CO466 Continuous Optimization

Riley Jackson

September 23, 2020

1 Introduction

Let $f : \mathbb{R}^n \to \mathbb{R},\ g : \mathbb{R}^n \to \mathbb{R}^m,\ h : \mathbb{R}^n \to \mathbb{R}^p$ all continuous and consider the problem

$$\begin{aligned} \inf \quad & f(x) \\ \text{s.t.} \quad & g(x) \leq \mathbb{0} \\ & h(x) = \mathbb{0} \end{aligned} \tag{P}$$

Write $S \coloneqq \{x \in \mathbb{R}^n : g(x) \leq \mathbb{0}, h(x) = \mathbb{0}\}$ for the set of solutions to (P). S is the feasible set of (P), also called the feasible region of (P).

Definition $\bar{x} \in \mathbb{R}^n$ is a **global minimizer** of (P) if $x \in S$ and $f(x) \geq f(\bar{x})$ for all other $x \in S$. Often we will simply call $\bar{x}$ a **minimizer** of (P).

Definition $\bar{x} \in \mathbb{R}^n$ is a **local minimizer** for (P) if $\bar{x} \in S$ and there is some neighborhood U of $\bar{x}$ such that $f(\bar{x}) \leq f(x)$ for all $x \in U \cap S$.

Definition $\bar{x} \in \mathbb{R}^n$ is a **strict local minimizer** of (P) if $\bar{x} \in S$ and there is some neighborhood U of $\bar{x}$ such that $f(\bar{x}) < f(x)$ for all $x \in U \cap S \setminus \{\bar{x}\}$.

Definition $\bar{x} \in \mathbb{R}^n$ is an **isolated minimizer** of (P) if $\bar{x} \in S$ and there is some neighborhood U of $\bar{x}$ such that $\bar{x}$ is the only local minimizer of (P) in $S \cap U$.

Of course, all isolated minimizers are strict local minimizers but the converse is not true (for a counterexample, consider the function $x^2 \cos(1/x) + 2x^2$).

Definition A **Continuous Optimization Problem** is a problem of optimizing (minimizing or maximizing) a continuous function of finitely many real variables subject to finitely many equalities and inequalities on continuous functions of said variables. In other words, a continuous optimization problem is a problem in the form of (P)

A natural question to ask is what kinds of problems can formulated as continuous optimization problems? Almost everything!

Example (Fermat's Last Theorem) "There do not exist positive integers x, y, z and an integer $n \geq 3$ such that $x^n + y^n = z^n$."

Consider

$$\begin{aligned} \inf \quad & f(x) \coloneqq (x_1^{x_4} + x_2^{x_4} - x_3^{x_4})^2 + \sin(\pi x_1)^2 + \sin(\pi x_2)^2 + \sin(\pi x_3)^2 + \sin(\pi x_4)^2 \\ \text{s.t.} \quad & g_1(x) \coloneqq 1 - x_1 \leq 0 \\ & g_2(x) \coloneqq 1 - x_2 \leq 0 \\ & g_3(x) \coloneqq 1 - x_3 \leq 0 \\ & g_4(x) \coloneqq 3 - x_4 \leq 0 \end{aligned} \tag{P}$$

The objective value is non-negative, and clearly is only zero if $x_1^{x_4} + x_2^{x_4} = x_3^{x_4}$ with x_1, x_2, x_3, x_4 integers. The constraints give the conditions that $x_1, x_2, x_3 \geq 1$ and $x_4 \geq 3$, so the minimum value of this continuous optimization problem is 0 and attained if and only if Fermat's last theorem is false.

In fact, it is easy to find a feasible sequence $\left(x^{(k)}\right)$ such that $f\left(x^{(k)}\right) \to 0$, hence Fermat's last Theorem is equivalent to determining whether the optimizer of (P) is attained.

From the above example we see that continuous optimization problems are notoriously hard even when the number of variables is small. Furthermore we can formulate discrete structure such as integral constraints in continuous optimization problems. Finally, we notice that we used highly "non-linear" functions.

Example (Combinatorial Optimization and $\{0,1\}$ Integer Programming) Let $m, n \geq 1$, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$ be given. Consider the $\{0,1\}$ integer program

$$\begin{aligned} \min \quad & c^T x \\ \text{s.t.} \quad & Ax \leq b \\ & x \in \{0,1\}^n \end{aligned} \tag{IP}$$

Notice that we can let $g(x) \coloneqq Ax - b$ so that $g(x) \leq \mathbb{0}$ covers the linear constraint. Furthermore, by letting

$$h(x) = \left(x_1(1-x_1), \ldots, x_n(1-x_n)\right)^T$$

we satisfy the binary constraint with $h(x) = \mathbb{0}$. Since the objective function is clearly linear and thus continuous, this problem can be formulated as a continuous optimization problem.

From the above example, we see that combinatorial optimization problems can be posed as continuous optimization problems with only "mildly" non-linear (quadratic) functions. Since many combinatorial optimization problems are NP-Hard, we see that continuous optimization problems can be very hard even with "simple" constraints.

Thus to successfully solve continuous optimization problems we must study the problem at hand and exploit special structure and properties.

1.1 Conic Form of Continuous Optimization Problems

Definition A set $K \subseteq \mathbb{R}^n$ is a **cone** if $\forall_{x \in K} \forall_{\lambda \in \mathbb{R}_+} \lambda x \in K$.

Definition A set $S \subseteq \mathbb{R}^n$ is **convex** if $\forall_{x,y \in S} \forall_{\lambda \in [0,1]} \lambda x + (1-\lambda) y \in S$, ie: convex sets contain line segments.

Definition A set $K \subseteq \mathbb{R}^n$ is a **convex cone** if it is both convex and a cone.

Let $g : \mathbb{R}^n \to \mathbb{R}^m$, $f : \mathbb{R}^n \to \mathbb{R}$ be continuous functions and consider

$$\begin{aligned} \inf \quad & f(x) \\ \text{s.t.} \quad & g(x) \preceq_K \mathbb{0} \end{aligned}$$

where $K \subseteq \mathbb{R}^m$ is a convex cone and for $u, v \in \mathbb{R}^m$, $u \succeq_K v \iff (u - v) \in K$. This is atleast as general as our earlier formulation as we can take $K = \mathbb{R}^m_+ \oplus \{\mathbb{0}\}$ (the $\mathbb{0}$ on the right is the origin in $\mathbb{R}^p$).

1.2 Calculus

Definition The **directional derivative** of $f : \mathbb{R}^n \to \mathbb{R}$ at $\bar{x} \in \mathbb{R}^n$ along the direction d is defined to be

$$f'(\bar{x}, d) \coloneqq \lim_{\alpha \to 0} \frac{f(\bar{x} + \alpha d) - f(\bar{x})}{\alpha} \qquad \text{(Gatineaux (directional) derivative)}$$

Exercise What is the directional derivative of $f : \mathbb{R}^n \to \mathbb{R}$ given by $f(x) = \|x\|_\infty$ for every $\bar{x} \in \mathbb{R}^n$?

Definition $f : \mathbb{R}^n \to \mathbb{R}^m$ is **differentiable** at $\bar{x} \in \mathbb{R}^n$ if there exists some linear $A : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$\lim_{\substack{h \to \mathbb{0} \\ h \in \mathbb{R}^n}} \frac{\|f(\bar{x} + h) - (f(\bar{x}) + A(h))\|}{\|h\|} = 0$$

Such A is called the **derivative** of f at $\bar{x}$ and is denoted by $Df(\bar{x})$ or $f'(\bar{x})$. We will also use $\nabla f(\bar{x}) = f'(x)$.

Now suppose that $f : \mathbb{E}_1 \to \mathbb{E}_2$, then we have

$$Df : \mathbb{E}_1 \to \mathcal{L}(\mathbb{E}_1, \mathbb{E}_2)$$
$$D^2 f : \mathbb{E}_1 \to \mathcal{L}(\mathbb{E}_1, \mathcal{L}(\mathbb{E}_1, \mathbb{E}_2))$$

If $f : \mathbb{R}^n \to \mathbb{R}$ then $D^k f(\bar{x})\left[h^{(1)}, \ldots, h^{(k)}\right]$ is the k-th directional derivative along the directions $h^{(1)}, \ldots, h^{(k)} \in \mathbb{R}^n$.

Theorem (Taylor's Theorem) *Let $U \subseteq \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$ be a $\mathcal{C}^r$ function on U. Let $x, d \in \mathbb{R}^n$; if $x, x + d \in U$ and the line segment from x to $x + d$ lies in U then there is some z on said line segment such that*

$$f(x+d) = f(x) + \sum_{k=1}^{r-1} \frac{1}{k!} D^k f(x) \overbrace{[d, d, \ldots, d]}^{k \text{ times}} + \frac{1}{r!} D^r f(z) \overbrace{[d, d, \ldots, d]}^{r \text{ times}}$$

Definition Let $U \subseteq \mathbb{R}^n$ be a closed set and let $f : U \to U$, f is called a **contraction mapping** if $\exists_{\lambda \in [0,1)}$ such that

$$\|f(x) - f(y)\| \leq \lambda \|x - y\|$$

for all $x, y \in U$.

We now present a plethora of fixed point theorems.

Theorem (Banach Fixed Point Theorem [1922]) *Let $U \subseteq \mathbb{R}^n$ be a closed set and let $f : U \to U$ be a contraction mapping. Then*

(i) The mapping f has a unique fixed point $\bar{x} \in U$.

(ii) For all $x^{(0)} \in U$, the sequence $\left(x^{(k)}\right)$ generated by $x^{(k+1)} = f\left(x^{(k)}\right)$ will converge to $\bar{x}$ and we have that $\left\|x^{(k)} - \bar{x}\right\| \leq \lambda^k \left\|x^{(0)} - \bar{x}\right\|$.

Theorem (Brouwer's Fixed Point Theorem [1910]) *Let $U \subseteq \mathbb{R}^n$ be a non-empty convex compact set and let $f : U \to U$ be a surjective continuous map. Then there is some $\bar{x} \in U$ such that $f(\bar{x}) = \bar{x}$*

Theorem (Kakutani's Fixed Point Theorem [1941]) *Let $U \subseteq \mathbb{R}^n$ be a non-empty compact convex set and let $f : U \to \mathcal{P}(U)$ be a set valued map on U. If the graph of f, $\{(x, v) \in U \oplus U : v \in f(x)\}$ is closed and $f(x) \neq \varnothing$ and is convex for all $x \in U$ then there is some $\bar{x} \in U$ such $\bar{x} \in f(\bar{x})$.*

Theorem (Borsuk-Ulam Theorem [1930-1933]) *Let $f : \mathcal{S}^n \to \mathbb{R}^n$ be a continuous map. Then there is some $\bar{x} \in \mathcal{S}^n$ such that $f(\bar{x}) = f(-\bar{x})$.*

1.3 Linear Algebra

We denote by $\mathbb{S}^n$ the set of $n \times n$ symmetric matrices, by $\mathbb{S}_+$ the set of $n \times n$ symmetric positive semidefinite matrices, and by $\mathbb{S}_{++}$ the set of $n \times n$ symmetric positive definite matrices.

Theorem (Spectral Decomposition) *For every $A \in \mathbb{S}^n$ there is some orthogonal $Q \in \mathbb{R}^{n \times n}$ such that $A = QDQ^T$ where $D \in \mathbb{R}^{n \times n}$ is diagonal.*

Proof. Let D be the matrix of eigenvalues and let the columns of Q be the associated unit eigenvectors. □

Definition A matrix $A \in \mathbb{S}^n$ is **positive semidefinite** if $h^T A h \geq 0$ for all $h \in \mathbb{R}^n$.

Definition A matrix $A \in \mathbb{S}^n$ is **positive definite** if $h^T A h > 0$ for all $\mathbb{0} \neq h \in \mathbb{R}^n$.

Definition A matrix A is **skew-symmetric** if $A = -A^T$.

Notice that if A is skew symmetric then we have

$$h^T A h = \left(h^T A h\right)^T = -h^T A h \implies h^T A h = 0$$

and therefore A is positive semidefinite.

Theorem (Choleski Decomposition) *Let $A \in \mathbb{S}^n$, then:*

(i) A is positive semidefinite iff there is some $L \in \mathbb{R}^{n \times n}$ lower triangular such that $A = LL^T$

(ii) A is positive definite iff there is some non-singular $L \in \mathbb{R}^{n \times n}$ lower triangular such that $A = LL^T$

1.4 Miscellaneous

Note that in Taylors Theorem above, we required that all functions be real valued, and indeed there is no natural generalization when $f : \mathbb{R}^n \to \mathbb{R}^m$ and $m > 1$, even if $r = 1$ (ie: we only require continuity). There is, however, the following which is similar

Theorem *Let $U \subseteq \mathbb{R}$ be an open set and let $f : U \to \mathbb{R}^m$ be $\mathcal{C}^1$ on U. Suppose for $\bar{x}, d \in \mathbb{R}^n$ we have that the line segment from $\bar{x}$ to $\bar{x} + d$ is contained in U. Then*

$$f(\bar{x} + d) - f(\bar{x}) = \int_0^1 Df(\bar{x} + \alpha d)\, d\, (\partial\alpha)$$

A consequence of this theorem is that if Df is Lipschitz continuous on U with constant L so that

$$\|Df(x) - Df(y)\| \leq L\|x - y\|$$

we get that

$$\begin{aligned}
\|f(\bar{x} + d) - f(\bar{x}) - Df(\bar{x})\, d\|_{\ell_2} &= \left\| \int_0^1 [Df(\bar{x} + \alpha d) - Df(\bar{x})]\, d\, (\partial\alpha) \right\|_{\ell_2} \\
&\leq \int_0^1 \|Df(\bar{x} + \alpha d) - Df(\bar{x})\|_2 \cdot \|d\|_{\ell_2} (\partial\alpha) \\
&\qquad \text{Lemma below and norm inequalities} \\
&\leq \int_0^1 L\|d\|_{\ell_2}^2 \alpha \partial\alpha \\
&= \frac{1}{2} L\|d\|_{\ell_2}^2
\end{aligned}$$

Lemma *Let $h \coloneqq \int_0^1 [Df(\bar{x} + \alpha d) - Df(\bar{x})]\, d\, (\partial\alpha)$, then we have*

$$\|h\|_2 \leq \int_0^1 \|[Df(\bar{x} + \alpha d) - Df(\bar{x})]\, d\|_2 (\partial x)$$

Proof. We have

$$\begin{aligned}
\|h\|_{\ell_2}^2 &= h^T h \\
&= h^T \int_0^1 [Df(\bar{x} + \alpha d) - Df(\bar{x})]\, d\, (\partial\alpha) \\
&= \int_0^1 h^T [Df(\bar{x} + \alpha d) - Df(\bar{x})]\, d\, (\partial\alpha) \\
&\leq \int_0^1 \|h\|_{\ell_2} \|[Df(\bar{x} + \alpha d) - Df(\bar{x})]\, d\|_{\ell_2} (\partial\alpha) \qquad \text{Cauchy Schwarz}
\end{aligned}$$

and the results follows by dividing by $\|h\|_{\ell_2}$. □

Say that $\|d\|_{\ell_2} < \varepsilon$, then the error in the first order estimate of $f(\bar{x} + d)$ is bounded above by $\frac{1}{2} L\varepsilon^2$. Note that we may replace f by Df^r in the above theorem (assuming $f \in \mathcal{C}^{r+1}$) and apply the same reasoning. Therefore it appears (and is the case) that this theorem is useful in the design and analysis of algorithms for continuous optimization.

Theorem (Inverse Function Theorem) *Let $U \subseteq \mathbb{R}^n$ be open, $f : U \to \mathbb{R}^n$ be $\mathcal{C}^1$, $\bar{x} \in U$, and $\det(\nabla f(x)) \neq 0$. Then there is some open neighborhood V of $\bar{x}$ and an open neighborhood W of $f(x)$ such that*

- *$f(V) = W$*
- *f has a local $\mathcal{C}^1$ inverse*
- *$f^{-1} : W \to V$*
- *For every $y \in W$ with $x = f^{-1}(y)$ we have $Df^{-1}(y) = [Df(x)]^{-1}$.*

Note that in the above, if $f \in \mathcal{C}^r$ then there is an $f^{-1} \in \mathcal{C}^r$.

Theorem (Implicit Function Theorem) *Let $h : \mathbb{R}^n \to \mathbb{R}^p$, $h \in \mathcal{C}^1$ in a neighborhood of $\bar{x} \in \mathbb{R}^n$ where $h(\bar{x}) = 0$. Suppose that $h'(x)$ has full row rank ($\operatorname{rank}(h'(x)) = p \leq n$) and define a partition $[B : N]$ of the columns of $h'(\bar{x})$:*

$$h'(\bar{x}) = \begin{bmatrix} | & | & | \\ h'_B(\bar{x}) & | & h'_N(\bar{x}) \\ | & | & | \end{bmatrix}$$

such that $h'_B(\bar{x}) \in \mathbb{R}^{p \times p}$ is non-singular. Partition $\bar{x}$ and x with the same $[B : N]$. Then there is a neighborhood U_B of $\bar{x}_B$ and U_N of $\bar{x}_N$ and a $\mathcal{C}^1$ function $f : U_N \to U_B$ satisfying

- $f(\bar{x}_N) = \bar{x}_B$
- $h\left(\begin{pmatrix} x_B \\ x_N \end{pmatrix}\right) = 0 \iff x_B = f(x_N)$ *forall* $x_B \in U_B, x_N \in U_N$.

Moreover, $f'(x_N) = -\left[h'_B(\bar{x})\right]^{-1} h'_N(\bar{x})$

These are super abstract (lol), lets see an easy example.

Example Consider the very special case where $A \in \mathbb{R}^{n \times p}, \operatorname{rank}(A) = p$ as given in

$$\begin{aligned} \min \quad & c^\top x \\ \text{s.t.} \quad & Ax = b \\ & x \geq \mathbb{0} \end{aligned}$$

Let $h(x) = Ax - b \implies h'(x) = A$ (has full row rank) and notice that $\bar{x}_B = A_B^{-1}b - A_B^{-1}A_N\bar{x}_N$ so $f(x_N) = A_B^{-1}b - A_B^{-1}A_N\bar{x}_N$. Furthermore, we have that $U_B, U_N = \mathbb{R}^p, \mathbb{R}^{n-p}$ respectively since these equations are valid for the whole space (because everything is linear). We can verify the conditions on h and on the derivative of f as well, they also work.

Lemma (Chain Rule) *Let $U \subseteq \mathbb{R}^n, V \subseteq \mathbb{R}^m$ be open sets and let $f_1 : U \to \mathbb{R}^m, f_2 : V \to \mathbb{R}^p$ be differentiable on U and V respectively such that $f_1(U) \subseteq V$. Then $(f_2 \circ f_1)$ is differentiable on U and for all $\bar{x} \in U$ we have*

$$D(f_2 \circ f_1)(\bar{x}) = Df_2(f_1(\bar{x}))\, Df_1(\bar{x})$$

Example (Line Search, Directional Derivatives) Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable on $\mathbb{R}^n$ and we have a point $\bar{x} \in \mathbb{R}^n$ and a "search direction" $d \in \mathbb{R}^n$. We define $\varphi : \mathbb{R} \to \mathbb{R}$ by $\varphi(\alpha) = f(\bar{x} + \alpha d)$, then $\varphi'(\alpha) = \langle \nabla f(\bar{x} + \alpha d), d \rangle$. If $f \in \mathcal{C}^2$ then $\varphi''(\alpha) = d^\top \nabla^2 f(\bar{x} + \alpha d)\, d$. Notice that if $\alpha = 0$ then we have $\varphi'(0) = \langle \nabla f(\bar{x}), d \rangle$ and $\varphi''(0) = d^\top \nabla^2 f(\bar{x})\, d$.

Corollary *Suppose h and $\bar{x}$ are as in the implicit function theorem and assume that $Z \in \mathbb{R}^{n \times q}$ (where $q \leq n - p$) is such that $h'(\bar{x})\, Z = \mathbb{0}$. Then there is a neighborhood U of $\mathbb{0} \in \mathbb{R}^q$ and a $\mathcal{C}^1$ function $t : U \to \mathbb{R}^n$ such that*

- $t(\mathbb{0}) = \mathbb{0}$
- $t'(\mathbb{0}) = \mathbb{0}$
- $h(\bar{x} + Zd_Z + t(d_Z)) = \mathbb{0}$ *for all* $d_Z \in U$

So the function t above gives a way of moving away from $\bar{x}$ (a solution of the non-linear system $h(x) = 0$) in a way that keeps feasible with respect to $h(x) = 0$. So whats the point of the Z matrix? It is a partial description of the null space of $h'(\bar{x})$ and it ensures "first order" feasibility is maintained.

Proof. Let $h, \bar{x}, Z$ be as in the assumptions. Using the partition $[B : N]$, define $Z = \begin{bmatrix} Z_B \\ Z_N \end{bmatrix}$ (recall that $h'(\bar{x}) = [h'_B(\bar{x}) \mid h'_N(\bar{x})]$). Let $U = \{d_Z \in \mathbb{R}^q : (\bar{x}_N + Z_N d_Z) \in U_N\}$ (here U_N is the neighborhood of

x_N given in the implicit function theorem). Define t by $t_N(d_Z) = 0$ and $t_B(d_Z) = f(\bar{x}_N + Z_N d_Z) - \bar{x}_B - Z_B d_Z$. Thus

$$h(\bar{x} + Zd_Z + d(t_Z)) = h\begin{bmatrix} \bar{x}_B + Z_B d_Z + f(\bar{x}_N + Z_N d_Z) - \bar{x}_B - Z_B d_Z \\ \bar{x}_N + Z_N d_Z + 0 \end{bmatrix} = h\begin{bmatrix} f_N(\bar{x}_N + Z_N d_Z) \\ \bar{x}_N + Z_N d_Z \end{bmatrix}$$

which is zero by the implicit function theorem. Also, we clearly have $t(0) = f(\bar{x}_N) - \bar{x}_B = 0$ and $t'_N(0) = 0$. We also have

$$\begin{aligned} t'_B(0) &= f'(\bar{x}_N) Z_N - Z_B && \text{Chain Rule} \\ &= -\left[h'_B(\bar{x})\right]^{-1} h'_N(\bar{x}) Z_N - Z_B && \text{Implicit Function Theorem} \\ &= \left[h'_B(\bar{x})\right]^{-1} \left[-h'_N(\bar{x}) Z_N - h'_B(\bar{x}) Z_B\right] \\ &= \left[h'_B(\bar{x})\right]^{-1} - h'(\bar{x}) Z \\ &= 0 \end{aligned}$$

□

Note In LP's, $t(d_Z) = 0$ because the nullspace of h' describes all possible feasible moves without needing to add a non-linear term.

Corollary *Assume h and $\bar{x}$ are as described in the implicit function theorem. Let $d \in \mathbb{R}^n$ be such that $h'(\bar{x}) d = 0$. Then there is some $\bar{\lambda} > 0$ and a $\mathcal{C}^1$ arc (directed curve) $\hat{t}$ with the properties:*

- $\hat{t}(0) = \bar{x}$
- $h\left(\hat{t}(\lambda)\right) = 0$ *for all* $\lambda \in \left[0, \bar{\lambda}\right)$
- $\hat{f}'(0) = d$

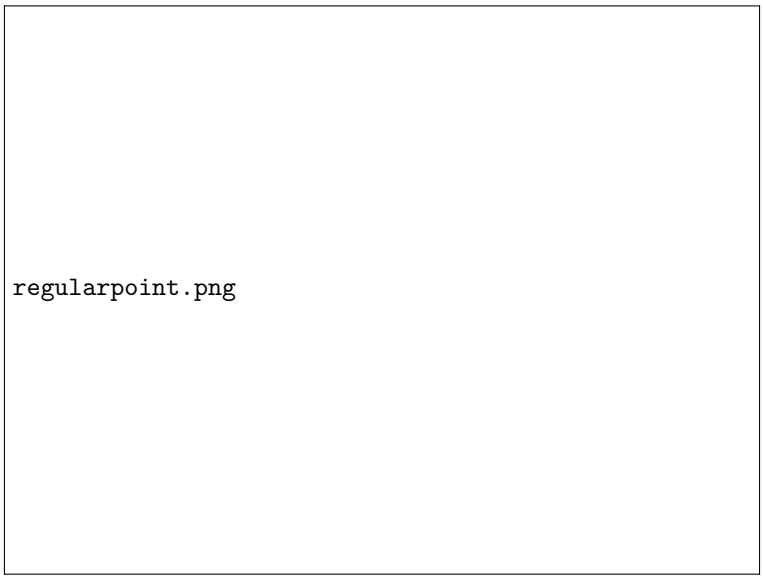
Proof. In the statement of the previous Corollary, let $Z = d$ and using the resulting t, let $\hat{t}(\lambda) = \bar{x} + \lambda d + t(\lambda)$. The corresponding neighborhood is essentially $\left[0, \bar{\lambda}\right)$. □

The picture is roughly

Definition Let $h : \mathbb{R}^n \to \mathbb{R}^p$ with $p \leq n$. We say that a point $\bar{x} \in \mathbb{R}^n$ is **regular** if $\operatorname{rank}(h'(\bar{x})) = p$ (ie: the Jacobian has full rank). We say that a point $\bar{y} \in \mathbb{R}^p$ is a **regular value** if $\forall x \in h^{-1}(\bar{y})$ are regular.

Note $h^{-1}(\bar{y}) = \varnothing$ implies that $\bar{y}$ is a regular value.

Here is a picture:



The set $h^{-1}(\bar{y})$ is thus

If h is affine, then $h(x) = Ax - b$ for some $A \in \mathbb{R}^{p\times n}, b \in \mathbb{R}^p$. Let $\bar{y} \in \mathbb{R}^p$ be given, then $h^{-1}(\bar{y}) = \{x \in \mathbb{R}^n : Ax = \bar{y} + b\}$.

Theorem (Sard's Theorem, Morse-Sard Theorem) *Let $h : \mathbb{R}^n \to \mathbb{R}^p$ where $p \leq n$ and $h \in \mathcal{C}^r$ with $r \geq n - p + 1$. Then the p-dimensional Lebesque measure of $\{y \in \mathbb{R}^p : y$ is not a regular value$\}$ is zero.*

Note Morse [1939] proved the $p = 1$ case, Sard [1942] proved the generalization above. Smale [1965] proved an infinite dimensional extension.

2 Unconstrained Continuous Optimization

Our model is the following:

$$\begin{aligned} \inf \quad & f(x) \\ \text{s.t.} \quad & g(x) \leq \mathbb{0} \\ & h(x) = \mathbb{0} \end{aligned} \tag{P}$$

where $f : \mathbb{R}^n \to \mathbb{R}, g : \mathbb{R}^n \to \mathbb{R}^m, h : \mathbb{R}^n \to \mathbb{R}^p$ and $S = \{x \in \mathbb{R}^n : g(x) \leq \mathbb{0} \text{ and } h(x) = \mathbb{0}\}$. For this section, we assume that $S = \mathbb{R}^n$.

Theorem (First-order necessary conditions) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be $\mathcal{C}^1$ and $S = \mathbb{R}^n$. If $\bar{x} \in \mathbb{R}^n$ is a local minimizer for (P) then $f'(\bar{x}) = 0$ (sometimes $\bar{x}$ is called a **stationary point of** f).*

Proof. Suppose that $f'(\bar{x}) \neq 0$, then there is some $d \in \mathbb{R}^n$ such that $\langle f'(\bar{x}), d\rangle < 0$ (ie: take $a \in \mathbb{S}^n_{++}$ and let $d = -Af'(\bar{x})$). Consider $\varphi : \mathbb{R} \to \mathbb{R}$ given by $\varphi(\alpha) = f(\bar{x} + \alpha d)$. Then $\varphi'(0) = \langle f'(\bar{x}), d\rangle < 0$. Thus for all sufficiently small, positive α we have that $f(\bar{x} + \alpha d) < f(\bar{x})$. Therefore $\bar{x}$ is not a local minimizer for (P). □

Optimality conditions are widely used in algorithm design. For example, many software use $\left\|\nabla f\left(x^{(k)}\right)\right\| < \varepsilon$ as a part of the stopping criteria.

Definition $d \in \mathbb{R}^n$ is a **descent direction for** f **at** $\bar{x} \in \mathbb{R}^n$ if $\langle f'(\bar{x}), d\rangle < 0$. $d \in \mathbb{R}^n$ is an **improving direction for** f **at** $\bar{x}$ if $f(\bar{x} + \alpha d) < f(\bar{x})$ for all sufficiently small $\alpha > 0$.

As we proved above, a descent direction is indeed an improving direction.

Theorem (Second-order necessary condition) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be $\mathcal{C}^2$ and $S = \mathbb{R}^n$. If $\bar{x} \in \mathbb{R}^n$ is a local minimizer of (P) then $f'(\bar{x}) = 0$ and $\nabla^2 f(\bar{x}) \in \mathbb{S}^n_+$.*

Proof. The fact that the gradient vanishes is immediate from the former result. Suppose for contraction that $\nabla^2 f(\bar{x}) \notin \mathbb{S}^n_+$. Since $f \in \mathcal{C}^2$, the Hessian is symmetric, hence there must be some $d \in \mathbb{R}^n$ such that $d^\top \nabla^2 f(\bar{x}) d < 0$. Define $\varphi : \mathbb{R} \to \mathbb{R}$ by $\varphi(\alpha) = f(\bar{x} + \alpha d)$. Then $\varphi'(0) = \langle \nabla f(\bar{x}), d\rangle = 0$ and $\varphi''(0) = d^\top \nabla^2 f(\bar{x}) d < 0$. Therefore, for all $\varepsilon > 0$ and sufficiently small we have $f(\bar{x} + \varepsilon d) < f(\bar{x})$ which contradicts minimality. □

Definition $d \in \mathbb{R}^n$ is called a **direction of negative curvature for** f **at** $\bar{x}$ if $d^\top \nabla^2 f(\bar{x}) d < 0$.

Theorem (Taylor's Theorem - implicit remainder version) *Let $U \subseteq \mathbb{R}^n$ be open, let $f : U \to \mathbb{R}$ be $\mathcal{C}^r$ on U, let $\bar{x}, d \in \mathbb{R}^n$, and assume that $[\bar{x}, \bar{x} + d] \subseteq U$. Then,*

$$f(\bar{x} + d) = f(\bar{x}) + \sum_{k=1}^{r} \frac{1}{k!} D^k f(\bar{x}) \overbrace{[d, \ldots, d]}^{k \text{ times}} + \mathcal{R}(\bar{x}, d)$$

where $\mathcal{R}(\bar{x}, \cdot) : \mathbb{R}^n \to \mathbb{R}$ is such that $\lim_{h \to 0} \frac{\mathcal{R}(\bar{x}, h)}{\|h\|^r} = 0$.

ie: if we are considering small values of d then the first r terms are a very good approximation.

Theorem (Second order sufficient conditions) *Let $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^2$, $S = \mathbb{R}^n$, and let $\bar{x} \in \mathbb{R}^n$. If $f'(\bar{x}) = 0$ and $\nabla^2 f(\bar{x}) \in \mathbb{S}^n_{++}$, then $\bar{x}$ is a strict local minimizer for (P).*

Proof. Let $\delta = \min\left\{d^\top \nabla^2 f(\bar{x}) d : \|d\|_{\ell_2} = 1\right\}$, by Courant Fischer this is $\lambda_{\min}\left(\nabla^2 f(\bar{x})\right)$. By the previous theorem, $\forall d \in \mathbb{R}^n, \|d\|_{\ell_2} = 1$ and $\alpha > 0$ small enough we have

$$f(\bar{x} + \alpha d) = f(\bar{x}) + \alpha \langle \nabla f(\bar{x}), d\rangle + \frac{\alpha^2}{2} d^\top \nabla^2 f(\bar{x}) d + o\left(\alpha^2\right) \geq f(\bar{x}) + \frac{\delta}{2}\alpha^2 + o\left(\alpha^2\right)$$

Choose a neighborhood U of $\bar{x}$ such that $\frac{\delta}{2}\alpha^2 > \left|o\left(\alpha^2\right)\right|$, then forall $x \in U \setminus \{\bar{x}\}$ we have that $f(x) > f(\bar{x})$ and $\bar{x}$ is a strict local minimizer for (P). □

How applicable is this last theorem?

Proposition *Let $f : \mathbb{R}^n \to \mathbb{R}$ be $\mathcal{C}^2$ and consider $\tilde{f}(x) = f(x) + c^\top x$ where $c \in \mathbb{R}^n$ is given. Then for almost all $c \in \mathbb{R}^n$, $\tilde{f}'(\bar{x}) = 0 \implies \nabla^2 f(\bar{x})$ is non-singular.*

Proof. Apply Sard's Theorem to $f(x) \coloneqq f'(x)$ with $r = 1$ and $p = n$. □