1.1 Necessary conditions

Definition 1.1 (descendant direction)
a vector $d \in \mathbb{R}^n$ is a *descendant direction* for function f in x if $\exists \delta > 0 : f(x + \alpha d) < f(x) \quad \forall \alpha \in (0, \delta)$. We denote with $D(x)$ the set of all descendant directions for f in x

Definition 1.2 (stationary point)
A point $x \in \mathbb{R}^n$ is a *stationary point* for f if $\nabla f(x) = 0$

Theorem 1.1 (Firs-Order Necessary Condition)
Let $f \in C^1$. If $\bar{x} \in \mathbb{R}^n$ is a local minimum for problem (1.1), then $\nabla f(\bar{x}) = 0$

Proof. Let $\bar{x} \in \mathbb{R}^n$ be a local minimum for problem (1.1). The proof is by contradiction, thus assume that $\nabla f(\bar{x}) \neq 0$. Define a direction $d^* = -\dfrac{\nabla f(\bar{x})}{\|(\bar{x})\|_2}$ and a point $y = \bar{x} + \alpha d^*$, for some $\alpha > 0$. It follows that $y \neq \bar{x}$ for any value of $\alpha > 0$.

For a sufficiently small value of α, one can approximate function f in y according to the Taylor series up to the first order as follows:

$$f(y) = f(\bar{x}) + \nabla f(\bar{x})^T (y - \bar{x}) + R_1(\bar{x}, \alpha) = f(\bar{x}) - \alpha \|\nabla f(\bar{x})\|_2 + R_1(\bar{x}, \alpha)$$

with $\lim_{\alpha \to 0} \dfrac{R_1(\bar{x}, \alpha)}{\alpha} \to 0$

Thus, for a sufficiently small value of α, the associated point y is such that $f(y) < f(\bar{x})$, giving a contradiction with the hypothesis that $\bar{x}$ is a local minimum □

Theorem 1.2 (Second-Order Necessary Condition)
Let $f \in C^2$ if $\bar{x} \in \mathbb{R}^n$ is a local minimum for problem (1.1), then

1. $\nabla f(\bar{x}) = 0$
2. $d^T \nabla^2 d \geq 0$

Proof. The first condition has alread been proved in the previous theorem

We now prove condition 2 by contradiction, and assume this condition is not satisfied by a local minimum $\bar{x} \in \mathbb{R}^n$. Thus, assume that $\nabla^2 f(\bar{x})$ is not positive semidifinite.

Since 2 is not satisfied, it is possible to find a vector $d^* \in \mathbb{R}^n$ such that ${d^*}^T \nabla^2 f(\bar{x}) d^* < 0$. Note that $d^* \neq 0$. For the sake of simplicity, assume that d^* has been normalized so as to have $\|d^*\| = 1$. Define a new point $y = \bar{x} + \alpha d^*$ for some scalar α, and note that $y \neq \bar{x}$ for all $\alpha > 0$

For a sufficiently small value of α, one can approximate function f in y according to the Taylor series up to the second order as follows:

$$f(y) = f(\bar{x}) + \nabla f(\bar{x})(y - \bar{x}) + \frac{1}{2}(y - \bar{x})^T \nabla^2 f(\bar{x})(y - \bar{x}) + R_2(\bar{x}, \alpha)$$

with $\lim_{\alpha \to 0} \dfrac{R_2(\bar{x}, \alpha)}{\alpha} = 0$

As condition 1 states that $\nabla f(\bar{x}) = 0$, and

$$(y - \bar{x})^T \nabla^2 f(\bar{x})(y - \bar{x}) = (\alpha d^*)^T \nabla^2 f(\bar{x})(\alpha d^*) = \alpha^2 d^{*T} \nabla^2 f(\bar{x}) d^* < 0$$

then we get

$$f(y) = f(\bar{x}) + \frac{1}{2}\alpha^2 d^{*T} \nabla^2 f(\bar{x}) d^* < f(\bar{x})$$

This implies that for any sufficiently small value of α there exists a point y for which $f(y) < f(\bar{x})$, which contradicts the hypothesis that $\bar{x}$ is a local minimum. □

Theorem 1.3 (Second-Order Sufficient Condition)
Let $f \in C^2$. A solution $\bar{x} \in \mathbb{R}^n$ that satisfies the following conditions:

1. $\nabla f(\bar{x}) = 0$
2. $\nabla^2 f(\bar{x})$ is positive definite

is a (strict) local minimum for problem (1.1)

Proof. Let $\bar{x} \in \mathbb{R}^n$ be a solution that satisfies conditions 1 and 2. Let $\rho > 0$ and define a neighbourhood of $\bar{x}$ with radius ρ as follows:

$$N(\bar{x}, \rho) = \{y \in \mathbb{R}^n : \|y - \bar{x}\| \leq \rho\}$$

Let $y \in \mathbb{N}(\bar{x})$ be a point in this neighbourhood that is distinct from $\bar{x}$, i.e., defined by some $d \in \mathbb{R}^n$ with $\|d\| = 1$ and some $\alpha > 0$. The Taylor series for function f in y up to the second order is:

$$f(y) = f(\bar{x} + \alpha d) = f(\bar{x}) + \nabla f(\bar{x})^T \alpha d + \frac{1}{2}(\alpha d)^T \nabla^2 f(\bar{x})(\alpha d) + R_2(\bar{x}, \alpha) = f(\bar{x}) + \frac{1}{2}\alpha^2 d^T \nabla^2 f(\bar{x}) d + R_2(\bar{x}, \alpha)$$

where the last equality derives from condition 1.

For a sufficiently small value of α, the last term $R_2(\bar{x}, \alpha)$ is negligible. Thus, recalling the properties of positive definite matrices we have

$$f(y) \geq f(\bar{x}) + \frac{1}{2}\alpha^2 \lambda_{min}$$

where λ_{min} is the smallest eigenvalue of matrix $\nabla^2 f(\bar{x})$. As this is a positive definite matrix, we have $\lambda_{min} > 0$. This implies that $f(y) > f(\bar{x})$ for sufficiently small $\alpha > 0$ □

1.2 Algorithms for unconstrained optimization

Iterative schemes:

$$x^{k+1} = x^k + \alpha_k d^k$$

- $d^k \in \mathbb{R}^n$, $\|d^k\| = 1$ search direction
- $\alpha_k \in \mathbb{R}_+$ step size

1.2.1 Line Search Algorithms

1. if x^k is optimal stop
2. determine a descendent direction d^k for the objective function
3. determine the step size α_k along direction d^k starting from x^k
4. define the nuew solution $x^{k+1} = \alpha_k d^k$ and iterate

Determining the search direction

Typically

$$d^k = -D^k \nabla f(x^k)^T$$

where D^k is symmetric and nonsingular. Whenever D^k is positive definite, d^k is a descendant direction.

The gradient method

based on the approximation of the objective function f according to the Taylor series up to the first order

$$f(x^k + \alpha d) = f(x^k) + \alpha \nabla f(x^k)^T d$$

considering this expression as a function of d we get a minimum for

$$d^k = -\frac{\nabla f(x^k)}{\|\nabla f(x^k)\|}$$

Newton's method

second order Taylor approximation:

$$f(x^k + h) = f(x^k) + \nabla f(x^k)^T h + \frac{1}{2} h^T \nabla^2 f(x^k) h$$

setting to zero the gradient wrt h:

$$h = -\nabla^2 f(x^k)^{-1} \nabla f(x^k)$$

so the algorithm takes:

$$d^k = -\frac{\nabla^2 f(x^k)^{-1} \nabla f(x^k)}{\|\nabla^2 f(x^k)^{-1} \nabla f(x^k)\|} \quad \text{and} \quad \alpha^k = \|\nabla^2 f(x^k)^{-1} \nabla f(x^k)\|$$

Modified Newton's method

Performance of Newton's method can be improved by calculating step size according to a line search algorithm

Quasi-Newton's method

To reduce computational effort one can compute the search direction as

$$d^k = -\bar{B}^{-1} \nabla f(x^k)$$

where matrix $\bar{B}$ is some approximation of the current Hessian matrix. In particular, we can write

$$\nabla f(x^{k+1}) \simeq \nabla f(x^k) + \nabla^2 f(x)(x^{k+1} - x^k)$$

hence

$$\bar{B}(x^{k+1} - x^k) \simeq \nabla f(x^{k+1}) - \nabla f(x^k)$$

Step size selection

Let

$$\phi^k : \mathbb{R}_+ \to \mathbb{R}, \alpha \to \phi^k(\alpha) = f(x^k + \alpha d^k)$$

The "best step size" for iteration k is

$$\alpha^k = \arg\min_{\alpha \geq 0} \phi(\alpha)$$

but can be computationally expensive

$$\phi'(\alpha) = \nabla f(x^k + \alpha d^k)^T d^k = 0$$

id $d^k = -\nabla f(x^k)$ the best step size policy gives $\nabla f(x^{k+1})^T \nabla f(x^k) = 0$, i.e. for every pair of consecutive iterations the gradients of function f are orthogonal to each other. Can produce fluctuations in the resulting objective function value by producing a new point that is quite far from the previous one.

limited best step size

A common choice is to impose a maximum value for the distance between consecutive points:

$$\alpha^k = \arg \min_{0 \leq \alpha \leq \bar{\alpha}} \phi(\alpha)$$

Constant step size

faster

Wolfe Conditions

Allow to determine an approximate solution for the problem; typically good performance in convergence and computing time

Conditions require that α be such that:

$$f(x^k + \alpha d^k) \leq f(x^k) + c_1 \alpha \nabla f(x^k)^T d^k$$
$$\nabla f(x^k + \alpha d^k)^T d^k \geq c_2 \nabla f(x^k)^T d^k$$

where $0 < c_1 < c_2 < 1$ are two parameters of the algorithm. The first condition is know as *Armijo condition* and can be rewritten as

$$f(x^k) - f(x^k + \alpha d^k) \geq -c_1 \alpha \nabla f(x^k)^T d^k$$

This condition ensures that α is improving wrt $\alpha = 0$ for function $\phi(\alpha)$, with a value reduction taht is proportional to α and to $\phi'(0) = \nabla f(x^k)^T d^k$. This condition does not ensure convergence, hence the second condition, known as *curvature condition*. We can rewrite it as

$$\phi'(\alpha) \geq c_2 \phi'(0)$$

When the condition is satisfied, it signifies that we cannot expect much decrease of the objective function by increasing α, and it is not satisfied for small values of α.

Algorithm:

1. set $i = 0$ and determine an initial value $\alpha(0)$
2. compute $f(x^k + \alpha(i) d^k)$
3. if $f(x^k + \alpha(i) d^k) > f(x^k) + c_1 \alpha(i) \nabla f(x^k)^T d^k$ set $\alpha(i+1) = \alpha(i)/2, i = i + 1$ and goto step 2
4. if $\nabla f(x^k + \alpha(i) d^k)^T d^k < c_2 \nabla f(x^k)^T d^k$ set $\alpha(i+1) = 2\alpha(i), i = i + 1$ and goto step 2
5. set $\alpha_k = \alpha_i$ and return

Typically, the value for c_1 is very small (e.g. $c_1 = 10^{-4}$), while c_2 is considerably larger (e.g. $c_2 = 0.9$). It can be proven that, beside pathological conditions, the Wolfe conditions define at least one interval $[\alpha_1, \alpha_2]$ that includes candidate values for the next step size.

1.2.2 Trust-region algorithms

A region T in which an approximation $\tilde{f}$ of the cost function is considered to be valid. The search direction is given by

$$p^k = \arg \min\{\tilde{f}(x^k + p) : x^k + p \in T\}$$

Typically the trust region is defined by all points within a distance of x^k and the approximating function $\tilde{f}$ is given by the Taylor series up to the second order. For this choice the determination of p^k requires optimizing a quadratic function over a convex set.

In practical algorithms the region size is chosen according to the performance of the algorithm during previous iterations: The size of the trust region is updated according to the ratio

$$r_k = \frac{f(x^k) - f(x^k + p)}{\tilde{f}(x^k) - \tilde{f}(x^k + p)}$$

Values close to 1 indicate that the model is consistently reliable and the trust region may be increased, whereas if r_k is small the model is an inadequate representation of the objective function over the current trast region which should be reduced in size.

It can be proved that, if the x^k points generated belong to a bounded set, then there exists a limit point of the sequence that satisfies the second order necessary conditions.

1.3 Constrained Optimization

Theorem 1.4 (Gordan's Theorem)
Let A be an $m \times n$ matrix. The system $Ax < 0$ has no solution iff there exists a $y \in \mathbb{R}^m, y \geq 0, y \neq 0$ such that $A^T y = 0$

Proof. Given the $m \times n$ matrix A, define the following problems:

$$P_1: \text{ is there an } x \in \mathbb{R}^n \text{ such that } Ax < 0?$$
$$P_1: \text{ is there a } y \in \mathbb{R}^m \text{ such that } y \geq 0, y \neq 0 \text{ and } A^T y = 0?$$

Observe that it cannot happen that both problems have answer "yes". Assume indeed that there exists both an $x \in \mathbb{R}^n$ such that $Ax < 0$ and a $y \in \mathbb{R}^m$ such that $y \geq 0, y \neq 0$ and $A^T y = 0$. We have $0 = 0^T x = (A^T y)^T x = (y^T A)x = y^T(Ax) = y^T z < 0$, where we introduced $z = Ax < 0$, and the last inequality derives from $z < 0, y \geq 0$ and $y \neq 0$

Now assume that problem P_1 has answer "no" and define the following sets:

$$S_1 = \{z \in \mathbb{R}^m : z < 0\} \quad \text{ and } \quad S_2 = \{z \in \mathbb{R}^m : z = Ax \text{ for some } x \in \mathbb{R}^n\}$$

As $S_1 \cap S_2 = \emptyset$ there should exist an hyperplane, associated with a vector $y \in \mathbb{R}^m$, that separates S_1 and S_2, i.e. such that

$$y^T z < 0 \quad \forall z \in \S_1 \quad \text{and} \quad y^T z \geq 0 \quad \forall z \in S_2$$

Vector y must satisfy $A^T y = 0$; indeed, if $A^T y \neq 0$, one could define $\bar{x} = -(y^T A)^T = -A^T y$, such that $\bar{x} = 0$. Imposing $y^T Ax \geq 0$ for $x = \bar{x}$ we should have $0 \leq (y^T A)\bar{x} = (-\bar{x}^T)\bar{x}$, while this is impossible as $\|\bar{x}\| > 0$

Furthermore, by definition y satisfies $y^T z < 0 \quad \forall z \in S_1$. In order to check possible y vectors that satisfy these conditions, let's impose this condition for different z vectors. In particular, we consider m distinct vectors $\tilde{z}_j \in \mathbb{R}^m$, one for each $j = 1, \ldots, M$, the j-th being defined as follows: $\tilde{z}_j = -\varepsilon 1^T - e_j$. For every $\varepsilon \in (0, 1)$, each vector z_j has all components that are negative, hence it belongs to S_1. Thus, for $j = 1, \ldots, m$ and $\forall \varepsilon > 0$ it should be $y^T \tilde{z}_j = -\varepsilon 1^T y - y_j < 0$, which implies that y cannot be the null vector and $y \geq 0$. Thus, y is a solution of problem P_2 that has anser "yes". Summarizing: if problem P_1 has answer "no", then P_2 has answer "yes". This concludes the proof. □

1.3.1 Fist-order necessary conditions

We consider optimization problems with explicit constraints

$$\begin{aligned} \min f(x) & \\ x \in \mathbb{R}^n & \\ g_i(x) \leq 0 & \qquad i \in I \\ h_j(x) = 0 & \qquad j \in E \end{aligned}$$

and we assume $f, g_i, h_j \in C^1$

Definition 1.3 (feasible direction)
A vector $d \in \mathbb{R}^n, d \neq 0$ is a feasible direction in $x \in F$ if $\exists \delta > 0 : x + \alpha d \in F \quad \forall \alpha \in (0, \delta)$

Definition 1.4 (descendant direction)
A vector $d \in \mathbb{R}^n, d \neq 0$ is a descendant direction for f in $x \in F$ if $\exists \delta > 0 :: f(x + \alpha d) < f(x) \quad \forall \alpha \in (0, \delta)$

Theorem 1.5
Let $f : F \to \mathbb{R}$ be a continuous function. if $\bar{x} \in F$ is a local minimum for problem (P), then $D(\bar{x}) \cap F(\bar{x}) = \emptyset$

Proof. The proof is by contradiction. Assume that the thesis is false: there exists a vector $d \in D(\bar{x}) \cap F(\bar{x})$ and two positive numbers δ_1, δ_2 such that $f(\bar{x} + \alpha d) < f(\bar{x}) \quad \forall \alpha \in (0, \delta_1)$ and $f(\bar{x} + \alpha d) \in F \quad \forall \alpha \in (0, \delta_2)$. It follows that, for every $\alpha \in (0, \min\{\delta_1, \delta_2\})$, the point $y = \bar{x} + \alpha d$ belongs to F and has $f(y) < f(\bar{x})$, i.e., $\bar{x}$ cannot be a local minimum. □

Special case: only inequalities

Definition 1.5 (set of active constraints)
we define with

$$I_a(\bar{x}) = \{i \in I : g_i(\bar{x}) = 0\}$$

the set of active constraints

Definition 1.6 (set of feasible directions)
We define with

$$F_s(\bar{x}) = \{d \in \mathbb{R}^n, d \neq 0 : \nabla g_i(\bar{x})^T d < 0 \forall i \in I_a(\bar{x})\}$$

Theorem 1.6
Let $f : F \to \mathbb{R}$ be a continuous function. If $\bar{x} \in F$ is a local minimum for (P), then $D(\bar{x}) \cap F_s(\bar{x}) = \emptyset$

Proof. Let $\bar{x} \in F$ be a local minimum. By contradiction, assume that there exists a vector $d \in \mathbb{R}^n$ such that $d \in D(\bar{x}) \cap F_s(\bar{x})$. As $F_s(\bar{x}) \subseteq F(\bar{x})$ it muyst be $d \in D(\bar{x}) \cap F(\bar{x})$, thus contradicting Theorem 1.5 □

Theorem 1.7 (Fritz-John conditions)
Let $f \in C^1$ and $g_i \in C^1 \forall i \in I$. If $\bar{x} \in F$ is a local minimum for f over F, then there exist scalar numbers λ_0 and $\lambda_i (i \in I)$ such that

1. $\lambda_0 \nabla f(\bar{x}) + \sum_{i \in I} \lambda_i \nabla g_i(\bar{x}) = 0$
2. $\lambda_i g_i(\bar{x}) = 0 \quad \forall i \in I$
3. $\lambda_0 \geq 0, \lambda_i \geq 0 \quad (\forall i \in I)$ and not all λ are zero

Proof. Let $\bar{x} \in F$ be a local minimum. Let $, = |I_a(\bar{x})|$, and define an $(m+1) \times n$ matrix A in which

- row 0 corresponds to $\nabla f(\bar{x})^T$
- each row $i (i = 1, \ldots, m)$ corresponds to $\nabla g_i(\bar{x})^T$

According to theorem 1.6 there exists no vector $d \in \mathbb{R}^n$ such that

$$\nabla f(\bar{x})^T d < 0 \quad \text{and} \quad \nabla g_i(\bar{x})^T d < 0 \quad \forall i \in I_a(\bar{x})$$

i.e., there exists no vector $d \in \mathbb{R}^n$ such that $A^T d < 0$

Using Gordan's Theorem 1.4, this implies the existence of $m + 1$ scalars $\lambda_i \geq 0 (i = 0, \ldots, m)$ that are not all equal to zero and such that

$$\lambda_0 \nabla f(\bar{x}) + \sum_{i \in I_a(\bar{x})} \lambda_i \nabla g_i(\bar{x}) = 0$$

Setting $\lambda_i = 0 \quad \forall i \notin I_a(\bar{x})$ we prove 1. By construction, thus, each constraint $i \notin I_a(\bar{x})$ has $\lambda_i = 0$; as each remaining constraint $i \in I_a(\bar{x})$ has $g_i(\bar{x}) = 0$, 2. follows. Finally, 3. is a direct consequence of Gordan's Theorem. □

Definition 1.7 (Fritz-John point)
A point $x \in F$ is a Fritz-John point if it satisfies the Fritz-John conditions.